THE VALUE OF THE EU PUBLIC DOMAIN

RUFUS POLLOCK[1], PAUL STEPAN[2], AND MIKKO VÄLIMÄKI[3]

Abstract. This paper reports results from a large recent study of the public domain in the European Union. Based on a combination of catalogue, commercial and survey data we present detailed figures both on the prices (and price differences) of in copyright and public domain material and on the usage of that material. Combined with the estimates for the size of the EU public domain presented in the companion paper Pollock and Stepan (2009) our results allow us to provide the first quantitative estimate for the 'value' of the public domain (i.e. welfare gains from its existence) in any jurisdiction.

Keywords: Copyright; Public Domain; Intellectual Property; Europe

1. Introduction

Interest in copyright, and IP more generally, has been growing in recent years, largely as a reflection of the growing importance of cultural 'production' in society and the economy. A natural counterpart to an interest in copyright is an interest in the public domain for the two are closely related: the public domain begins where copyright ends. Moreover, this is a relationship of mutual interdependence: much copyrighted work builds, directly or indirectly, on public domain material; and the public domain of tomorrow is the copyrighted content of today. Thus, study of the public domain is as natural and necessary part of our research efforts as the study of copyright.[1]

[1] Mead Fellow in Economics, Emmanuel College, University of Cambridge
[2] Erasmus University Rotterdam and Austrian Society for Cultural Economics and Policy Studies
[3] Adjunct Professor Helsinki University of Technology and Fellow at University of Eastern Piedmont, Italy
Date: 10th December 2009.
Corresponding author: Rufus Pollock: rp240@cam.ac.uk or comments@rufuspollock.org. The results presented in this paper derive from work conducted as part of 'Public Domain in Europe' project funded by the European Commission DG InfoSoc and run by Rightscom Ltd. We would like to thank the European Commission, Rightscom, and other project members, for making this research possible. We also wish to express special thanks and appreciation to the many organizations who provided us with data including: AMG, BUMA/STEMRA, Philip Harper, Naxos, Nielsen, SIAE, SCPP, YLE, the Slovak National Library and the library of Trinity College Dublin.

[1] In fact they form such natural complements that the two will often be inseparable: for example, in studying cultural production, as we do here, we will necessarily encounter both public domain and copyrighted

The work presented here is part of this effort. Here our focus is on the 'value' of the public domain, that is the social welfare generated by public domain material. As discussed in detail below value could be interpreted broadly or narrowly in this context. Broadly it would mean the social welfare associated with the use of public domain works. Narrowly it would be the social welfare generated by public domain works over and above that generated if those works were in copyright. In general we focus on the narrow interpretation.

Estimating *value* of the public domain is the more important, and the more difficult, aspect of this study. Estimation of value is difficult, especially from an empirical point of view, because it requires data on usage *and* price on some set of works. Furthermore, as we focus on net value (see below), we need to be able to compare the in-copyright and out-of-copyright (public domain) situation.

Thus getting at value requires us to look at three distinct but related areas. First, the usage of public domain material – that is how much public domain material is bought, broadcast, downloaded etc. Second, what are price differences between copyright and public domain material. Third, what are the *differences* in usage corresponding to these differences in price – to what extent are public domain works more used as a result of being more cheaply and easily available?

It should be clear, especially for the last two items, that any results we obtain will be as relevant to analysing copyright as to analysing the public domain – the price reduction for public domain items is the price increase under copyright, etc.

The existing empirical work on copyright is relatively limited and that on the public domain even more so. For example, we know of no work looking at the usage of public domain material.[2] Work on the price effects of copyright is also limited. With the main existing papers being that of Heald (2006) and Liebowitz (2008).[3] Both these authors look at books in the US (these are both discussed below in the relevant sections). On the related matter of availability there is the work of Brooks (2005) which looks at the comparative availability of public domain versus in copyright sound recordings.

material; in studying the optimal term of copyright we must consider works both when in copyright and when they enter the public domain; etc.

[2]There is, unsurprisingly, substantially more work related to the usage of (demand for) copyright material, for example, Hui and Png (2002); Ghose et al. (2004); Gaffeo et al. (2008); Png and hong Wang (2007); Hendricks and Sorensen (2009), though most of this 'indirect' in the sense that usage itself is not the central item of interest.

[3]Again there are a greater number of works dealing with pricing of copyright works, especially in relation to demand. See, for example Bittlingmayer (1992); Clerides (2002); Ringstad and Løyland (2006).

There is also the recent and (relatively) extensive literature on unauthorised file-sharing. This is not directly relevant but has an indirect relationship if one considers unauthorised copies as being have some approximate similarity to public domain copies.[4] Thus, for example, the estimates on demand effects and willingness-to-pay (see e.g. (Le Guel and Rochelandet, 2005; Rob and Waldfogel, 2006)) could potentially be extrapolated to the case of public domain material.

2. Theory

The theory we use here is entirely standard (see e.g. Landes and Posner (1989); Watt (2000)). We shall therefore not go into great detail but rather will give brief overview of how the theory will apply in our case and, in particular, how it relates to our 'empirical strategy'.

2.1. **The General Concept of Value.**

When we talk of 'value' it is important to be clear about what we mean. Here the term value is interpreted to mean the social value (or welfare), which is the usual meaning attributed to the term value by economists. This value, be it of an apple or a novel, is the value derived by a user from its employment or enjoyment – often approximated in monetary terms by willingness-to-pay (WTP) – net of the costs of producing the good. For goods with an associated price, social value may in turn be divided into 'user value' (consumer surplus) – defined as the value to the user of the good net the price paid for it – and 'commercial value' (producer surplus) – defined as the price minus the cost of producing the good (the seller's profit).

Thus the value of a good may be quite different from the price paid for it. This is an important distinction to keep in mind, for it is not unusual to see the value of an activity being equated to its revenue rather than to the utility generated for society. To illustrate this difference consider the case of a novel that goes out of copyright and enters the public domain. Suppose before this occurred the novel was sold for £10 in bookshops but afterwards it is sold for £5 and is also available for free on the internet. Sometimes it is suggested that this results in a reduction in the value of that work for society since before the work was 'worth' £10 but now is 'worth' only £5 or even nothing.

[4]Of course this approximation can be rather crude. Unauthorised file-sharing is limited almost exclusively to digital material. Furthermore, accessing unauthorised copies rather than public domain ones is likely to incur significant costs (actual or potential) due to their unauthorised status (for example, the risk of prosecution, the difficulty of locating material etc).

From the above this can be seen to be completely false. The value of the work has not changed at all. All that has happened is that the price has dropped. A consumer who previously valued the book at, say, £15 and who paid £10 and was left with £5 of 'consumer surplus', now pays £5 (or £0) and is left with £10 (or £15) of 'surplus'.

Furthermore, the reduction in price means that consumers who valued the work at less than £10, and therefore did not buy at the original price, will now be able to purchase it. For each such consumer society gains the entire value they put upon the good (net of costs). Aggregating the valuations of all of these individuals who only get access at the lower price gives the total value to society of having this lower price. Conversely when a monopoly – or some other regulation – restricts access to a good there is a consequential loss to society (termed the deadweight loss).

2.2. **Valuing a Work, and Hence the Public Domain.** The value of a work (at a given price) then is the value of that work to each user summed across those users who gain access to that work (at that price).

This is illustrated in terms of a classic linear demand curve in Figure 1. The first of these shows a single price and quantity (here superscripted IP but this can be ignored at the present). The value of this work would then be equal to the gross surplus generated by the work (equal to the area under the demand curve up to q^{IP}) minus the cost of producing it. However here we are not simply interested in the value of the work but the value of public domain.

The value of the public domain could be given two interpretations, one broad, one narrow. The broad interpretation would be to take the value of the public as the social value (welfare) associated to all the works in the public domain. The narrow interpretation would be define the value of the public domain as the social value (welfare) generated by the work when it is in the public domain minus its value when under copyright. In what follows we shall favour the narrower interpretation and so the value of the public domain will be its 'net' value, that is the value generated by the work being in the public domain over above that it would generate under copyright.

As shown by the figure when a work enters the public domain we expect its price to drop. This has two effects. First, there is a transfer of value from the current owner of the exclusive right to users, second there is a gain in value as new users who previously

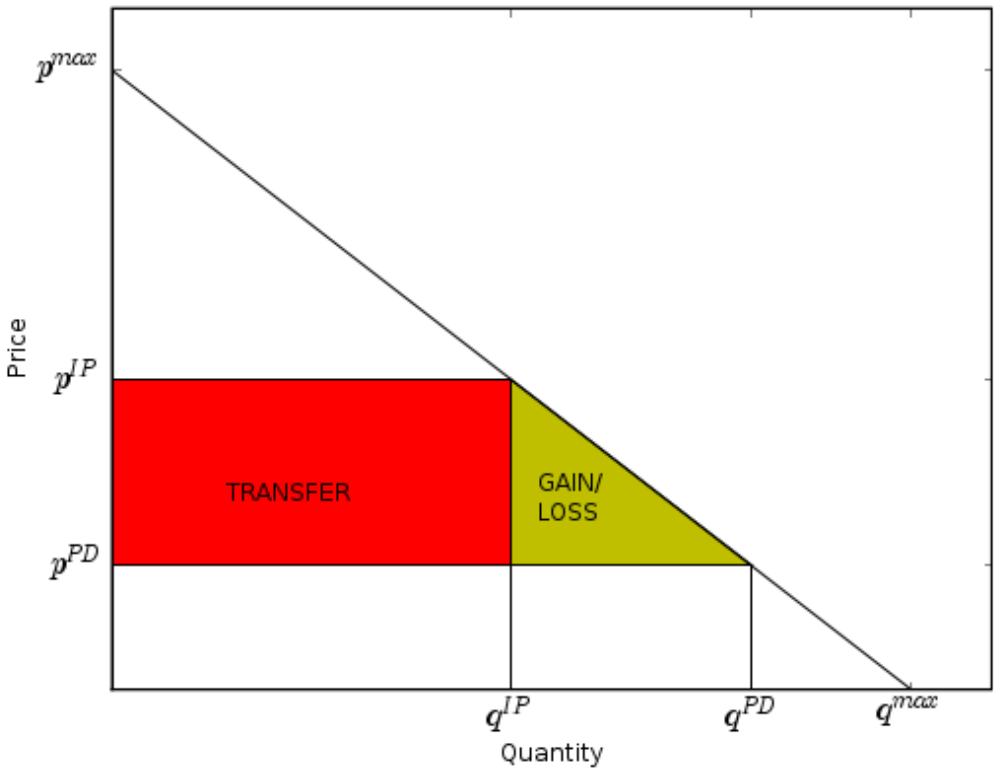

FIGURE 1. The Demand Curve for a Work and the Effect of an IP Right

did not purchase at p^{IP} gain access to the work (this 'gain' is also often referred to as the deadweight loss as this value is *lost* when the work when subject to the IP right).

Thus, the value of the public domain is this gain from greater access to this work. In addition, strictly we should also take into account two additional related effects. First, any reduction in the production of work arising from the income foregone by the owner of the IP right when the work enters the public domain. However, given the current length of copyright (and any related rights) this effect is so slight that it can be ignored for the purposes of calculations in this study. Second, entry in the public domain of a work permits not only greater access, but greater reuse. The benefits of such reuse may not be adequately captured in the basic demand curve for the work[5] and so the simple calculation suggested above would lead to an understatement of the public domain's value. Unfortunately, while this effect may well be important – after all there is substantial anecdotal evidence for the

[5]This is for several reasons, most obviously that downstream re-users are unlikely to be capturing the full surplus generated by their activities.

importance of public domain material in the production of new works – it is extremely difficult, at least with present data, to quantify.[6] Hence, this effect, just like the first, will generally be ignored.[7]

Finally, having established how to determine the value of the public domain for a given work, the next step would be to aggregate these values across all works in the public domain to obtain a single figure.

$$\text{Value of ith work} = \text{Welfare under PD - Welfare under IP} \quad (1)$$

$$v_i = w_i^{PD} - w_{IP} \quad (2)$$

And then:

$$\text{Total value} = \text{Sum of Individual Values across all works} \quad (3)$$

$$V = \sum_i v_i \quad (4)$$

2.3. **Empirical Strategies.** *In theory* then matters are very simple: we start by calculating a demand curve for each individual work in the public domain. At the same time one would determine the price and level of access both when in the public domain and when access is restricted by an IP right. Together this would permit us to calculate the net welfare gain and hence its public domain value. Finally we would then aggregate these values across all works.

Unfortunately, however, we live in a far from perfect world. Given the state of available data the methodology just described is clearly impossible because we do not have data at anything like that level of disaggregation or have the facility to neatly compare prices of the same work, at the same time, when in and out of copyright. Our aim then must be to derive as good an *estimate* as possible given the material available.

[6]Remember that is not sufficient to simply show that reuse of public domain material occurs but it is necessary to calculate the net benefit: i.e. the increased cost (and hence reduced creativity) of going to the next best alternative source.

[7]It should be emphasized that other parts, especially those providing case-studies, will be dealing with this in some detail.

First, it will likely be necessary to focus down on few key types of works that are widely available and used, for example books and sound recordings. Next for each type of work it will probably be best to divide the analysis into two parts:

(1) Estimate the deadweight loss of copyright (conversely the value of the public domain) generally. That is, estimate this as a function of a few key parameters defining the demand curve.
(2) Combine this figure with estimates of how these underlying parameters vary across the population of works.

To elaborate a little. On the first point the key issue will be to estimate the demand curve, or some portion thereof, for the set of cultural works under consideration along with the likely price changes. Where data is insufficient to estimate the full demand system (as is likely to be the case) it may be possible to fall back to a reduced form approach in which one seeks direct estimates for key variables (price changes, demand elasticities etc) and then combines these with a particular parametrisations for the demand curve. Even in this reduced case, it is highly unlikely that one can proceed on an individual work by work basis. Instead it will be necessary to use a whole set of works in order to deal with the very heterogeneous nature of cultural production.

For example, suppose one wishes to calculate the effect of copyright on price (an area on which little empirical work has yet been done). There are two obvious approaches to take. One approach is to track works as they move from being under copyright to being in the public domain. The second approach is try and select a set of 'matching' works some in the public domain and some not and then compare the prices of the two sets of work. In both cases one needs a fairly large sample size to get any kind of robust estimate.[8] The end result of this first process will be some form of average estimates for the impact of copyright on price, sales etc.

At this point we come to the second item. Combining these average estimates for the impact of copyright with information both on the distribution of sales across works and the distribution of sales across time (i.e. how cultural decay affects works), we can estimate the net value of the public domain across whole cohorts of works. Finally, putting this

[8]There is a very large degree of price variation found for these types of products which is *not* a result of the work's copyright status but is due to other factors. To 'filter' this out then requires a large dataset (putting it crudely).

together with our figures for the size of the public domain, we would obtain an overall estimate for the total value of the public domain in the EU.

3. Empirics

As discussed in the theory section it will be impossible to do demand-curve estimation for every single public domain work. Instead we are going to have to come at the problem from a whole variety of different angles, piecing together each piece of information to form an overall picture.

3.1. **Data.** One of our greatest challenges has been obtaining data with which to perform our analysis. An overview of the data we have obtained (as well as some examples of data we requested but did not obtain) is provided in Table 1. In general, it has proved hard to obtain data, especially across a broad range of jurisdictions. The reasons for this are not hard to discern: detailed data on sales and prices is commercially valuable and those who possess it either: a) make significant charges for access or b) are reluctant to provide it at all. Furthermore, and most significant for our work, these 'commercial' datasets are often poorly oriented to investigation of the copyright/public domain: not only do they never contain a PD status indicator but frequently they lack the data needed that would allow us to calculate that status – a matter we now discuss in greater detail.

