THE SIZE OF THE PUBLIC DOMAIN

RUFUS POLLOCK, PAUL STEPAN & MIKKO VALIMAKI

JULY 8, 2009

Abstract. This paper reports results from a large recent study of the public domain in the European Union. Based on a combination of catalogue and survey data our figures for the number of items (and works) in the public domain extend across a variety of media and provide one of the first quantitative estimates of the 'size' of the public domain in any jurisdiction.

Keywords: Copyright; Public Domain; Intellectual Property
JEL Classification:

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 the world. 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]

The work presented here is part of this effort – as well being part of a larger project to examine the overall *value* of the public domain. Here our focus is on the size and scope of the public domain, and provides quantitative estimates for number items in the public

Corresponding author: Rufus Pollock: rp240@cam.ac.uk or rufus@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 the other participants in the project, for making this research possible.

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

domains across a variety of EU countries and different media types. We have already made clear why the study of the public domain is important, but, why, specifically, are issues of size and scope interesting and important? There are several reasons.

First, in charting the current, and future, dimensions of the public domain we are estimating (historical) levels of cultural production. Not only is this interesting in itself, but as cultural economists such estimates are increasingly essential if we are to understand where we have come from and where we are going to (and this is important not only for cultural economists: culture and information playing a growing role in society as a whole).

Second, an examination of the public domain, and past cultural production more generally, allows us to make precise assessments of the impact on society of changes in copyright, especially copyright term. Recent years have seen several term extensions and even now a term extension for sound recordings is being considered within the EU. Estimates of the size of the public domain provide quantitative information on the number of works affected by such changes and could be an important input into policy-making in this area.

Third, and perhaps most importantly, is the relevance of our work to the widespread efforts to digitize cultural material and make it available online. Museums, libraries and archives are now considering,[2] or in the process of, digitizing large quantities of material with the intention of making it available to the public. Though both copyrighted and public domain material can be included in such efforts, the public domain is especially attractive because it can be made freely available without the need to identify rightsholders or pay royalties. This advantage is such, that for many projects, especially in the public sector, *only* public domain material is being considered. However, such projects require answers to questions such as: how much material could be digitized (from a given period)? How much, and what kind of, material be directly identified as public domain?[3] This paper provides detailed material that can help each of these questions.

Lastly, our results are a first, key, step in the larger effort to determine the overall 'value' of the public domain. In any area, 'count' statistics, such as the size estimates presented here, provide, the first, albeit perhaps crude, measures of significance. As discussed below, obtaining even these relatively basic figures is not a trivial matter and they form an essential base on which any subsequent work must build.

[2]There are also many commercial companies involved in these activities as well as public institutions.
[3]Similar issues also arise in the orphan works debate.

Before entering into the main body of the paper we should mention two important limitations of our investigations. First, the intricacies and variations of copyright law across member states (even post-harmonization) is such that the exactly delineating the public domain in each jurisdiction would be a major task. However, since our aim here is to obtain general quantitative *estimates*, a 'broad' approach which ignores these subtleties and uses an 'approximate' algorithm will be sufficient for our purposes.[4]

Second, and relatedly, it is obviously necessary, just as for copyright, to address the public domain separately for each separate media types: text, sound, moving images, etc. Furthermore, within each of those categories it may be useful, and important, to distinguish further, for example between books and other printed media such as journals or newspapers, between sound recordings and compositions, between multi-media and films etc. However, it should be mentioned from the outset, that though this is what is optimal in *theory*, in *practice* we will often be forced by the data available to narrow our coverage or limit the distinctions that can be made.

1.1. **Related Literature.** There is, at present, almost no *quantitative* research on the size of the public domain and we are aware of only one recent publication which bear directly upon this issue. This is David and Rubin (2008) which examines the impact of US copyright extensions on (reducing) the size of the public domain. Specifically, using the registration data uniquely available in the US, they calculate quantitative estimates of how many additional books would have been in the public domain if each of several 20th century extensions (ending in the 1998 CTEA) had not occurred. Here, our main focus is on the size of the public domain per se, not the impact of legislative changes upon that size – a task that is significantly harder in the EU compared to the US as registration information is absent and rules determining copyright status for older works is more complex.

2. The Size of the Public Domain

For our purposes the *size* of the public domain will equate to the number of works which are in the public domain – i.e. whose copyright has expired.

[4]Given the many other approximations that will no doubt be necessary the error induced by this particular assumption is likely to be, relatively, very small.

It must be noted that this simple definition conceals substantial complexity: "There is no clear and generally accepted definition of the concept of the public domain (le domaine public). The international conventions do not provide one and the vast majority of national laws do not do so either. [(Paul Torremans, 2008)]" This, combined with the significant jurisdictional variation in copyright – for example that moral rights survive the term of copyright in some jurisdictions – mean that the exact delineation of the public domain as a legal concept is not a trivial matter.[5] However, we need not concern ourselves greatly with these problems here and shall proceed on the basis that this simple definition suffices.[6] It should also be emphasized that our focus is on cultural material and will not include, for example, 'public sector information' which in many jurisdictions is either public-domain (e.g. the US) or almost so (the UK).