Title	Source	Country	Amount	Bks	Rec	Other	Size	Usage	Price	PD	(C)	Description/Comment
Recordings Usage	Buma/Stemra	NLD	M		Y			Y			Y	3 years of royalty data on a random selection of items. Only title for work provided (no author, date etc)
Nielsen Sales Data	Nielsen	UK	L	Y				Y	Y	Y	Y	Comprehensive sales data
BNF Info	BNF	FR	XS	Y	?	?	Y			Y		Very limited
NGCOBA	Philip Harper	MULTI	L	Y			N/A	N/A	N/A			New General Catalogue of Books and Authors (list of authors with death dates)
GfK Sales Data	GfK	FR	L	Y	Y			Y	Y	?	Y	Received but unsatisfactory
List of Composers	classical-composers.org	MULTI	M		Y		N/A	N/A	N/A			List of composers with death dates
Re-release of PD recordings	Naxos	UK	S		Y			?	?	Y		List of re-released albums (but no associated sales data)
Price data for recordings	AMG	UK/MULTI?	L		Y				Y	?	Y	Comprehensive classical and pop recordings data
Distribution to members	Literatur-Mechana	AT	M	Y				Y?		Y	Y	disbursement per member (anonymized but with birth/death dates) for all members (10k) 1994-2007
Distribution for all works	AKM	AT	L		Y						Y	Data for all works for 2007
PD/Copyright Music Usage	YLE	FI	M		Y			Y		Y	Y	PD/copyrighted minutes of music on various YLE's (national broadcasting company's) TV and radio channels
Recordings Price Data	Muze	UK	L						Y			
Not Received												
Library loans data	Finnish National Library	FI	?	Y				Y		Y		
Airplay data	Nielsen Music	MULTI	?		Y			Y		Y	Y	
Chart Record Sales	Official Chart Company	UK	?					Y			Y	Received a sample but just listed recordings with no info on price or sales
Recordings airplay data	PPL	UK	?		Y					Y?	Y	
Downloads of PD books	Project Gutenberg	MULTI	?					Y		Y		
Many more ...												

TABLE 1. Data Obtained and Requested

4. Calculating Public Domain Status

Our approach to calculating public domain (equivalently copyright) status was laid out in Pollock et al. (2009). There we explained (see section entitled 'Determining Public Domain Status'):

> Formally, our algorithm for computing the public domain status is:
>
> (1) Given information on an item match it to a work (or works).
> (2) Compute public domain status of the work(s) using, in our simple approach, author death dates and, possibly, (first) publication date.
>
> This seems very simple. Unfortunately, both steps present serious difficulties from a data perspective. As already discussed, associating items to works is hard. That said, for many items all that will matter is the authorial death date and library records do list the author. Unfortunately, the author records often do not have sufficient information with which to compute PD status with certainty – in particular, as we discuss in greater detail below, author death dates are frequently absent.[9] Thus, it may be necessary to fall back on some approximate method [in which PD probabilities are calculated based on publication date] ...

Like the data we were considering there (library catalogue records), all the data available to us for our work on value is about *items*. Unfortunately, our task here is even harder than for the size analysis for several reasons:

- Author dates are entirely absent and are not unambiguously identified. In library catalogues significant effort is made to unambiguously identify authors and author dates (be it only birth dates) are frequently provided. In 'commercial' datasets, by contrast, these features aren't required by users and as are therefore absent.
- Publishers and record labels tend to reissue the same item fairly frequently (if it is not discontinued) thus the publication/release date attached to a given item gives little or no guide to when it (or, more importantly, its associated work) was originally produced.[10] As a result, the publication date of an item gives us

[9]Libraries often record birth and death date only in order to disambiguate two authors of the same name. Furthermore, they add the birth date first (for obvious reasons!) and only add the death date if the birth date turns out later to be insufficient to disambiguate.

[10]This is worse than the simple fact that the same work may get issued in different items. Publishers will often create a new ISBN for a new printing of the same book.

absolutely no guidance as to when the underlying work was originally released. It is worth remembering that the release date is *the* determining factor for copyright in a recording and even for a book it can be extremely useful in inferring PD status when authorial information is absent (e.g. a book published in 1870 is almost certainly in the public domain).

- Even general item information is not as standardized as it is in catalogue data. For example, often the exact form of the title, even for the same item will vary between datasets and that leaves aside the possibility of varying titles for different titles related to the same work (is it "Hamlet" or "William Shakespeare's Hamlet" or "Hamlet by William Shakespeare" or "Hamlet, Prince of Denmark" etc).
- For our work here it is essential that PD status be calculated individually. Thus, unlike for 'size' calculations no approximate cohort-based method based on publication date or the like.

The implication of these facts is that to determination of PD status can not be based on the information in the dataset alone. Instead we are forced to a) obtain some other database of items (or works) which does record the information relevant for calculating PD status b) match records between the two databases.

This is no small requirement. While we can tolerate some level of error in matching and hence status determination, this error cannot be too large or it will render any analysis based on those results unreliable. At the same time, due to the size of the datasets involved, the speed of our algorithms also matters – for example, the Nielsen dataset for 2007 contains over 64,000 titles and if computing public domain status for each title takes 1 second then a full run will take 18 hours. If it takes 30s per title it will take 22 days. These two pressures operate in opposite directions: making the algorithm more accurate slows it down and vice-versa. The practical implications of this trade-off will be seen below where we attempt PD classifications (see e.g. the section on the Nielsen data).

5. Impact on Price

A variety of anecdotal information collected during the course of the research indicated that the copyright status of a work did have a substantial impact on price. For example,

one organization,[11] involved in providing recording for soundtracks to films and televesion, indicated that out-of-copyright recordings would be 70% cheaper than in copyright ones (£20 thousand to £6 thousand fee). Another source suggested that authorial copyright, via its impact on the number of recordings available, might raise the synchronization fee of a classical recording by around 50%.[12]

In Austria, discussions with theatres, opera houses and concert halls, indicated, very approximately, that royalty for a 'PD' work (the actual edition may be in copyright due to critical notes, specific layout etc) is a half of that for an in copyright one (e.g. 15% royalty for in copyright works but only 5% or 6% for works that are in PD – albeit in an edition which may obtain copyright due to editing, critical corrections etc).

While 'anecdotal' evidence is certainly useful, and perhaps indicative, one would prefer proper quantitative estimates based on a substantial dataset. As discussed above obtaining reliable quantitative estimates is not easy. In particular, care is needed when controlling for unobserved differences between works in and out of copyright – for example, the presence of additional material such as notes or an introduction in a public domain book, or differing print and paper quality. Furthermore, the high variance of price data mean that a large dataset is likely to be necessary if the tests are to have sufficient power to be useful. We have attempted to acquire several datasets covering a range of jurisdictions and media but were ultimately successful in obtaining only two substantial datasets both related to the UK: one from Nielsen relating to books, and one from Muze/AMG relating to recordings.[13] We will examine both in detail below.

5.1. **Availability.** Closely related to price is the question of availability. Many older works – both those in the public domain and those not – are simply not available.[14] When a work is not available at all this clearly has an even more dramatic affect on welfare than a simple change in price. Thus, it is important to consider availability as well as price when looking at the impact of copyright status.

[11]The organization explicitly requested to remain anonymous due to fears that a attributed statement could jeopardize his relationship with the large music labels.

[12]Though here one must be cautious since newer work may also be less popular and hence less recorded independent of any copyright impact.

[13]The number of providers of this kind of data is extremely limited. In addition to the data from Nielsen and Muze/AMG we also sought, and obtained, data from GfK in France. Unfortunately, the data from GfK data proved to lack the necessary information required for our work.

[14]Such a situation can be incorporated within the traditional price/demand framework by positing that such works have an 'infinite' (or very high) price – i.e. they are not available at any price.

Again, there is reasonable anecdotal evidence that a work being in the public domain increases the likelihood it is available. However, here we also have some quantitative evidence.[15]

For example, we have data on recordings from SCPP in France (discussed in detail below in its own section) indicates a 50% increase in availability (1680 tracks versus 1098 available). Other evidence is provided by Naxos, whose Historial label reissues public domain recordings. An analysis they performed, indicated that, of the 192 recordings in their historial catalogue originally issued by EMI, only 44 are currently available from EMI. [16]

5.2. **The Nielsen Dataset.** The Nielsen dataset was the most detailed and comprehensive dataset we obtained. It gives almost complete sales data for all books sold in the UK in a given period.[17] Their dataset contains price (RRP and actual), sales (units and revenues), as well as detailed information on the items themselves such as title, author name, ISBN, format, category etc.

5.2.1. *Computation of PD Status.* We began by focusing on a single year (2007) and examining all books selling at least 5 copies – there were 64,000 such publications in total. We tried various approaches, utilizing a variety of datasets and algorithms, for computing a PD status flag for each item in the dataset. These described fully in the appendix. In addition to two automated methods based on the Open Library and NGCOBA dataset, we also classified the entire dataset by hand thanks to the very generous assistance of Mr Harper. This last approach is obviously by far the most reliable – as well as providing much additional information (e.g. indications where an item, though perhaps not PD

[15]There is also some existing work such as that of Brooks (2005). His study focused on recordings in the United States and found that copyright had a significant negative impact on availability compared to the EU.

[16]Naxos have also traditionally marketed their historical (public domain) at very competitive prices (often significantly lower than the price available from the original issuer). The effect, in at least one case according to Naxos, has been to drive down the prices charged by the original record label: "In November 2003 EMI launched a new series entitled EMI Classics Historical – 'a new series documenting the most important historical recordings from its rich catalogue'. This series was marketed at the same price as Naxos and covers recordings made in 1947-54 which had just or were just about to fall into the public domain and have been or were to be released by Naxos." (Footnote 3 of Naxos' submission to the Gowers Review).

[17]Nielsen coverage of the first-hand market is not quite complete: sales through some independent booksellers may not be included. However, they estimate they cover well over 90% of all books sold and include all the major booksellers both online and off.

itself, was a translation of a PD work or where critical notes had been added). However, it is also laborious (we had received annual datasets covering the full 2001-2008 period).

Our hope therefore was that an automated approach would give good performance as we could then this automated method to the larger datasets available. Unfortunately none of our automated approaches appeared to have performed well: our NGCOBA-based automated classifier only had a correlation of 0.4 with our hand classifications (and the Open Library based results were clearly worse). Such a low correlations clearly make such classifications inadequate for use in our analysis. While, no doubt, the algorithms could be improved (perhaps to a point of usability), the effort required to do so was likely to have been significant. Given the time and resource constraints of this project this is not something we were able to do. We have therefore limited ourselves to analyzing the 2007 data for which have good hand classifications. We would note that even with this constraint the dataset we are analyzing is of a size and comprehensiveness much greater than anything done before.[18]

Public Domain?	**No. of Items**	**%**	**No. of Items (Pbk)**	**%**
N	56857	88.95	39605	93.21
Y	3868	6.05	1873	4.41
Y?	235	0.37	56	0.13
YN	12	0.02	0	0.0
YT	2864	4.48	940	2.21
YT?	80	0.13	17	0.04
Unknown	1	0.0	1	0.0
Total	63917	100	42492	100
PD (Broad)	7059	11.04	2886	6.79

TABLE 2. PD Status Counts for Nielsen Dataset. N=No, Y=Yes, YT=Translation of PD work (may or may not be PD itself), YN=PD work with notes, '?' indicates there may be some uncertainty.

The first results from this dataset are shown in Table 2 which shows the number of items with each PD status in the 64k dataset. The various classifications are explained in the caption, however one thing we should add is that the notes category ('YN') refers

[18]The only two existing papers in this area we are aware of are Heald (2006) and Liebowitz (2008), both of which are focused on the US (book) market. Heald (2006) studies bestselling books from the US and finds that public domain works both have more editions available and are significantly lowered priced compared to in copyright works ($3.85 versus $8.05). Liebowitz (2008) also looks at books across the 1923 divide (all books before 1923 are in the public domain in the US) and finds no significant difference in price between in and out of copyright works based on a standard regression (altering the regression to weight works based on sales alters this to a significant 14.5% difference in price. In both cases the sample size is small compared to the dataset available to us.

only to books which are predominantly 'notes' (e.g. the York Notes series). Many of the 'pure' PD items (classified 'Y') will include an additional introduction and critical notes – and the PD classification refers to the main text contained therein and not this additional material.

As the table shows nearly **4,000 items, corresponding to 6% of the total dataset, are definitely PD, with this figure rising to over 7,000 (11%) if we take a broad approach to the public domain definition** – in particular, including translations of public domain works (e.g. the Odyssey, Madame Bovary etc). This is a substantial number and interestingly corresponds, very approximately, with the 'size' of the public domain (as a proportion of all work) computed in our earlier chapter.

In addition to the total figures we have have also shown figures for paperbacks alone (based on the ONIX code classification). As the figures show the public domain proportions here are somewhat lower at around 4.4% for 'pure' PD and 6.8% for 'broad' PD, still quite a substantial number. Adding in pure hardbacks (classified BB by ONIX) raises the total items to 47,616 or which 4.5% are pure PD and 6.8% are broad PD.[19]

5.2.2. *Price.* Our next step is to look at price and sales (usage). Disaggregating by format here is important since comparing the price of hardbacks and paperbacks is not appropriate.[20] This is part of a more general point: in comparing the prices of our two groupings (PD versus in-copyright) we should try to ensure that all 'other things are equal' (or explicitly control for those differences in the regression). This is for two reasons. First, and more importantly, if there are variables (other than PD status) in the dataset which are both correlated with price and with status this could bias the results (for example, if all public domain books are twice as long as in copyright ones). Second, greater homogeneity (e.g. all paperback) increases the power of the test.

The Nielsen data provides a wealth of information for each item including total sales (units sold) and two pieces of price information: recommended retail price and average

[19]This shows that a good proportion of the Nielsen catalogue consists of items which either have unclassified format or are not 'pure' books (e.g. they audiobooks, mixed media etc). Calculations show that the vast majority (14,049) are unclassified. A brief examination by hand indicate that many of these are normal books – leaving the reason for their lack of classification something of a mystery. Interestingly then, the implication of our PD counts, is that the proportion of these unclassified items which are PD is significantly higher than the level for formally classified as 'pure' books.

[20]Including both paperback and hardback together would not necessarily invalidate the results though it certainly would weaken the power of the test by increasing the variance of the samples.

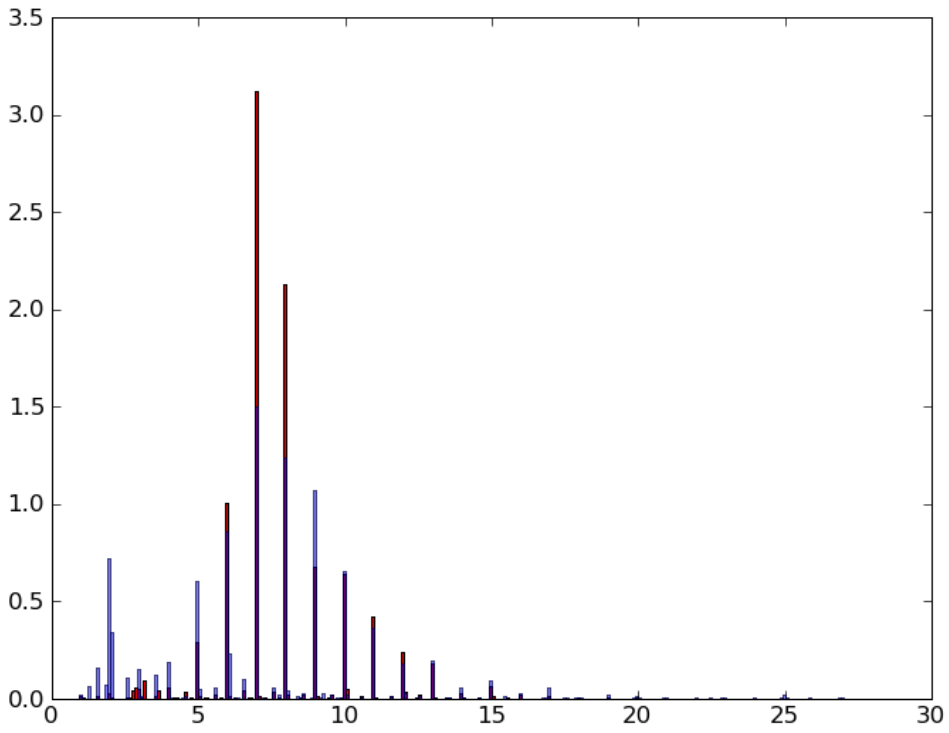

FIGURE 2. RRP (recommended retail price) distribution for PD (blue) and copyright (red) paperback books.

selling price (ASP) which is calculated by taking total sales revenue and dividing by the number of units sold. Figures 2 and 3 show the price distribution for both RRP and ASP. The distribution's have been 'normalized' and this allows us to plot the (pure) PD and copyright price distributions on the same graph (remember that copyright books outnumber PD ones by 20 to 1).

Interestingly, in both cases, the distribution appears to be (very) approximately normal (or log-normal in the case of ASP). As is to expected the RRP price distribution is rather sparser with clumping at regular intervals of a pound (to be precise it will be at £X.99). There are also obvious differences between the PD and copyright price distributions in both graphs. In the RRP one PD price distribution though strikingly similar in overall shape is significantly more dispersed than the copyright one. The same is true of the ASP

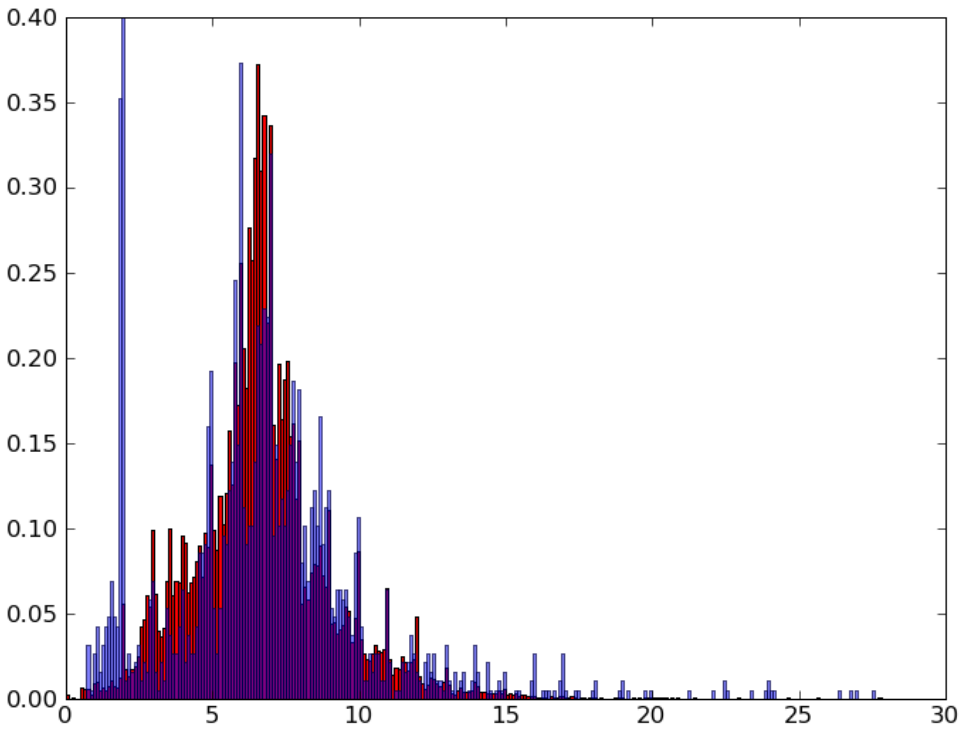

FIGURE 3. ASP (average selling price) distribution for PD (blue) and copyright (red) paperback books

graph with the PD distribution having noticeably higher densities at the upper end of the distribution.[21]

Strikingly in both ASP and RRP graphs there is a very significant anomaly with a spike in the PD distribution at the low end of the price range around the £2-3 level. Both the greater dispersion and this 'spike' are what we would expect: for public domain works anyone can produce an edition and this will permit entry at both ends of the market: at the top-end with 'luxury' and special (e.g. large-print, bespoke-format) editions, and at the lower end with budget editions marketed at the very lowest cost possible. This is borne out by an examination of the raw data: the large huge number of public domain titles clustered around the £1.99 level, many from from well-known classics publishers/brands

[21]Note that both distributions have been truncated at the £30 price level. For both PD and in-copyright books there were a few items with very high prices but the densities above 30 were very low. The distributions have also been truncated at the 0 level. Rather surprisingly it turns out that books can end up with an ASP below zero. This occurs due to 'returns' which result in a negative entry in the ASP . If these occur in sufficient numbers and a book was originally sold at a discounted price then the net effect can be negative (number sold at discounted price minus number of returns times full price).

such as Penguin Popular Classics, Wordsworth Classics, Dover Thrift Editions, Pocket Classics, Everyman Poetry etc. Similarly at the high-end most editions were 'special' in some way:[22] for example, the most expensive paperback (£26.95 RRP) was a large-print edition of Trollope's Barchester Towers, similarly a £23 edition of Jane Austen's Emma etc etc.[23]

This dispersion, though expected, is problematic for our price analysis. To illustrate consider this hypothetical example: an in-copyright work has one edition ('item') with a price of 7.99. When it goes public domain the old edition continues to be sold at its original price but two new editions come on the market one at at 4.99 and one at 20. The result: in copyright the average price was 7.99 but in the public domain it has actually increased to 10.99! What this brings to the surface is a general and important point: in estimating the price effect of the PD (or conversely copyright) we are really interested in the 'price' for works not for items. Of course a work price doesn't exist but must be constructed from the price(s) of the associated item(s) – the most obvious measure being the minimum price (for reasons provided by that hypothetical example). It implies that a simple mean price analysis based on items will yield an upwards biased PD price estimate (relative to the correct works-based one) and hence an upwards biased estimate of the PD effect on price.[24]

With this in my mind let us now look at the summary statistics for the distributions shown in the previous figures provided in Table 3 and 4. In the RRP case, there is a clear

[22]'Luxury' editions were primarily hardbacks. For example, the most expensive PD book was the 'The Nonesuch Dickens Boxed Set', a hard-back with an RRP of £125 – though the seven Dickens novels included in this set could each be picked up for £2 or less. Similarly, the highest price *item* was an deluxe audio version of 'The Complete Sherlock Holmes' costing £180. However, of the complete works alone there were 3 book versions (1 hardback, 2 paperback), the most expensive being a Collector's library edition £29.99 hardback and the cheapest a 2-volume Barnes and Noble edition costing £8 total (the other was a Penguin edition retailing for £18.99). Moreover the majority of the Holmes oeuvre is available separately in editions costing under £2 with the 'Adventures of the Dancing Men and Other Sherlock Holmes Stories' available at an RRP of 95p!

[23]Trollope's Barchester Towers could be found in 4 other editions: one hardback at £10.99 and 3 paperbacks the cheapest of which cost £6. For Austen the situation was similar: 5 editions alone below £3.50 starting with a Wordsworth Classics at £1.99, a Penguin Popular Classics at £2, an Everyman at £2.99, a Dover Thrift at £3.50 and a Penguin Classics for £3.50. At least the last of these (the only one we checked in this detail) included along with the text, an introduction, a chronology, notes and suggestions for further reading.

[24]Another potential upward bias comes from the existence of 'multi-work' items, i.e. editions including multiple individual books – e.g. 'The Complete Works of Jane Austen'. While these can exist for both in-copyright and public domain works it is likely these occur more frequently for public domain material. Unfortunately there is no flag in the dataset indicating multi-volume works. However, crude methods based on searching for specific phrases (e.g. 'in 1:' or 'The Complete') indicated that this conjecture was correct.

	No. Items	Mean	SD	Median	Min	Max
Copyright	35807	7.92	3.59	6.99	0.0	308.0
PD	1863	7.3	3.79	6.99	0.95	26.95
All	37670	7.88	3.6	6.99	0.0	308.0

TABLE 3. Summary Statistics for Paperback RRP Data. Number of items may differ from Table 2 as some items do not have an RRP.

	No. Items	Mean	SD	Median	Min	Max
Copyright	39605	6.71	2.49	6.57	0.0	86.01
PD	1873	6.67	3.54	6.63	0.75	27.5
All	41478	6.71	2.54	6.57	0.0	86.01

TABLE 4. Summary Statistics for Paperback ASP Data

difference in average prices of around 8% which a basic t-test confirmed was significant at a very high level. In the ASP case, the mean prices are almost identical (the difference is not statistically significant) with the standard deviation for PD books significantly higher – bearing out our concern about the effects of high-end price entry under the public domain.[25]

There are various ways to address the 'entry' problem a) use a *work* price computed as the minimum of the prices of all associated editions ('items') b) weight by prices by sales on the assumption that new low-price editions will sell more heavily than the new luxury editions c) right truncate the price distribution discarding all prices above a given level. The first of these options is rendered impractical by the difficulty of doing work-item matching.[26] Option 2 is a possibility but we are concerned about its robustness given the multiple ways price and quantity are related.[27] We have therefore chosen the last option.

The logic of this approach is that entry into the public domain is unlikely to result in an increase in the (minimum) price for a work. Thus, right truncation, in discarding high

[25]The difference between the RRP and ASP cases appears to be explained largely by the larger discounting of in-copyright books relatively to PD ones (the right tail for PD prices is significantly fatter relative to in-copyright prices in the ASP figure than in the RRP figure). This makes sense: high price PD items are usually 'special' and are much less likely to be the only edition of that work than in-copyright items. Thus, discounting for such high-price PD is less likely since buyers will be less price-sensitive (there are already cheaper editions available) and the item may already be selling close to cost (such editions are likely to have small print-runs).

[26]This problem was discussed at some length in the 'size' section of this report.

[27]Calculating mean price per sale yields £4.96 for public domain versus £5.65 for in-copyright material. However, the distribution of test statistic such as this difference in means, is hard to calculate since the theoretical correlation of price and quantity is non-obvious. There is also the difficult of the causal relationship of the two variables: lower prices would be assumed to increase sales, while the existence of fixed costs, mean that lower anticipated demand will lead to reduced print-runs and higher costs.

RRP Limit 10						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
Copyright	31431	7.22	1.57	6.99	0.0	10.0
PD	1618	6.28	2.55	6.99	0.95	10.0
All	33049	7.17	1.65	6.99	0.0	10.0
RRP Limit 15						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
Copyright	35411	7.76	2.17	6.99	0.0	15.0
PD	1804	6.89	3.05	6.99	0.95	14.99
All	37215	7.72	2.23	6.99	0.0	15.0
ASP Limit 10						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
Copyright	36433	6.23	1.79	6.45	0.0	10.0
PD	1683	5.86	2.41	6.31	0.75	9.99
All	38116	6.22	1.82	6.45	0.0	10.0
ASP Limit 15						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
Copyright	39363	6.63	2.25	6.56	0.0	15.0
PD	1826	6.34	2.87	6.56	0.75	14.95
All	41189	6.62	2.28	6.56	0.0	15.0

TABLE 5. Summary Statistics for Paperback with Truncation

price public domain and high price in copyright editions equally, will introduce no PD 'bias'. However, it will help correct – though not completely eliminate – the 'entry' bias since we will have eliminated high-priced editions of PD works (which occur, relative to in-copyright works, more frequently for the reasons discussed).

The results when performing this right truncation are shown in Table 5 for truncation at both £10 and £15. The percentage number of items truncated is relatively similar across PD and copyright groups[28] but the impact is substantial with price differences increasing to **around 11-13% for RRP and 4-6% for ASP (these differences all being highly statistically significant)**. These figures are reasonably close to standard figures given for the 'copyright' fraction of a books cost (5-20% – the bulk of a book's retailed price being accounted for in printing, distribution, marketing and retailer margin). Furthermore, we must reiterate that this sort of calculation almost certainly **under-estimates** the impact of the public domain on price and especially welfare.

[28]This similarity in percentages is not inconsistent with our earlier point about the greater dispersion in PD distribution as the distribution of items within the truncated section regions will be different between PD and in-copyright material.

This is not only because average price is a poor proxy for minimum price but because we are unable to make a perfectly like-for-like comparison. The 'comparison' problem is two-fold: first, public domain editions frequently have additional material (notes, introduction etc); second, because we are doing a cross-sectional analysis, we have the problem of not knowing whether the two distributions we have are generally comparable are comparing are suited to a simple comparison of *means*.[29]

To make clear what this means look back at Figure 3 and consider the many ways we could add, move, or remove 'mass' from the red distribution (copyright) in order to make the 'blue' (public domain).[30] It is clear there are many ways for this to occur which result in the average price reduction for a given work being substantial (say 20% or more) but the average impact is zero. For example, consider a book which is only made available once it is in the public domain (for example, that large print edition of Emma discussed previously). The previous price for this work then was 'infinity' (or some very large amount). Thus, the price reduction here is significant even though it will appear in the distribution as a relatively expensive item. Similarly, looking again at the Figure 3, it is noteworthy that the *modal* public domain price is £2 but £6.40 for in-copyright. If one were to suppose these to be the comparable work sets then the price reduction would have been a huge 70% – and even being more conservative and taking it to be in the £4-5 region for copyrightable material this would be over 50%.[31]

5.3. **Muze/AMG.** We also obtained detailed price data on recordings from Muze/AMG. Unlike Nielsen's book dataset this material no usage information but it does contain detailed item metadata (title, composer, performer, release date etc) along with information on the dealer price for these recordings (all data as of early 2009). The dataset contains the majority of recordings commercially available in the UK in standard formats (CD etc).

[29]The obvious way to address this problem is to conduct a longitudinal study in which one can follow works as they enter the public domain. This however has difficulties of its own, most importantly that the proximity of the public domain will itself have an affect on price due to anticipation by an owner of copyrights expiry as well as competition from temporally proximate but already public domain works.

[30]We can add and remove mass because the set of in-print public domain and in copyright books do not necessarily match.

[31]If one considers this over-generous it is worth considering that the various different Harry Potter books sell for between £5.99 and £7.99 (RRP with ASP being not much reduced from this). Given their popularity these are clearly books that could be printed in bulk and, hence, if in the public domain, would be offered at a budget paperback price of around £2.

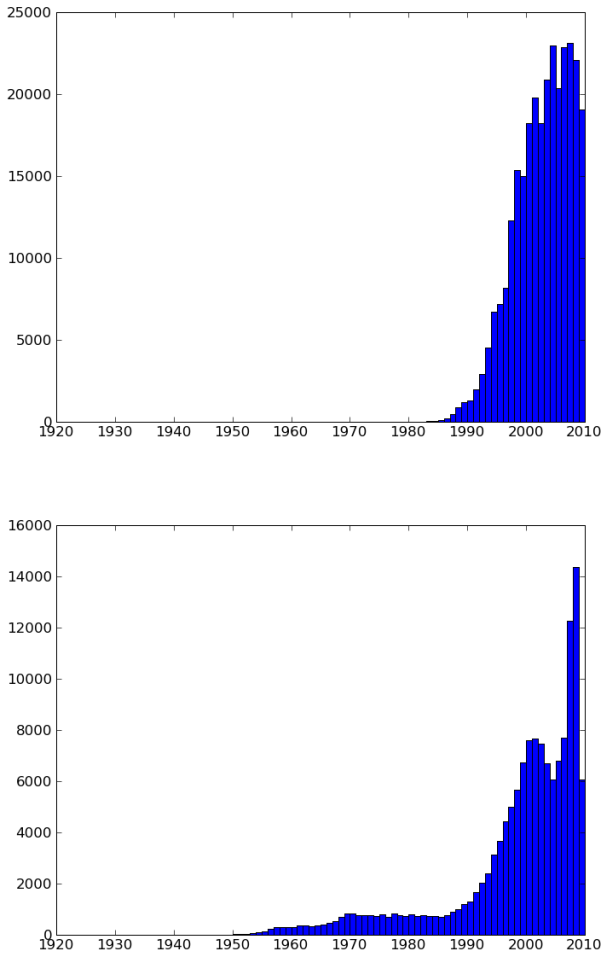

FIGURE 4. Counts by release date (top) and original release date (bottom) for UK recordings from Muze/AMG dataset

In our analysis we focused exclusively on material available on CD.[32] The dataset is also divided into classical and pop music categories which we analyzed separately.