There is one important subtlety in this definition that we cannot ignore and that is around the term 'works'. In general, when we think of cultural material we think of their physical (or digital) instantiation: a book, a CD, a film. However, a moment's reflection reveals some subtleties: the same text, say Shakespeare's Hamlet, may appear in several different books, a recording may appear on many different CDs etc.

To keep our thinking clear we shall adopt terminology derived from the 'Functional Requirements for Bibliographic Records' cataloguing schema.[7] We shall mean by an 'item' something equating to a publication/release, for example a book with a given ISBN, a particular CD release, a given DVD of a film. By a 'work' we shall mean the underlying 'platonic' work instantiated in that item, i.e. the text, sounds etc in abstraction. As such a given work may have many associated items: for example, Shakespeare's play "Hamlet" is a single 'work' but there are many associated items (publications). We can also have derivative works, such as translations: these count as new works in themselves – though with some dependency on another work.[8] We also have a similar situation with a recording

[5]For more details see the work of Paul Torremans (2008), prepared as part of the Public Domain in Europe project of which this research was part.

[6]We should also note that we are explicitly *not* adopting a broad definition of the public domain which includes works still under copyright but where those rights have been largely waived using an 'open' license (this broad approach is adopted in, for example, Pollock (2006)).

[7]See `http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records` for more info.

[8]Another, slightly more subtle is that of a new edition, where the text has been amended in some ways. This would again count as a new, derivative, work though in this case the differences from the original work may be very minor.

which strictly should be seen as consisting of a work corresponding to the composition and a work corresponding to the performance (of that composition).[9]

Our purpose for introducing this distinction is two-fold. First, public domain status is properly seen as an attribute of a work not an item – and so the public domain status of an item is the status of its associated work(s).[10] Second, we now have two distinct meanings for size: the number of manifestations or the number of works. Which of these should we use?

In our view, it seems more correct, and appropriate, to base size calculations on works not items. After all the republication of an existing work does not add anything to stock of human knowledge – the republication in 1880 of Shakespeare's plays (first published in the Folio of 1623) should surely not be counted as new production for 1880.

The only possible reason not to use works would be the fact that many new publications of existing works frequently claim a new copyright based on slight emendations or typography. However, these alterations are so slight that even though sufficient to obtain a new copyright it seems difficult to see how a quantitative assessment should accord them the same standing as the original work itself.[11]

Unfortunately, almost all the data available only gives information on items. In particular, library catalogues – the primary source for our work here – only provide this kind of data (because items are what they hold in their archives).[12] Furthermore, deriving works from items in an *automatic* way is a non-trivial matter – as we shall discuss further below.

Thus, despite having just made compelling arguments as to why works are the best subject for any size calculation we must acknowledge that, in what follows, by and large,

[9]In some cases we may also wish to divide the composition into two distinct works corresponding to the lyrics and the music.

[10]We have to be a bit careful here since the copyright status of an item and its work may not be exactly the same: for example, even books containing pure public domain texts may have copyright in their typesetting. Also there may be additional non-PD material such as an introduction or commentaries, though, in this case, at least theoretically, we should say the item contains 2 works a) the original PD text b) the non-PD introduction.

[11]In some respects this is one aspect of the more general 'weighting' problem that occurs in any quantitative assessment of knowledge goods because of their marked heterogeneity. Simple counts, as presented here, are problematic because they take no account of this heterogeneity – many books are never read, even upon their release, while some remain perennially popular.

[12]This focus on items is even more true of databases listing information related to the value of works such as prices and sales.

We'd also note that this deficiency is even true where registration information is available, as in the US. This is because, as discussed above, most new publications, even those which are simply reissues of previous material, will usually claim some kind of new copyright and will therefore be registered. Having acknowledged this deficiency.

when we talk of the *size* of the public domain we shall mean the number of public domain items and not the number of public domain works.[13]

3. Determining Public Domain Status

With our terminology in place determining public domain status is, in theory, a simple case of applying copyright law. There are of course some subtleties, for example the various "special cases" and one-off exceptions in copyright, such as the fact that, in the UK, the Copyright Designs and Patents Act para 301 contains a special provision so that *Peter Pan* by J.M. Barrie will remain in copyright forever (with royalties payable in perpetuity for the benefit of Great Ormond Street Hospital).[14] However, we are not seeking to compute the status of individual works but rather to obtain gross estimates and as such we are willing to accept a public domain 'algorithm' that is not perfect but only, say 99.99% accurate.

As such, for our purposes we shall adopt a very simple approach, and proceed on the basis that, in the EU, published literary, musical and artistic material by authors who died more than 70 years ago is now be in the public domain. For recordings, the recording itself (as distinct from the composition) will be public domain if more than 50 years old – and the entire as a whole will be public domain if both the recording itself and the composition are public domain.