Already aware the problems of doing automated author identifaction and lacking a detailed composer database we focused our attention on 'recording copyright'. Thus when we speak of public domain recordings we mean those whose recording copyright has expired.[33]

[32]Obviously price comparison can only be done across material in the same format. Moreover, the vast amount of the data related to CDs.

[33]Thus, our public domain recordings need not be 'fully' public domain i.e. having had both their 'recording' copyright *and* authorial copyright expire. For more discussion on this distinction see the earlier section on the size of the public domain.

In the UK, at the present time, recording copyright expires 50 years after first publication. Thus, in theory, determination of public domain status only requires to have available the release date of the recording. Unfortunately this is not easy.

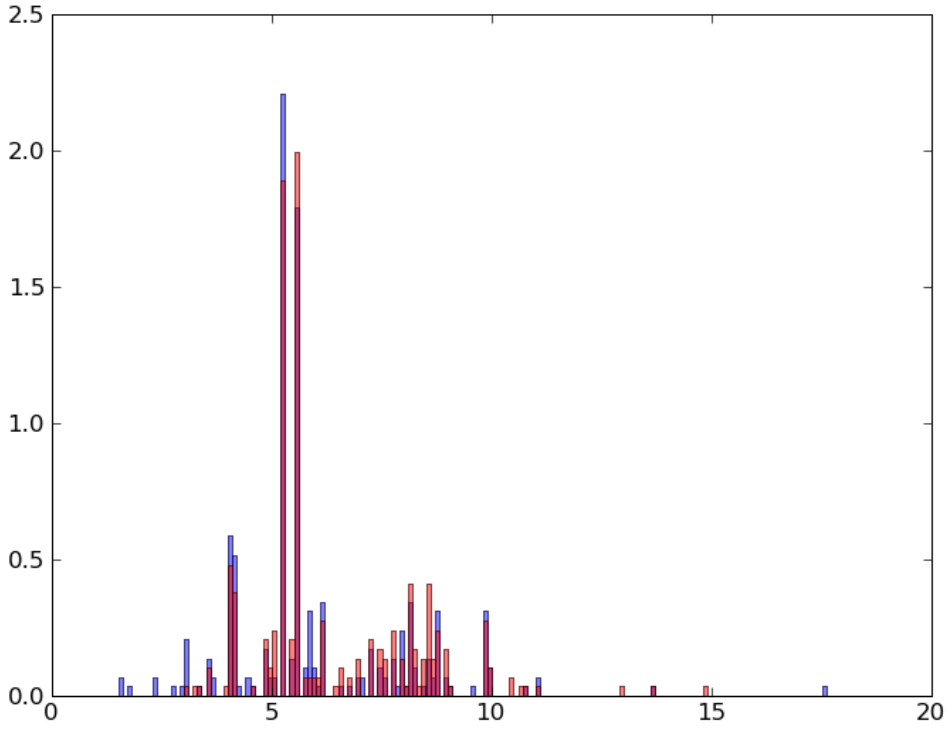

FIGURE 5. Price distribution for public domain (blue, originally released 1958) and in-copyright CDs (red, originally released 1960).

As discussed above in section 4, there is an important distinction to be drawn between 'items' and 'works'. Our dataset contains information on 'items'. Their release date is usually the date on which that item (i.e. that particular CD) was first put into circulation. This need not necessarily have any relation to when that particular work (i.e. the underlying recording) was *first* released – and for older recordings this is especially true since CDs have only been available since the 1980s, and hence almost all 'release dates are post-1980.

For copyright status we therefore require information on the *original* release date of the recording. Fortunately, some of the Muze data had an 'original release date' field. However, this field was not available for classical music and was only available for approximately

140 thousand out of the total of 290 thousand items. Figure 4 show the distribution of recordings by year based on release date and original release date. Of these 140 thousand items with an original release date, 1318 (1%) were public domain (released before 1959). This 1% figure could be taken as an estimate of the proportion of *commercially available* recordings which are public domain. However, we would be cautious in doing this given the very large number of recordings lack any original release date.[34]

1958 v. 1960						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
PD	290	5.97934482759	1.96563982346	5.55	1.53	17.5
Not PD	291	6.38855670103	1.87234806165	5.55	3.0	14.85
1957 v. 1959						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
PD	279	6.03781362007	2.162704538	5.55	1.38	17.5
Not PD	283	6.50462897527	3.03312142339	5.55	1.53	38.4

TABLE 6. 1 Year Comparisons for Pop Music Recordings

We focus then on this set of 140 thousand pop recordings for which relevant data is available. Our next step is to compare the prices of public domain and in copyright material. In order to limit variation in price caused by other factors[35] the sets of PD and copyright recordings are chosen to be relatively proximate in time: specifically we choose sets of years either side of the PD cutoff date. In our case the PD cutoff date is 1959 and we have taken PD recordings to be those from pre-1959 (1958 and before) and in copyright recordings as those from post-1959 (1960 and after).[36] In addition, given the potential lag in data we have also presented results for the case where 1958 is the cutoff date (so PD is 1957 and before and in-copyright is 1959 and after).

The results of comparing prices of 1 year post and 1 year pre copyright expiry are shown in Table 6 (the full price distribution is shown in Figure 5). As this shows, **in copyright CDs are, on average, 41p (6.7%) more expensive than public domain recordings** (for 1957 vs. 1959 it is 47p and 7.7%). Performing a standard t-test for difference in means

[34]In fact, it seems likely that presence of an original release date field is correlated with the original release date itself: recordings (originally) released recently are surely more likely to have an original release date than older recordings. As such, we would imagine this 1% figure *underestimates* the public domain proportion of commercially available recordings.

[35]For example, it is likely that the period and age of a work may influence price in and of itself.

[36]Strictly, recordings from the cutoff year are in copyright since recording copyright expires at the end of the calendar year in which the recording was made. However, by omitting the cutoff year allows for a cleaner division that reduces the likelihood that we have included PD items as in copyright or vice-versa.

shows that this difference is **highly significant (p-value of 0.010)** (p-value 0.04 for 1957 vs. 1959).[37]

1954-1958 v. 1960-1964						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
PD	1018	6.00974459725	1.95436972832	5.55	1.38	17.5
Not PD	1616	6.30943069307	2.05773463531	5.55	1.75	26.25
1953-1957 v. 1959-1963						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
PD	778	6.10015424165	2.40047879261	5.55	1.38	41.0
Not PD	1562	6.33829705506	2.26971426876	5.55	1.53	38.4

TABLE 7. 5 Year comparisons for pop music recordings

To test the robustness of these results we have performed a similar analysis with an extended sample period either side of the cutoff date. Extending the sample window obviously has the advantage of increasing the sample size. It also has the obvious downside of reducing the comparability of the two samples since the recordings being compared are, on average, further apart in time.[38] Results are show in Table 7. Here price differences are 30p (5%) (for 1953-1957 vs. 1959-1963 recordings it is 23p and 4%). This difference is again statistically significant with a p-value of 0.0002 – which is even lower than before despite the slightly smaller difference in means as a result of the substantially larger sample size.[39]

Pre-1959 v. Post-1959						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
PD	1226	6.16347471452	2.28046294735	5.55	1.38	34.99
Not PD	131144	7.032425883	2.46341819619	7.29	0.09	39.99

TABLE 8. Full comparisons of public domain and in copyright recordings

[37]One concern here could be the large number of recordings being omitted from the analysis due to a lack of release date. However, this omission should not bias the results unless recordings without a release are in some way systematically correlated with PD status and price (e.g. all high price PD recordings lacked a release date) which does not seem particularly likely.

[38]The basic point here is that unbiased estimates of PD vs. non-PD price difference requires that there are other factors affecting price which are correlated with PD status. Since it is likely that the age of recordings has some impact on price, one would like to keep the samples as close in age as possible.

[39]For 1953-1957 vs. 1959-1963 the p-value is 0.02 – also highly significant.

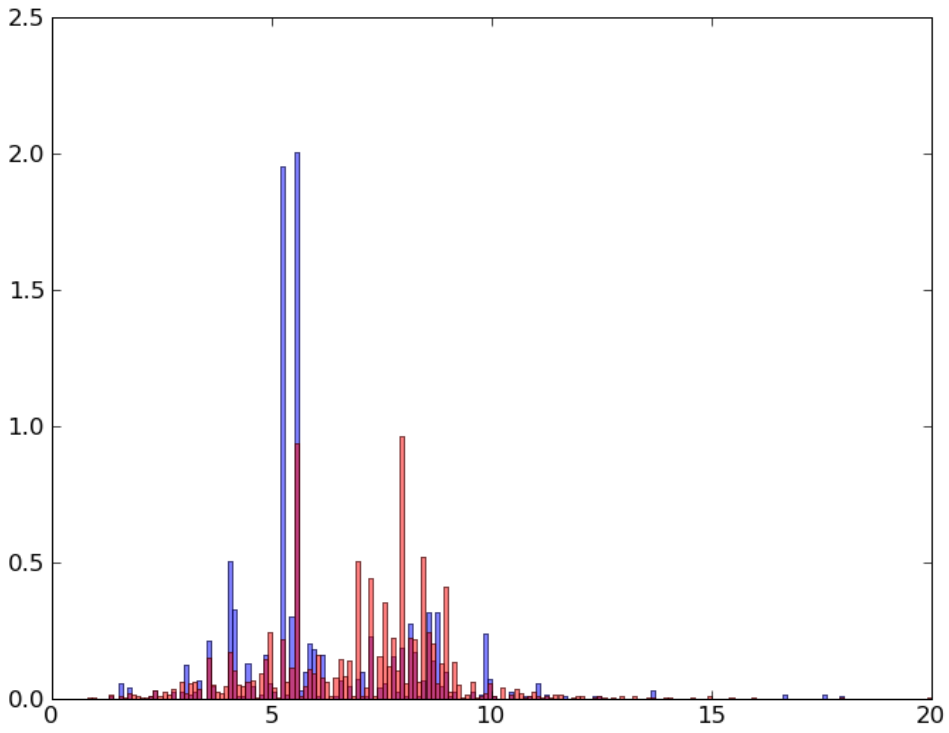

FIGURE 6. Price distribution for public domain (blue, originally released 1958 and before) and in-copyright CDs (red, originally released 1960 and after)

Lastly, Table 8 and Figure 6 show the price comparison for the full dataset.[40] Here, in-copyright recordings are on average 87p (14.1%) more expensive than public domain ones, a substantially greater difference than in the previous cases (t-stat of -12.0 with p-value $< e^{-16}$!). However, as just discussed, there are reasons to be cautious about PD versus in-copyright price comparisons the greater the discrepancy in average age of the two sets of recordings.[41]

[40]To reduce bias from outliers in the form of especially expensive items we have removed all CDs with a price above £40. This removed approximately 300 items (out of 300 thousand) in the in-copyright category and 1 in the PD category. Including these items increased the PD average price by 3p and the in-copyright average price by 11p.

[41]It is also important to note that keeping the sets of PD and in-copyright recordings proximate in vintage may introduce its own (downward) bias in our estimate of price differences. In particular, in-copyright work that is 'close' to the public domain may have a reduced price for two reasons (reduced, that is, relative to the counter-factual scenario where the work is identical except that is no longer 'close' to the public domain). First, copyright owners anticipating the entry of the work into the public domain may drop the price in order to extract as much of the residual demand as they can. Second, in-copyright material close to the public domain may be competing to some extent with existing (lower-priced) public domain material, which would be likely to exert a downwards pressure in price.

To sum up, using a large dataset on dealer-prices for popular music recordings we have found highly significant price differences between public domain and in-copyright recordings of similar ages.[42] In our results in-copyright recordings are, on average, between 4-14% (depending on comparison sets) more expensive than public domain recordings – corresponding to a price increase of 23-87p relative to the public domain price of approximately £6). These numbers are not insubstantial, and are in line with quote estimates of the 'copyright' proportion of a CD's cost – like books, CDs are a physical media and therefore the 'copyright' proportion of the total cost tends to be relatively low.[43]

6. Usage

We have anecdotal evidence from two record labels who publish music where copyright or neighboring rights have expired. Of these labels Ondine publishes mainly classical and VL-musiikki, the more interesting one, focuses on popular music. For example, VL-musiikki has published "greatest hits" CDs of Elvis Presley and Edith Piaf in August 2008. In both cases the neighboring rights of major labels have expired. In only two months, the Elvis album sold over 11,000 copies and Edith Piaf over 7000. With these figures the albums would most probably have been in the Finnish official top 10 annual charts if VL-musiikki were a member of IFPI (they do not wish to be).[44]

Petteri Laiho, the manager from VL-musiikki, explained that his company alone sold roughly 13% of the total volume of IFPI's official figures (8 million albums) and predicted that the real figure of CD sales in Finland could be 50-100% more than IFPI says. VL-musiikki is quickly growing the "neighboring rights business". In 2007, they made around 100,000 euros from the sales of "neighboring rights" albums, which was roughly 7% of their total sales. In 2008, they predict to increase total turnover 20-30% and increase the share of "neighboring rights" music towards 20% of total sales. Their strategy – if

[42]This result is in marked contrast to the finding of no price difference in the European Commission's impact assessment in relation to the proposed term extension for sounds recordings. That finding was largely based on a single piece of evidence: the PwC report for the Gowers Review of Intellectual Property prepared by PwC on behalf of the British Phonographic Industry (PwC, 2006). We would note that that report also found negative price impacts in several of their samples but given their limited size none of them were significant.

[43]We would also echo the points made above in relation to books that simple comparison of means may not lead to an accurate estimate of the true effect of copyright status on price, in particular that there is several reasons that price differences may be *underestimated*.

[44]In the first six months of 2008, the album in the 10th position in the official IFPI's list (which now includes both CDs and downloads) had sold a little over 16,000 copies and the album in the 20th position had sold a little over 12,000 copies.

the EU keeps the term of neighboring rights in 50 years – is to grow the "neighboring rights" business to 30-40% of the total turnover by 2012, which would mean 3-4 million euros more in turnover. They have planned to expand the business to Baltic states and other Nordic Countries but do not start sales abroad until EU makes a final decision on neighboring rights. Overall, Laiho stressed that his label is growing in the traditional CD business where the majors are struggling. He stated that with neighboring rights in existence, the major labels simply refuse to license the recordings to republishers, or set unacceptable terms and therefore that the expiry term has a significant, immediate and beneficial impact on their business.

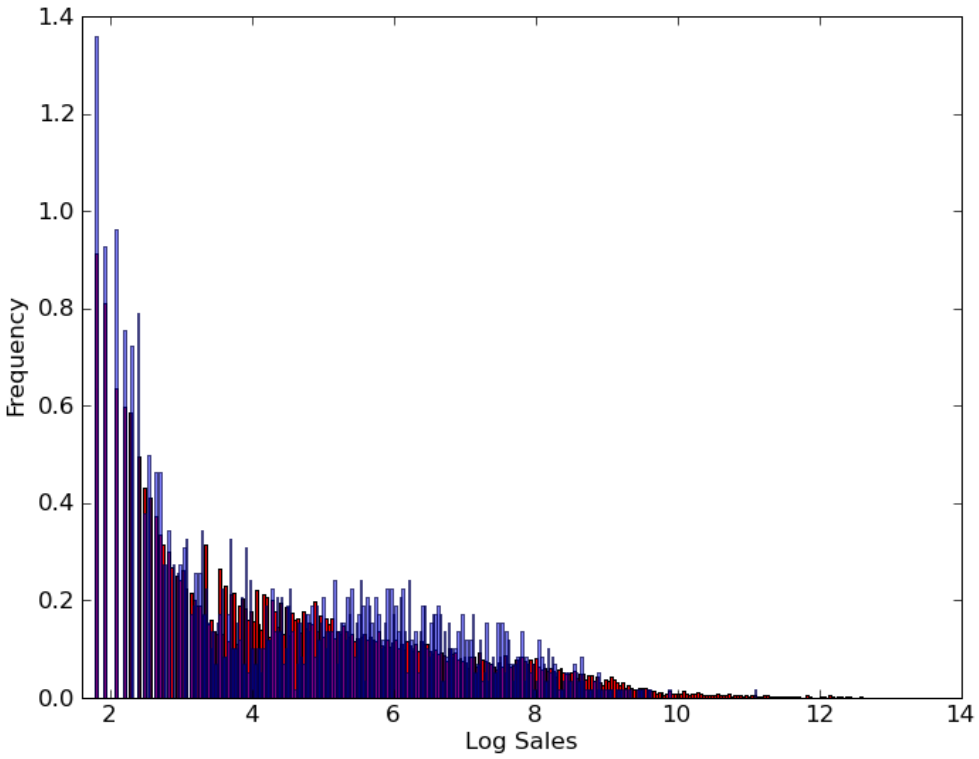

FIGURE 7. Distribution of paperback sales for PD and in copyright books

6.1. **Nielsen.** We also have detailed sales data from Nielsen including both units sold and revenue. Figure 7 and 8 show the distribution of sales (units) and revenues for PD and copyright items. Starting first with unit sales, the general shape in both cases is

approximately as one would expect:[45] (very crudely) negative exponential or power-law like,[46] with the greatest frequency (i.e. most products) is at the lowest sales levels with the number of items with a given level of sales declining sharply as the number of sales increase. To put this in concrete terms: there were over 12,000 items selling the minimum 5 units[47] but only 11 items with sales over 300,000 units, 4 over 500,000 and one (the top-selling item) with just over 800,000 units.

There are some noteworthy differences between the PD and copyright material. In particular, the PD distribution has a significant deviation from the smooth decay of the copyright distribution with a clear 'hump' in the middle region of the distribution with a peak at 6 (400 units).[48] Additionally, the PD distribution has both a greater density of low-sellers (selling less than 10 copies) and a smaller tail of high selling titles (the PD tail ends roughly 2 orders of magnitude below that for copyright at 10 (20,000 units) rather than 12 (160,000 units)). This is what we would expect: more low-sellers because PD works will have more low-selling 'specialist' and obscure titles, fewer massive-sellers both because sales of given work are split between several editions and because the work is old – which both reduces demand in and of itself, but, more importantly (at least in relation to the bestseller category) means that a PD book market is limited to those who do not already have a copy while new new release can sell to the entire population.[49]

Interestingly moving to revenues the picture changes quite significantly. Both PD and copyright distributions now have a short left-tail where densities drop off sharply from the peak around 4 (£54). This, we hazard, is a consequence of the fact that a small fraction of items have very low ASPs (zero in some cases!) and the low-selling portion of these then

[45]See e.g. Gaffeo et al. (2008).

[46]It is important to keep in mind that we have log sales on the x-axis so the decline (in non-log) terms is even greater than shown there. Simple examination of log-log and log plots of the sales distribution indicate that neither an exponential or a power-law fit particularly well: for the exponential the fit is reasonable up to about a 100,000 sales but (as is typical) fails very poorly to match the tail above that (though it is worth noting that this tail consists of not much more than a hundred or so items there being only 11 items with sales above 300,000); while for the power law a reasonable fit is only achieved over limited portions of the distribution e.g. in the 50-3000 unit sales range, and again (though independently) over the 8000-160,000 range.

[47]The dataset had been truncated at 5 units of sales.

[48]Interestingly this shape appears similar to that for the mechanical rights (compositions) usage in the Netherlands, see the data from BUMA/STEMRA below in Figure14.

[49]The obvious way to address this would be to match older in-copyright books with public domain material – an approach that would also be valuable in the price analysis. Unfortunately the limitations of time (and automated classifications) already discussed prevent pursuing this line of enquiry here. However, it is an obvious and useful piece of future research.

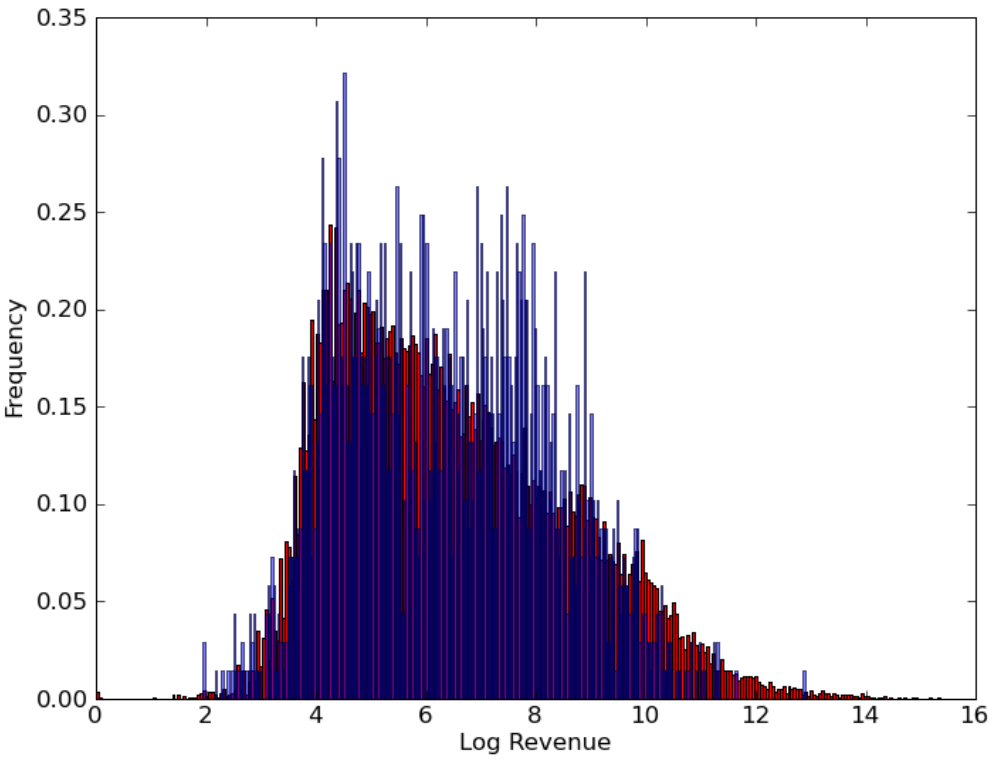

FIGURE 8. Distribution of paperback PD revenues for PD and in copyright books

make up the short left-tail visible in the revenue figure. More generally, revenue is the product of price times units sold so we would expect this figure to reflect the interaction of the ASP price distribution (Figure 3) and the unit sales distribution (previous Figure). Given the price distribution's skewed normal shape, a negative correlation between price and sales will result in the revenue distribution being more 'centralized' than the unit sales distribution.[50] A plot of sales against price is shown in Figure 9 and indeed clearly indicates a mild negative correlation.[51]

Thus we would expect to seem centralization – and indeed we do. There are also some noticeable differences between the copyright and PD distributions. The copyright distribution one at its peak around 4 (£50) falls steadily – almost linearly – until 12 (£160k) and then flattens out for the final run up to just over 15 (£4.6m) – again this

[50]Revenue = price x sales, so log(revenue) = log(price) + log(sales). If prices were all the same then the revenue distribution would just be the sales distribution shifted by the standard price. However, if they are negatively correlated then high prices will go with low sales and vice-versa and the revenue distribution will be more centralized.

[51]This is confirmed by a basic regression.

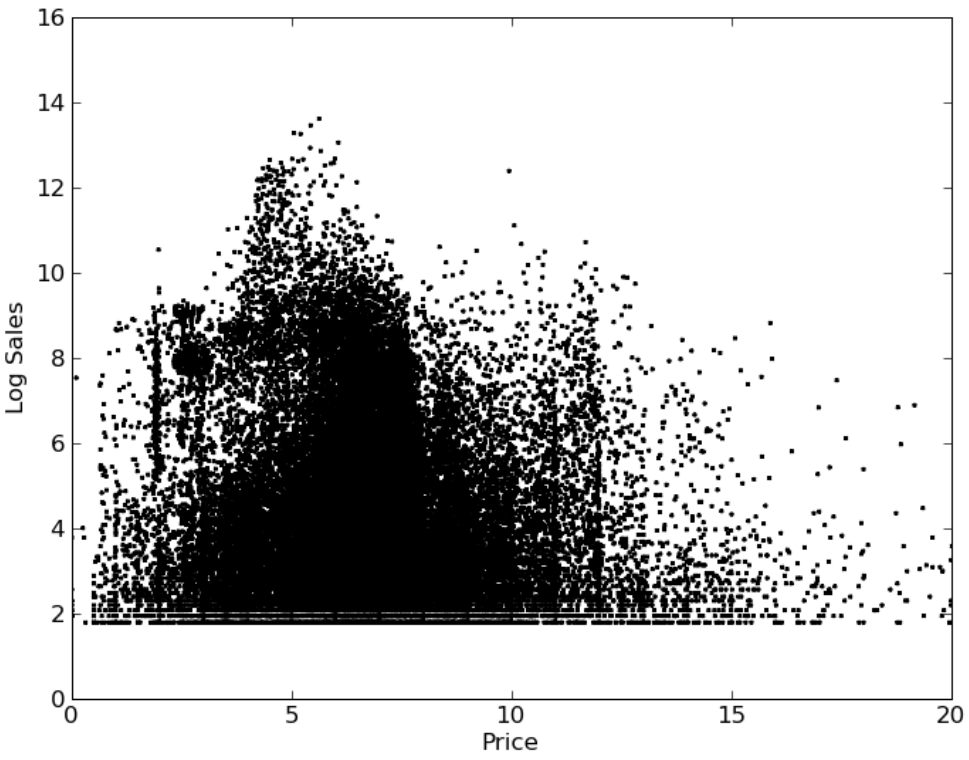

FIGURE 9. The Relationship of Price and Sales

upper tail is very small (there are under 50 items will sales over £1m – roughly 14 on the scale, 5 with sales over 2.5m, 2 with sales over 3m and just one with sales over 4m). The PD distribution, though having a somewhat similar trend, had a much greater variation over the central portion of distribution with densities often at a level close to the peak even around 9 (being more than 4 orders of magnitude greater than at the peak at 4). It also did not have much of a right tail – no doubt reflecting the same feature in the sales distribution – most best-selling titles have an average price and hence this right tail portion of the revenue distribution should follow that of unit sales very closely.

In summary, for both the sales and revenue distribution public domain material has a greater density in the middle of the distribution and a smaller tail of 'best-sellers'. Table 9 presents basic statistics for both the unit sales and revenue distributions. The two differences between PD and copyright material is apparent here: median sales (and revenues) of PD items are higher relative to copyrighted items but mean sales and revenues – which will heavily influenced by the right tail – are lower. Specifically mean sales and

Unit Sales						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
Copyright	39605	1612.21	12738.21	71.0	6.0	819426.0
PD	1873	627.55	2102.97	90.0	6.0	66785.0
All	41478	1567.75	12456.98	71.0	6.0	819426.0
Revenue						
	No. Items	**Mean**	**SD**	**Median**	**Min**	**Max**
Copyright	39605	9102.39	67797.67	458.25	0.0	4605174.12
PD	1873	3112.24	11788.17	556.32	6.0	397370.75
All	41478	8831.89	66308.25	461.5	0.0	4605174.12

TABLE 9. Summary Statistics for Paperback Sales Data

revenues for PD items are a little over a third of those for in copyright material. This means that **'pure' public domain works account for approximately 2% of total book sales (by volume)**.[52]

6.2. **Libraries.** In addition to asking libraries about their holdings we also asked for information on their usage (in the form of lending and the like). We only received information from a few libraries and detailed information from only one: the Slovakian National Library.

The Slovak National Library provided us with complete library loan data by item for 2007.[53] The results of analyzing this are shown in Figure 10 and Table 10. As these show usage is heavily skewed towards the present.[54] **Total usage from the pre-1920 period (the majority of such material being public domain) is 3,051 out of total of just over 62,000 loans for the entire period (5%).**

Trinity College Dublin also provided some interesting summary information. Usage of their books is detailed in Table 11. As this shows usage grows from 8,000 (6,000 withdrawals, 2,000 requests to view) for 1870-1880 material (almost all of which is public domain) to 300,000 for the 1950-1960 (almost none of which is public domain). Note that the number of items grows over the corresponding period from 15,000 to 500,000 so **usage per item is actually fairly flat (40% versus 60%) – in fact, for the 1880**

[52]A revenue comparison would not make much sense here given our earlier results on the price differences between PD and in-copyright works.
[53]Such data will, of course, not include usage within the library that did not result in a loan.
[54]The distribution in usage in fact looks remarkably similar to the general distribution of publications presented in our earlier section on the size of the public domain.

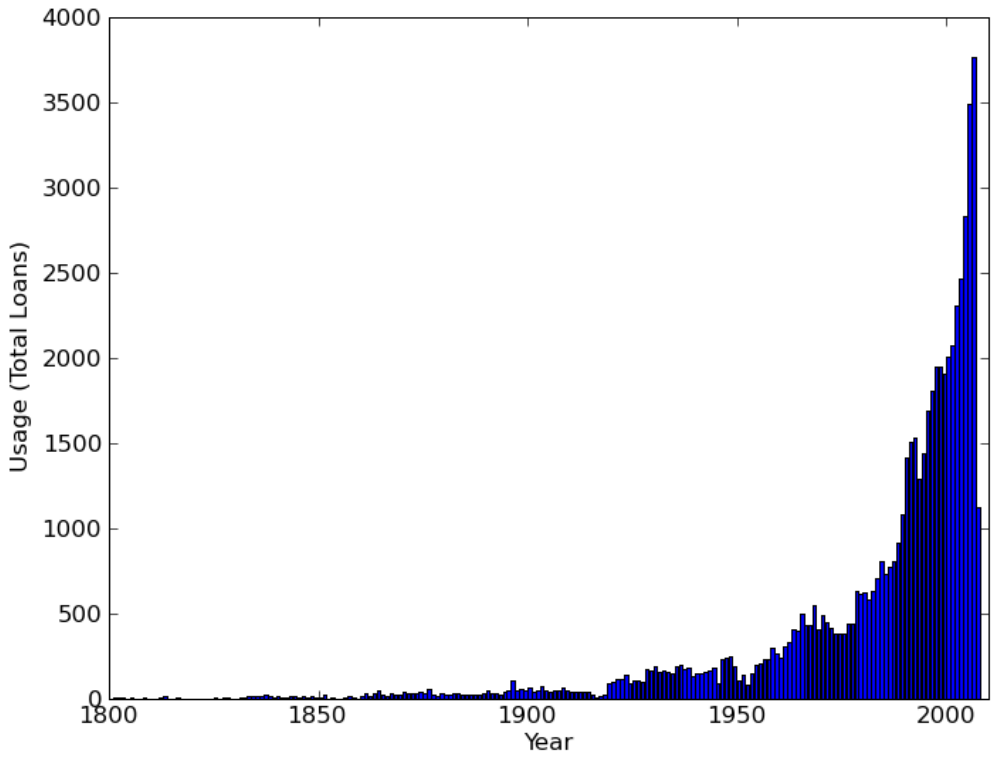

FIGURE 10. Usage (Loans) of Books from the Slovak National Library

Period	Total Usage
Pre-1900	2146
Pre-1910	2681
Pre-1920	3051
Pre-1930	4277
Pre-1940	5982
Pre-1950	7802
Pre-1960	9740
To Present	62663

TABLE 10. Usage (Loans) of Books from the Slovak National Library

and 1890s periods (also mainly public domain) usage per item is substantially above 1 and higher than any later period!

We also have information on photos. For this material they recorded: no usage prior to 1930; for 1930-1940: 1,000,000 euros in licence fees for use of WWII photo collection (used in 20th century history TV programme); for 1940-1950: distribution of 22,500,000 postcards generating 200,000 euros plus licence income for use of photographs in books of

30,000 euros; for 1950-1960: 300,000 posters plus 20,000,000 postcards to value of 1,200,000 euros. Of course by no means all of this material will be in the public domain. However it gives some indication of archive material, at least some of which is likely to be public domain.

Period	**Usage**
Before 1850	Not Available
1850-1870	Not Available
1870-1880	6,000 withdrawals and 2,000 requests to view
1880-1890	13,000 withdrawals and 20,000 requests to view
1890-1900	15,000 withdrawals and 50,000 requests to view
1900-1910	10,000 withdrawals and 23,000 requests to view
1910-1920	25,000 withdrawals and 20,000 requests to view
1920-1930	30,000 withdrawals and 75,000 requests to view
1930-1940	20,000 withdrawals and 120,000 requests to view
1940-1950	70,000 withdrawals and 100,000 requests to view
1950-1960	100,000 withdrawals and 200,000 requests to view

TABLE 11. Usage of Books from Library of Trinity College Dublin

6.3. **YLE (Finnish National Broadcaster).** Table 12 presents information from YLE (the Finnish National Broadcaster) on the usage of 'free' (PD) and 'non-free' (copyright) music on their various music and TV stations. As this shows, **for each of the 4 years, usage of PD material was around 10% of the level of copyrighted material**. This overall figure did mask some significant variation. For example, on one radio station (YLE Radio 1) PD material usage was higher than copyrighted material by almost 50% in each of the four years, while, by contrast, on YLE Maakuntaradio (Regional Radio) in 2007 there were only 4446 minutes of PD material but over 1.38 million minutes of copyrighted material (a ratio of 0.3%).

6.4. **Data from Società Italiana degli Autori ed Editori (SIAE).** We obtained several pieces of data from Italy the majority of which, and the most useful, coming from the Società Italiana degli Autori ed Editori (SIAE).

A particularly interesting dataset was available to us thanks to a curiosity of Italian copyright law. Until 1996, Italian Copyright Law (law decree n. 633/1941 repealed by law n. 30/1997) contained provisions establishing a regime of 'paying-public-domain'. Specifically, a fee had to be paid to the State for every performance or broadcast of a

Station/Channel	Type	2007	2006	2005	2004
Varsinainen TV1	Free (PD)	1590	2274	2859	2100
	Non-Free (C)	15390	21728	22250	25275
Varsinainen TV2	Free (PD)	1079	1340	1410	1936
	Non-Free (C)	34695	41609	37682	35875
TV-elokuvat TV1	Free (PD)	2666	3212	2336	2935
	Non-Free (C)	64837	48294	54101	44550
TV-elokuvat TV2	Free (PD)	1300	1002	924	1292
	Non-Free (C)	70672	50982	52232	48597
Yle Teema	Free (PD)	1592	1690	1777	0
	Non-Free (C)	11640	14216	9483	0
Yle FST	Free (PD)	695	486	447	0
	Non-Free (C)	10926	6338	6515	0
Yle Extra	Free (PD)	55	0	0	0
	Non-Free (C)	13260	0	0	0
Yle Teema elokuva	Free (PD)	4775	5768	1336	
	Non-Free (C)	54559	45540	21264	0
Yle FST elokuva	Free (PD)	932	783	587	242
	Non-Free (C)	23569	16309	16766	242
Yle Extra elokuva	Free (PD)	32	0	0	0
	Non-Free (C)	4119	0	0	0
YLE Radio 1	Free (PD)	108604	108485	107861	111317
	Non-Free (C)	78311	73306	76808	76249
YleX	Free (PD)	68	155	82	182
	Non-Free (C)	231655	231455	234478	237337
YLE Radio Suomi	Free (PD)	1060	857	467	595
	Non-Free (C)	100568	90650	86837	79283
YLE Radio Vega	Free (PD)	12463	12399	14170	14614
	Non-Free (C)	131338	124090	117083	115959
YLE Radio Extrem	Free (PD)	47	128	95	325
	Non-Free (C)	244192	233813	241525	236175
YLE Yöradio	Free (PD)	90856	92682	91156	92855
	Non-Free (C)	74601	71755	72137	89708
YLE Maakuntaradio	Free (PD)	4446	4053	5318	6848
	Non-Free (C)	1380640	1365931	1621024	1756843
YleQ	Free (PD)	0	753	1005	1138
	Non-Free (C)	0	157810	220529	217830
Total		232260	227340	227683	236137
		2544972	2593826	2890714	2963923

TABLE 12. Usage (in minutes) of PD and in-copyright music on YLE (Finnish National Broadcaster) Stations

public domain work (and elaborations thereof).[55] This fee varied across types of work,

[55]Article 175: For every performance or broadcast of a work suitable for public performance, or of a musical work, when such work is in the public domain for any reason any person who performs or broadcasts such work shall, in accordance with the rules contained in the Regulations, be required to pay to the State a domanial fee based upon the gross receipts or that part of the receipts which is proportionate to the part that the work occupies in the performance or broadcast, irrespective of the purpose of the performance or broadcast and of the country of origin of the work.
The amount of the nominal fee shall be determined by a Presidential Decree.
The amount of the domanial fee for separate portions of musical works or for short music compositions, shall be determined by the "S.I.A.E." in accordance with the provisions of the Regulations and upon the basis of the amount of payment normally required by the said body in respect of protected works performed under similar conditions.

being: 10% of total revenues for representations of musical works; approximately 5% for representations of theatrical works, operas, ballets and choreographic works; and 2.5% for representations of critical editions of public domain works. In the case of in-copyright works the percentage was 10% in all cases (the last category not existing for in-copyright works).

Thus, by comparing SIAE's record of revenues for the PD categories with that for all material (copyright and PD) it is possible to estimate usage. This is done in Table 13. Before analyzing these results, first note some minor points: a) total fees for musical works exclude revenue from foreign countries and fees from mechanical reproduction since these occurred only for copyrighted works b) we had no data on the proportion of critical editions and have made the conservative assumption there was no usage of critical editions (if, by contrast, we were assume a 100% critical editions them we would have to multiply implied PD usage for musical works by 4 and for 'other' types by 2); c) for reasons that were not entirely clear we do not have total revenues for the 'Other' category until 1974 (though we do have PD revenues) d) PD usage percentage for the 'Musical Works' category is exactly the same as the revenue percentage (and hence is not shown separately).

These minor points aside, the figures provide us with some really interesting information. Specifically, around **2% of performances or broadcasts of pure 'musical works' are of PD material**. Moreover, this is likely to be an underestimate as the usage of critical editions in this area is likely to be high as almost all public domain compositions will be of 'classical' music. If 50% of performances used such critical editions the correct figure would be around 5%. For the **'Other' category (the dominant item of which is almost certainly opera), the implied usage is significantly higher, ranging from a low of 6.4% to a high of 27% – over a quarter of all performances or broadcasts! – the average being just under 13%**.

We also obtained data on the number of public domain works being performed in Roman theatres. As ETI (Italian Theatre Institute) were unable to furnish us with any data in this regard, team members examined the forthcoming seasons of 20 of the top Roman

Article 176: The domanial fee shall also be payable in respect of public performances and broadcasts of protected elaborations of works in the public domain specified in the preceding Article. In such a case, and without prejudice to the rights of the author of the elaboration, the amount of the domanial fee shall be determined on the basis of one half of what would have been due if the performance or broadcast had had as its object the work in the public domain in its original form.

Year	Musical Works (Total)	PD	PD %	Other (Total)	PD	PD %	PD Usage %
1946							
1947	1196.9				12.2		
1948	1747.5				39.2		
1949	2263.3				72.7		
1950	2722.2				55.8		
1951	3074.2				171.2		
1952	3508.3				92.8		
1953	3884.7				98.2		
1954	4270.1				115.7		
1955	4654.4				107.1		
1956	4947.6				101.7		
1957	5041.2				119.8		
1958					183.6		
1959	6260.3				79.0		
1960	6985.3				134.6		
1961	7701.3				136.5		
1962	7908.8				130.9		
1963	8824.2				263.5		
1964	10326.9				306.5		
1965	10921.4				332.9		
1966	12691.8				422.6		
1967	13759.9	255.3	1.9		387.6		
1968	14868.8				408.2		
1969	16827.7				505.8		
1970	17441.7				581.1		
1971	19152.6	226.3	1.2		546.3		
1972	21059.2	256.7	1.2		753.5		
1973	23525.1	275.0	1.2		672.0		
1974	28083.0	359.7	1.3	11915.2	927.4	7.8	15.6
1975	32499.9	455.2	1.4	14469.1	1078.4	7.5	14.9
1976	35134.3	533.7	1.5	19009.2	1199.4	6.3	12.6
1977	48517.7	595.8	1.2	26126.5	1114.7	4.3	8.5
1978	51067.6	777.1	1.5	24427.7	1254.7	5.1	10.3
1979	61848.4	1083.9	1.8	32145.0	1621.2	5.0	10.1
1980	70107.1	1270.2	1.8	36869.9	2046.4	5.6	11.1
1981	88753.1	1494.3	1.7	40848.4	2643.9	6.5	12.9
1982	113007.0	1774.0	1.6	53741.8	3013.6	5.6	11.2
1983	124943.5	2250.4	1.8	57893.7	7800.7	13.5	26.9
1984	145279.9	2603.4	1.8	82224.2	6726.2	8.2	16.4
1985	167143.6	4741.7	2.8	85865.6	8470.6	9.9	19.7
1986	214325.3	2017.7	0.9	90201.3	7359.2	8.2	16.3
1987	227320.4	4770.0	2.1	98770.3	8997.0	9.1	18.2
1988	259876.2	4814.3	1.9	110891.5	4753.0	4.3	8.6
1989	286455.2			123170.8	5055.4	4.1	8.2
1990	302560.0	6150.7	2.0	143368.9	5439.5	3.8	7.6
1991	340065.6	7391.4	2.2	165923.4	5273.5	3.2	6.4
1992	341840.8	7270.0	2.1	175795.0	7088.6	4.0	8.1
1993		7868.7			7280.0		
1994	406267.1	8814.8	2.2		7603.5		
1995	416072.8	8663.2	2.1				
1996	433380.6	8719.6	2.0				
1997	414920.1	8861.6	2.1				

TABLE 13. SIAE revenues (millions of lira) and implied usage for performance and broadcast of PD and copyright works in Italy. 'Other' category is theatrical compositions, operas, ballets, choreographic works (for which PD fees were paid at the lower rate of 5%).

theatres. Of 273 anticipated productions, 12 were of public domain works (4%) while a further 33 (12%) were adaptations of public domain works. Thus in total 16% (45) productions utilized public domain works directly or indirectly.

6.5. **SCPP.** SCPP is a French collecting society for phonograms. They very kindly supplied with two sets of data on the usage (sales) of recordings. The reason they hold

usage/sales data is that they use this to apportion the monies they receive. As they explained to us: "We distribute private copying levies mainly on the basis of unit sales multiplied by duration. In order to do this, we ask our members to declare to us their unit sales (physical products and digital)."[56]

The first dataset contained total annual revenues from 1991 to 2006 for all phonograms released between 1945-1960 disaggregated by 'vintage' (i.e. year of release) but without any disaggregation by item (i.e. we only received totals all sales of a given vintage in a given year). The second set consisted of a complete breakdown of unit sales (by item and artist) for two vintages (1950 and 1954) for the sales period 1995-2005 (the neighbouring, 'recording', rights in these phonograms expiring in 1950 and 1954 respectively). We shall examine the first of these in this section and the second in the following section on the distribution of sales/usage.

6.5.1. *Annual Sales.* The first dataset containing annual sales revenue is shown in Table 14 with a summary in Figure 11.

[56]'Special' products – e.g. those sold to a company at a very low price to be given away as a gift – are not included in these figures. Similarly if a newspaper were to give away a set of free tracks with the paper these would not register in these sales statistics.

Année Vente	1945.0	1946.0	1947.0	1948.0	1949.0	1950.0	1951.0	1952.0	1953.0	1954.0	1955.0	1956.0	1957.0	1958.0	1959.0	1960.0
1991	512.9	550.8	528.5	712.6	380.6	422.1	1125.0	2601.8	2978.9	2269.2	2355.6	2809.2	2036.1	4160.1	2612.2	5020.3
1992	389.5	468.2	725.4	774.6	540.3	607.3	990.9	1770.0	2548.2	2019.2	1903.6	1960.8	1705.9	2958.0	2019.8	6653.4
1993	272.0	1183.5	1038.7	745.0	1211.1	1063.8	1859.7	2038.3	3529.1	2591.6	3522.4	3082.2	2605.1	4384.4	3412.3	4807.1
1994	195.8	732.2	592.2	516.1	617.5	695.5	1029.2	1769.0	2483.3	2606.3	3175.6	2772.1	2777.0	3889.6	2897.5	3811.1
1995	187.7	579.0	706.8	645.2	851.0	577.8	1116.9	1860.9	2549.8	2341.0	2515.3	2402.6	2467.0	3913.7	2641.4	4472.6
1996		573.7	906.4	568.5	706.7	771.0	1162.0	2213.7	3333.7	3460.4	3800.1	2921.2	2424.7	4468.1	3375.2	6878.7
1997	31.2	66.4	861.9	594.4	690.0	749.0	976.6	1353.9	2796.5	3011.2	3463.2	2363.0	2786.2	3816.2	3233.5	5919.8
1998		20.8	7.2	402.7	491.8	770.9	878.0	1292.1	2499.8	2572.9	3018.4	1959.7	2566.9	3932.4	3521.9	5447.2
1999	337.9	508.7	563.6	502.8	651.4	1109.8	908.4	2115.6	3247.3	3740.5	3324.0	4521.9	4015.3	5497.2	4754.2	6301.0
2000	590.2	1307.6	1689.8	749.0	1198.7	1507.7	1197.4	1782.1	3075.1	4082.9	3951.2	4163.3	4154.2	5748.1	5375.9	6570.7
2001	687.9	1348.5	1830.7	792.5	1204.1	1630.7	1402.8	3514.3	4497.3	5046.8	3893.0	5411.7	5140.5	7314.2	5649.6	8136.0
2002	756.8	663.4	1097.6	1772.4	2080.0	1096.6	1103.1	2139.0	3128.6	3692.2	3927.9	4225.3	3814.5	5293.1	4731.5	6454.1
2003	697.0	743.0	1518.8	1372.8	1600.4	1447.5	2330.3	2335.8	2922.9	2887.8	3770.5	3527.7	2860.6	4590.9	4688.2	5076.2
2004	9.0	9.5	25.5	20.1	18.3	9.1	12.0	37.4	64.2	2261.6	2328.9	2892.4	2342.1	3163.9	3227.7	4849.8
2005			2.8	0.4	0.2	12.8	1.3	3.1	1.5	185.3	2384.6	2760.7	2294.7	2841.5	2964.6	3773.7
2006			0.1									3477.4	3553.2	4627.4	4087.2	4714.8

TABLE 14. Sales revenue (thousands) for 1945-1960 releases in recent years from SCPP data

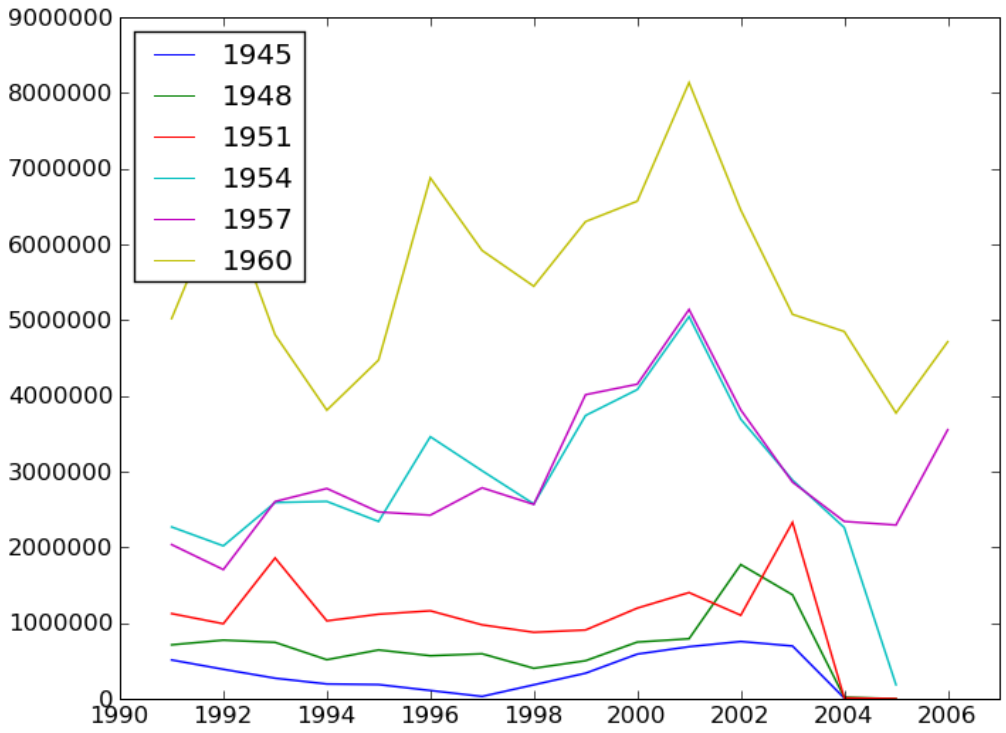

FIGURE 11. SCPP total sales by release year. only 5 years are shown to avoid 'cluttering' the image.

The alert reader should already be wondering about the data on public domain recordings (e.g. post 1995 for 1945 recordings, 1996 for 1946 recordings etc). After all, once a recording[57] enters the public domain SCPP should no longer be paying out (or receiving) revenues in relation to that recording. It is certainly noticeable in the dataset that revenues fall away precipitously around 2004 for all vintages which are already PD by that point. This sharp drop-off may be the result of issuers of PD material no longer registering it with SCPP and/or paying fees.[58] Nevertheless it is still surprising that there are any revenues post PD-entry and it is unclear why the sudden drop-off happens in 2004 across a whole range of vintages.[59] We did raise this issue but we were able to resolve the question

[57]Strictly here we mean the recording right in the recording as the right in the composition may still be operative.

[58]Note also that this dataset contains revenues not unit sales which is likely to intensify any under-counting since sales of PD recordings (even when registered) are like to generate less revenue per item than sales of in copyright recordings.

[59]These drops are so sizeable that they cannot explained by simple variation. For example 1945 revenues go from nearly 700k in 2003 to 9k in 2004 and then zero (unrecorded thereafter).

completely.[60] As a result it is necessary to assume that all revenue figures for a given vintage post-entry into the PD may be unreliable – and hence underestimates.[61]

This is rather unfortunate as it renders the comparison of usage pre and post public domain of dubious validity.[62] That said, since we can assume post PD entry is under recorded any results where PD entry does have a positive effect are, a fortiori, still significant. However, even a brief glance at the data would indicate that those huge drops for post-PD vintages around 2004 would likely nullify any analysis.

Year	Sales of all Phono	Sales 1945-1960	Sales PD	PD %
1991	831.14	31.08	0	0.0
1992	798.76	28.04	0	0.0
1993	889.9	37.35	0	0.0
1994	1015.38	30.56	0	0.0
1995	1070.87	29.83	0	0.0
1996	1174.26	37.56	0	0.0
1997	1193.87	32.71	0.1	0.01
1998	1160.85	29.38	0.03	0.0
1999	1621.75	42.1	1.91	0.12
2000	1634.23	47.14	5.54	0.34
2001	1836.02	57.5	7.49	0.41
2002	1822.1	45.98	8.57	0.47
2003	1479.47	42.37	12.05	0.81
2004	1251.59	21.27	0.2	0.02
2005	1208.63	17.23	0.21	0.02
2006	1300.64	20.46	0.0	0.0

TABLE 15. Revenues (millions) for phonograms in France 1991-2006

Given these difficulties we have refrained here from conducting a formal impact analysis. However, the data does allow us to look at PD usage relative to non-PD material – see

[60]SCPP explained to us that: "They [record companies] do this [declare their sales to us] for all of their products, whether they are in public domain or not. Also, on a compilation, some tracks can be in the public domain, some can be still protected. So, normally, the sales data that we have includes protected tracks as well as public domain tracks." However, there are PD specialists (who only release PD material) and many PD releases may consist only of PD tracks. It is therefore not clear whether PD sales are fully declared to SCPP. We would also note that 'free' distribution (e.g. with a newspaper) is not included in the SCPP figures and this is an area where one would expect PD to be relatively more heavily used since no payments need be made (at least for the recording right).

[61]There are other slight oddities: for example, the absence of any 1996 or 1998 data for the 1945 vintage, or the dramatic drop for the 1947 vintage in 1998.

[62]And additional problem is what the comparison period should be. In our discussions with SCPP one thing they mentioned was that "Most record companies release compilations of their back catalogue the year before some tracks enter the public domain." If so then pre versus post comparison will be upwards biased (and estimated 'PD effect' will be downwards biased). One solution would be to extend the comparison periods further back either side of the PD transition. However, this creates other problems: sales are clearly affected by many variables other than PD status and as the comparison periods become less proximate the more difficult it comes to make an assumption that 'all other things are equal'.

Table 15. This shows annual revenues for all phonograms, phonograms from 1945-1960 and PD phonograms for each year 1991-2006. Based on these figures PD phonograms account for a very small fraction of overall usage 0.5% or less in all cases. However, as just discussed, we are concerned that the estimates for PD revenues may be significantly underestimated – a view reinforced by this table (the number of public domain recordings is increasing from 2001 to 2006 yet revenues fall from over 12 *million* in 2003 to 200 *thousand* in 2005 and just 123 (!) in 2006).

6.5.2. *Availability.* Our second SCPP dataset gives unit sales (not revenues) for two 'vintages' (1950 and 1954) from 1991-2005 *disaggregated* by item (track) – i.e. with have full sales information down to the item/track level. The 1954 data is limited by the fact that we only have one year of PD data (and, as already discussed above, data for PD material post 2004 seems especially problematic). Thus we shall focus on the 1950 data.

We have already discussed the overall impact of the public domain on sales in the previous section so here we focus on slightly different matters. First, the disaggregated dataset allows us to look at the *availability* of recordings. We approximate availability of a recording by it having at least one sale in a given period (if a track has no sale within the entire 1991-2005 period it would not appear in our dataset at all).[63] We compared 4 year periods pre and post PD (1996-1999 and 2001-2004). Over the entire period there were 2022 tracks were 'available' (had at least one sale). In the pre-PD period 1098 were 'available' (54%) while in the post-PD period 1680 were 'available' (83%). Of course we must be cautious in interpreting such results. Other significant changes, especially regarding digital technology, were occurring over the period examined. Nevertheless the increase is substantial and is certainly suggestive.

7. Distribution of Sales/Usage

A variety of our datasets allow us to look at the distribution of sales, that is, how many items[64] sell one copy, how many sell two, etc. We have already seen some data on this area

[63]Obviously it is not a perfect proxy but it seems reasonable for our purposes. We'd like to thank SCPP for suggesting this approach.

[64]Again we have data on 'items' not 'works' though we would note that it is likely for the majority of items, especially those selling few copies, that the work and item can taken as identical since the work will never be issued more than once.

in the section on the Nielsen dataset above. Here we present some of the other datasets that shed light on this area.

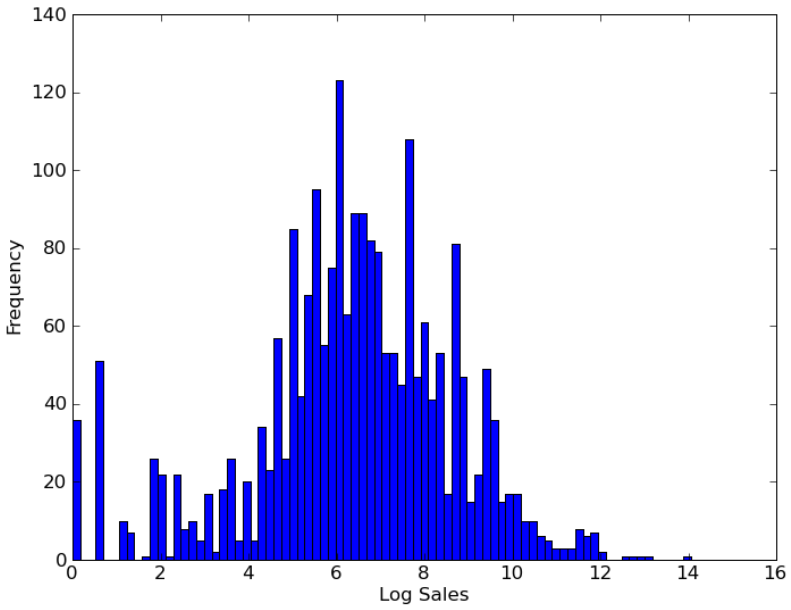

FIGURE 12. Distribution of Sales of 1950 Recordings 1996-2003 in France.

7.1. **SCPP.** The second SCPP dataset, as already discussed is disaggregated by item and provides a good dataset with which to look at the distribution of sales across items. Figure 12 shows the distribution for the 1950 vintage for all sales between 1996 and 2003 (we have aggregated across several periods to provide some smoothing as there are large year to year fluctuations in sales of individual items). As can be seen from the histogram the sales distribution is (very) approximately log-normal.

7.2. **BUMA/STEMRA.** We received data from BUMA/STEMRA, a dutch collecting society for phonograms and compositions. The primary datasets consisted of a random selection from their full catalogue listing the names of tracks/compositions together with usage (number of times and duration) for 2005-2007. Unfortunately the dataset gave no information on a track other than its name, in particular we had no information on the year of release which means we were unable to make any public domain/in copyright distinction. Nevertheless, the dataset does allow us to compute the distribution of usage across works. We have different types of usage information available here. For performing

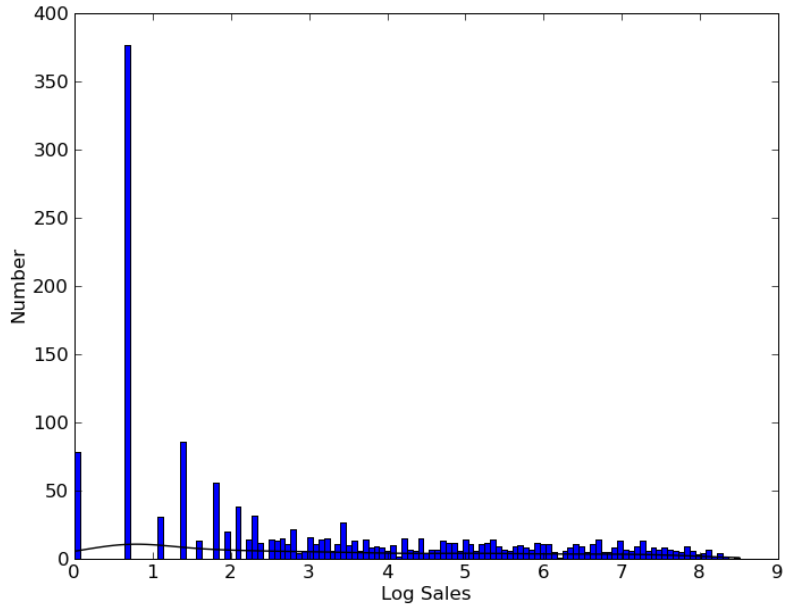

FIGURE 13. Distribution of performing rights usage from BUMA/STEMRA data

rights we have number of times used (and duration), for mechanicals we have number of times used and total external payment. The distribution based for these three different figures are shown in Figure 13 and 14.

7.3. **Literar-Mechana (LIME) Austria.** We received data from Literar-Mechana, an Austrian collecting society which administers the rights of authors, journalists, scientists and translators as well as the later owner of the rights. The Literar-Mechana accounts for: mechanical rights, broadcasting in semi-public places (bars, hotels, ...), levy on blank tapes and other carriers for readings of literary works, mechanical copying of licenses of video and CDs onto AV-carrier, reprography, of literature for private use, copying of works in school books, library lending, public readings and broadcasting rights (except for Austria). The dataset was an anonymized listing of disbursements to members in particular years. The distribution for authors for 1994 is shown Figure 15 and appears, very approximately to be a right-skewed lognormal, a little similar to the distribution for SCPP and BUMA/STEMRA (payments) distributions.

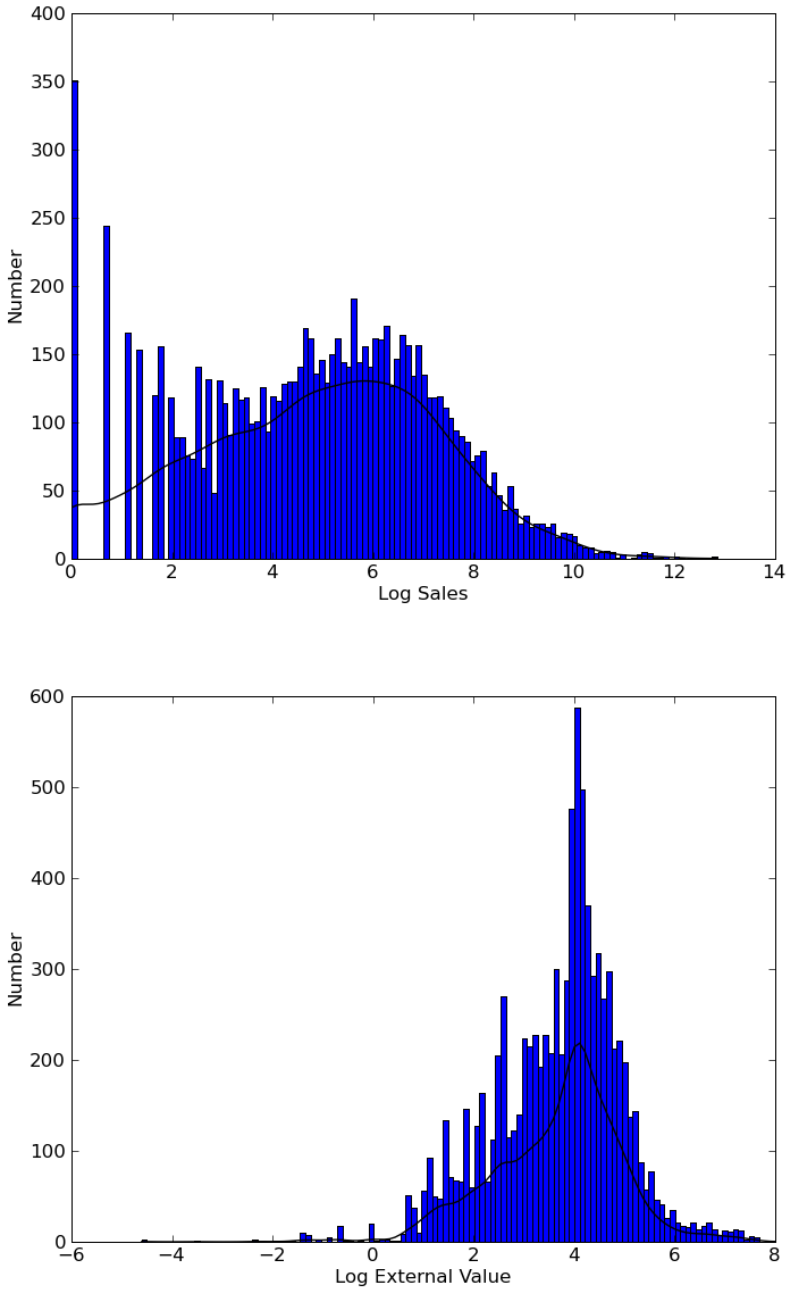

FIGURE 14. Distribution of mechanical rights usage (top) and payments (bottom) from BUMA/STEMRA data

8. DIGITIZATION

One striking omission from the preceding sections is any significant discussion of digitization and 'digital' material. The digital aspect is especially relevant to the public domain for two reasons. First, the reproduction cost, i.e. the marginal cost of producing new

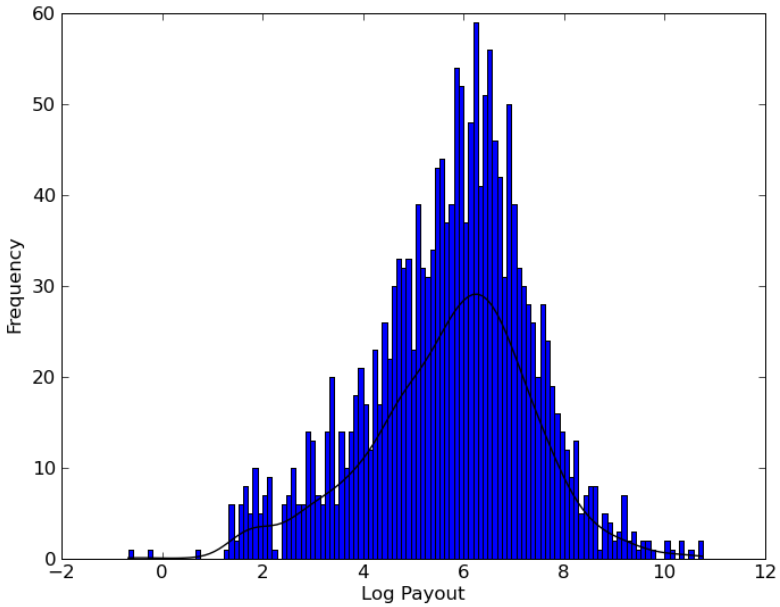

FIGURE 15. Distribution of LIME Disbursements to Authors in 1994

copies, is approximately zero. This has strong implications for the price of public domain material in digtial form – in general we would expect price to be at or close to zero. Second, issues related to IP rights bulk large when considering the digitization of non-digital material such as the content of existing libraries and archives. For public domain material these rights issues are absent while for in copyright material they can be so significant as to prevent, or at least substantially hold-up, digitization.[65]

Taken together these suggest that the 'value' of the public domain will be significantly higher in a digital world because a) the price difference – or, at least, the *percentage* price difference – between public domain and non-public domain works will increase (if digital public domain material is avaiable at 0 then the price difference must be 100%) b) as the technological costs of digitizing and making available material drop (as they have been doing) the 'transaction costs' associated with clearing rights – which are absent for public domain material – become a relatively larger issue.

[65]For evidence on this socre one need only look at the various existing digitization efforts such as that being carried out or considered by Google Books, the Open Content Alliance, Europeana etc.

These are obviously important considerations, and likely to grow more so.[66] It is therefore important to seek existing evidence on their impact, evidence which will also allow us to make some tentative extrapolations into the future.

Unfortunately, we do not have available here any comprehensive datasets of the kind discussed above for traditional media. However, we do have a significant amount of evidence, anecdotal and otherwise that take together give a clear picture of what digitization will mean.

We start with price. Here the evidence from several different sources strongly indicates that digital public domain material will be available at zero price. For example, Project Gutenberg offers all of its more than 26,000 public domain ebooks available for download at zero cost. Similarly, archive.org offers more than 1 million scanned public domain books for download at no cost, as well as providing large amounts of other public domain (and open) material for free access.

Good statistics on usage have been hard to find but Gutenberg provides data on downloads over the last month. These show that in November 2009 from their own servers they saw more than 2.4m books downloaded.

We also did a partial comparison of prices of ebooks on Amazon. Interestingly for every public domain title there was at least one version available at 0 price – as well, of course as editions available at a non-zero price.

9. Conclusion

9.1. **Price Impact.** Our estimates based on Nielsen data suggest the average retail price impact for books to be around 5-15%, though we would emphasize that, as discussed above, such average figures can be misleading and may well underestimate the overall impact of the public domain on access. Moreover, it should also be remembered that for pure digital products such as ebooks, digital downloads etc, the impact is likely to be much higher than for physical products where the bulk of the price is accounted for by manufacturing and distribution costs. Unfortunately we did not have the data or the resources to investigate these types of markets though a brief examination of public domain ebooks on Amazon

[66] In fact, given existing trends, it is likely that soon digital will be (by far) the most important area. However, at the present time, even in music which is the most easily digitizable 'physical' media still account for the majority of authorised sales (of course, unauthorised copying via file-sharing networks is entirely digital and on some estimates already rivals or exceeds authorised usage).

indicated that most well-known PD works were available at a price of 0 (often sourced it seems from Project Gutenberg) which was not the case for in copyright works.

9.2. **Usage.** We had data on usage of public domain (and copyrighted) material from a very wide variety of areas. A summary is provided in Table 16.

Area	**Source**	**PD (%)**	**Comments**
Paperback Book Editions	Nielsen	4.4-6.8	For 'pure' PD (for 'broad' PD 6.8). Including unclassified formats raised figure to 6 (11 for broad)
Paperback Book Sales	Nielsen	2	Only included pure PD items (would be 50% to 100% higher with 'broad' PD
Book Borrowing	Slovak National Library	5	
Book Borrowing	Trinity College Dublin Library	3-8	24% relative to pre-1960 books (total to present likely to 4-8 times as much)
Music on Radio and TV	YLE (Finland)	9	
Musical works	SIAE (Italy)	2	Likely to be a partial underestimate (figures up to 8% possible)
Other works (esp. Opera)	SIAE (Italy)	6-27	
Recordings Revenue	SCPP (France)	<1%	Likely that public domain revenues are significantly underestimated

TABLE 16. Summary of Usage Data. PD column gives usage of public domain material as percentage of total usage.

As this shows, the proportion ranged quite substantially from a low of under one percent to a high of over 25%. However, it is likely that the extreme values at either end are not a good general guide – the lowest estimate comes from SCPP data which as discussed above is likely to be a significant underestimate, while the highest value is for a rather special case (opera in Italy).

Leaving these extremes to one side, across a variety of jurisdictions and media, usage runs at 2-10%. Interestingly this is, very approximately, the same level as our estimates of the size of the public domain, suggesting that average usage per work may actually be the same between PD and in-copyright material.

9.3. **Distribution of Sales Across Works.** Here our preference would have been to estimate distributions for disaggregated data, especially by cohort/vintage – i.e. a set of works produced at approximately the same time – and by public domain status.

However, limitations of data has made this impossible though we did estimate: a) separate distributions for PD and in-copyright books in the UK (Nielsen data); b) the distribution for a given vintage (1950) of recordings in France (SCPP); c) the distribution for all extant recordings in the Netherlands (BUMA/STEMRA).

Simple examination of the various distribution indicated some significant variation with no obvious pattern emerging.[67] What was clear in all cases was that there was very significant variation in the usage of works with the bulk of works – copyright or PD – having little usage and a few being 'best-sellers'.

9.4. **The Value of the Public Domain.** Taken at face value, our results based on the Nielsen data – the only dataset on price-impact we have – suggest that, for books in the UK, public domain items are cheaper relative to in-copyright ones by around 5-15%. We must reiterate the concerns expressed previously that this 'average' figure could be a significant underestimate of the overall impact of the public domain on price. Nevertheless, proceeding on the basis of this figure, with total sales of PD paperbacks in our sample at £5.8m (1.2m units), this would generate a consumer gain of around £580k.

Of course, this is not the value of the public domain since that is solely the reduction in the deadweight loss (that being the gain in welfare to society). To calculate this requires an estimate of the *impact* of public domain on usage. This is something which it has not been possible to estimate given our available data. Instead, one option is to make an assumption about the elasticity of demand and use that to estimate the size of the deadweight loss triangle. Assuming an elasticity of 2-4 implies a deadweight loss of £58k-116k (1-2% of PD revenues).[68]

These figures are obviously small, reflecting the small price-effect which is, in turn largely a reflection of the fact that the bulk of a book's price are accounted for by production

[67]Only the distributions would be comparable – actual levels of usage would not be given the different activities being measured (retail sales, collecting society usage etc).

[68]These figures for the elasticity seem a reasonable range based on the existing literature, see e.g. on books Bittlingmayer (1992); Ringstad and Løyland (2006), on music, e.g. Peter J. Alexander *Music Recording* in James Brock, Ed. the Structure of American Industry (2005), p.127, Gasser et al. *Content and Control: Assessing the Impact of Policy Choices on Potential Online Business Models in the Music and Film Industries.*

and distribution costs – not the 'cost of copyright'.[69] We should also reiterate here that this kind of estimate is very sensitive to the incidence of price changes across works – for example if popular works see particularly large price reductions this simple approach will be an underestimate.

Being bold, we can attempt to extrapolate this figure to the full set of public domain works. Our results suggest a price-impact 5-15% (though, as noted above, this may be an underestimate). Taking a point estimate of 10% and combining it with a PD usage ratio of 10% and an elasticity of 2-4 this yields a public domain value equal to 1-2% of current PD revenues and 0.1-0.2% of total revenues (i.e. revenues for all cultural works both public domain and in copyright).

Again this may not seem like a large number. However, the reader must remember that 'value' is a surplus measure and relative to revenues is always likely to be relatively small because costs are substantial. This same logic would apply when estimating the 'value' of copyright. For example, with linear demand the *surplus* generated by a copyrighted work is equal to the 'value' of the public domain (itself being the deadweight loss).[70] In this case then the 'value of copyright' would also be 1-2% of current PD revenues and therefore 'small'. This just illustrates the large gap between 'value' (taken to be welfare) and revenues.[71]

Thus, in deciding whether this figure for the value of the public domain is large of not we really want to compare it with some other measure of surplus, the most obvious being the 'value of copyright' – this being the additional welfare generated by the existence of copyright compared to its absence (or compared to it being at some very low level – say one year in length). Unfortunately estimating this value is, if anything, even more fraught with difficulty than our endeavours here (how does the supply of creative work vary with anticipated profits?). We must therefore remain satisfied with the figures we have. For, imperfect and imprecise as they may be, they are the best we can currently obtain.

[69]This in turn is why any estimate of the 'value of copyright' certainly *cannot* be based on revenue figures and the like.

[70]The copyright acts as a monopoly and hence with linear demand the markup over cost will be such that the deadweight loss and consumer surplus triangles are the same. Assuming all the monopoly surplus goes to pay for the creator's efforts then surplus is just consumer surplus – if not the surplus will between one and three times the deadweight loss.

[71]And shows why revenue derived estimates of the 'value', be it of copyright or the public domain are of dubious validity.

Appendix A. Determining Public Domain Status for Nielsen Data

A.1. **Matching Against Open Library.** We ran a matching algorithm[72] against the Open Library database via their web api.[73] This was a very slow process. First, because web apis are relatively slow and second because, perhaps due to overloading, the OL API would stop responding at some point and a manual reboot would be required (to try avoid overloading the API we'd already added a significant delay between requests – another reason the process was quite slow). Overall it took more around 10 days to run through the whole 64k item dataset. The results were as follows:

```
Total Items: 63917
Total PD: 2206.0
Percent PD: 3.450
Percent Matched: 58.847
```

As this shows matching was not that successful with only around 3/5 of items successfully matched. The reasons for this relatively poor match rate is likely due to two factors. First, in order to keep algorithm run-time within reasonable bounds, we had to limit the number of title matches to 10 in the first step of the algorithm. For many titles there were more than 10 matches and thus we may discarding the correct match when we truncate to the first 10 results. At the same time the lack of standardization in titles and author names[74] mean that matching between the two datasets based on these attributes can still fail even when the same item is present in both – and in making the algorithm 'fuzzier' to cope with this one inevitably degreeds performance both in terms of speed and false positives. Furthermore, and ironically, the Open Library data often suffered from the very problems as the original (Nielsen) dataset: authors were not consolidated and authorial death dates were largely absent.[75]

[72]See 'OLQuery.by_work' method in `http://knowledgeforge.net/pdw/hg/file/tip/pdw/getdata/openlibrary.py`.

[73]We initially thought this would be straightforward – and fast – because we could match on ISBN which was present in both datasets. Unfortunately, Open Library is US oriented and it turned out that the 'same' book issued in the US and the UK will often receive different ISBNs with the result that matching on ISBN yielded very few matches.

[74]To give two very simple examples: titles in the Nielsen dataset had been normalized to move 'The' and 'A' to the end of the title (e.g. 'Christmas Carol, A') while the Open Library dataset did not. Similarly for names Nielsen had adopted Last, First but Open Library had adopted First Last.

[75]This is perhaps not surprising as it appears the Open Library have scraped a substantial amount of their data from Amazon which has no reason to record this kind of information.

Overall, approximately 3.5% of all items were identified as PD (that being 5.8% of those actually matched). The PD determination algorithm was a conservative one with an item labelled as PD only if all authors were positively identified as PD.

Thus, this figure is likely to be a lower bound (at least assuming the match process was reasonable – and allowing for the fact that some PD items included non-PD material such as commentaries). It was certainly clear from basic eyeballing that a substantial number of PD works were either not matched or not computed as PD (because of incorrect authors or missing death dates).

A.2. **Matching Against NGCOBA.** The New General Catalogue of Books and Authors (NGCOBA) is a database created by Philip Harper.[76] It is specifically oriented to determining PD and copyright status and therefore has excellent death date information. It is author oriented but also records work information (title with date of first publication). We ran the following algorithm using Philip Harper's NGCOBA database:[77]

(1) Matched by title and authors.
    - If match: compute PD status strictly (all death dates known and all less than 1937)
    - Else: continue
(2) Pick first author and find all (approx) matching authors (allow extra first names)
    - If no match: Not PD
    - Initialize PD score to 0
    - For each matched author alter score in following manner:
    - If author PD: +1
    - If not PD: -3
    - If unknown (no death date) -0.5
    - PD if score < 0 (Else: Not PD)

This algorithm took a few hours to run (this could likely be improved with a some DB optimization). The results were:

```
Total Items: 63917
```

[76]It is available online at `http://www.kingkong.demon.co.uk/ngcoba/ngcoba.htm`.
[77]To do this required downloading, parsing, normalizing and then loading the NGCOBA data into a database. This was a significant operation as the NGCOBA data format is a non-standardized and activities such as name and date normalization are a non-trivial endeavour.

```
Total PD: 6404.0

Percent PD: 10.019
```

As can be seen the fraction PD here was substantially higher at around 10%.

A.3. **Determination by Hand.** Thanks to very generous assistance of Mr Harper we were also able to classify the entire dataset by hand (using the NGCOBA and our own knowledge). The results of this are shown in the main text.

References

Bittlingmayer, G. (1992). The Elasticity of Demand for Books, Resale Price Maintenance and the Lerner Index. *Journal of Institutional and Theoretical Economics*, 148:588–588.

Brooks, T. (2005). Survey of Reissues of US Recordings. Copublished by the Council on Library and Information Resources and the Library of Congress.

Clerides, S. (2002). Book value: intertemporal pricing and quality discrimination in the US market for books. *International Journal of Industrial Organization*, 20(10):1385–1408.

Gaffeo, E., Scorcu, A. E., and Vici, L. (2008). Demand distribution dynamics in creative industries: The market for books in Italy. *Information Economics and Policy*, 20(3):257–268.

Ghose, A., Smith, M., and Telang, R. (2004). Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact.

Heald, P. (2006). Copyright Ownership and Efficient Exploitation: An Empirical Study of American Works Before and After They Enter the Public Domain.

Hendricks, K. and Sorensen, A. (2009). Information and the Skewness of Music Sales. *Journal of Political Economy*, 117(2):324–369.

Hui, K.-L. and Png, I. P. L. (2002). On the Supply of Creative Work: Evidence from the Movies. *American Economic Review*, 92(2):217–220.

Landes, W. and Posner, R. (1989). An Economic Analysis of Copyright Law. *Journal of Legal Studies*, 18(2):325–363.

Le Guel, F. and Rochelandet, F. (2005). P2P Music-Sharing Networks: Why the Legal Fight Against Copiers May be Inefficient? *Review of Economic Research on Copyright Issues*, (2):69–82.

Liebowitz, S. J. (2008). Is the Copyright Monopoly a Best-Selling Fiction?

Png, I. and hong Wang, Q. (2007). Copyright Duration and the Supply of Creative Work. Levine's Working Paper Archive 321307000000000478, UCLA Department of Economics.

Pollock, R. and Stepan, P. (2009). The Size of the EU Public Domain. Technical report.

Pollock, R., Stepan, P., and Välimäki, M. (2009). The Value of the EU Public Domain. Technical report.

PwC (2006). The Impact of Copyright Extension for Sound Recordings in the UK. A Report for the Gowers Review of Intellectual Property prepared by PwC on behalf of the BPI.

Ringstad, V. and Løyland, K. (2006). The demand for books estimated by means of consumer survey data. *Journal of Cultural Economics*, 30(2):141–155.

Rob, R. and Waldfogel, J. (2006). Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students.

Watt, R. (2000). *Copyright and Economic Theory: Friends or Foes?* Edward Elgar, Cheltenham, UK.