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

[13]We were able to perform some fairly limited analysis around works. In particular, using the 'raw' catalogue data from CUL (see below for details) and a basic work generation algorithm, we established that approx 13% of all items could be matched to a work (i.e. an object to which two or more items could be attached) with the at least half of these being works with just two items. Such figures would suggest that based estimates on items rather than works would not lead to too great a bias (though we would note that it is precisely those works with many items that are likely most 'valuable'). However, in the absence of very substantial additional labour, we could not establish how good this result was – in particular, whether the other 87% of items were in fact works (i.e. were the publications of works which were only ever published once so the work and item were identical for our purposes), or whether we were simply failing to match many items. As such an estimate of public domain 'size' based on works is an ongoing research topic.

[14]For a jurisdiction by jurisdiction analysis see the work of Torremans et al.

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.[15] Thus, it may be necessary to fall back on some approximate method.

One possibility, and that which we shall adopt, is to base PD status on publication dates: if a book was published, say, 140 years ago it is almost certain it is in the public domain – for it to be in copyright its author must have lived more than 70 years after the book was published (copyright lasts for life plus 70 years in the EU). Conversely, any publication less than 70 years old is unlikely to be in the public domain. For periods in between we can assume some proportion of publications are PD starting close to zero for more recent items and rising towards one for older ones.

This option is attractive because, in contrast to the sparseness of authorial dates, publication dates are almost always recorded in catalogues.[16] Thus, in what follows it is the main approach used. Of course, this approach is sensitive to the PD weightings/proportions used for the different publication 'vintages'. We will therefore seek to check these weightings for robustness as well as basing them, where possible, on available authorial information.

4. Data

Cultural institutions, primarily libraries, have long compiled records of the material they hold in the form of catalogues. Furthermore, most countries have had one or more libraries (usually *the* national library) whose task included an archival component and, hence, whose collections should be relatively comprehensive, at least as regards published material.

[15]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.

[16]It was this fact, and its implications for the ease of extraction from library information systems, that the summary information we asked for from libraries only required that for counts be broken down by publication date.

The catalogues of those libraries then provide an invaluable resource for charting, in the form of publications, levels of information production over time, and hence the size of the public domain.[17]

Our main task then has been to obtain from such institutions – usually national libraries – whatever relevant data they have. We requested two levels of information from the libraries we contacted. First, and most basic, was simple summary information listing approximate holdings by decade and type of content – books, recordings etc.[18] Second, we asked whether libraries would provide us with 'raw' (MARC) catalogue data with which to perform a more comprehensive analysis based on our own computations.

It did not prove easy to obtain data. Many of the libraries contacted did not respond at all. Of those that did respond the majority only provided summary information with only a few (two in the UK and on in Slovakia) being able, or willing, to provide us with the 'raw' catalogue data. [19]

In these circumstances we had to 'take what was given' in terms of the coverage of different types of media. In particular, the majority of the data we have relates to books, with some (more limited) information on compositions and recordings and no information at all on, for example, photographs or films.[20] Table 1 gives a summary of the data we have been able to obtain (more detailed information may be found in the appendix).

In what follows the majority of our analysis will focus on the raw data. This is for two reasons: first, 'raw' data provides information at a much greater level of granularity (individual entries); second it is only in the raw data that we get on authors, and, in particular on their birth and death dates.

4.1. **Data Processing.** In a data-oriented project such as this data-processing is a very substantial part of the work involved. The scale of the task was substantial: Cambridge University Library (CUL) pre-1960 holdings totalled over 1 million records while the British Library's (BL) had more than 4.2 million for the same period. Processing and

[17]Subject, of course, to the obvious caveats about coverage – though we would point out that, at least from the perspective of today, that material not available in the archives of our major libraries is likely lost to us forever.
[18]Specifically we sent libraries a simple spreadsheet questionnaire. See appendix for details.
[19]In one particular case a library explicitly declined to assist us citing confidentiality agreements they had signed with Google.
[20]Though in the case of films we have been able to extract data from `http://www.imdb.com/`.

Country	Work Types	R/S	Comments
Bulgaria	Bk, Ph	S	Bulgarian National Library
Slovakia	Bk etc	R + S	Slovakian National Library
Czech Republic	Bk	S	National Library of the Czech Republic
Poland	Bk, Rec, Ph etc	S	National Library of Poland
France	-	-	Received general information from BNF but no specific data.
Italy	Bk, Rec	S	Various libraries
UK	Bk, Rec	R + S	British Library and Cambridge University Library
Ireland	Bk, Rec, Ph	S	Trinity College Library, Dublin
Finland	Bk	S	National Library
Netherlands	Films	S	Netherland's Filmmuseum
Hungary	Films	S	Hungarian National Film Archive

TABLE 1. Library catalogue information provided to the project. Bk = Books (may also include periodicals), Rec = Recordings, Ph = Photos. R/S = Raw/Summary data.

analysing material, particularly on this scale, presents significant challenges.[21] To illustrate, consider one of our most basic tasks: loading information from the MARC format catalogue dumps into our DB for analysis. This had the following steps:

(1) Do a simple load: i.e. for each catalogue entry create a new Item *and* new Persons for any authors listed.
(2) "Consolidate" all the duplicate Persons, i.e. a Person who is really the same but for whom we create duplicate DB entries in part 1 (we can do this because MARC cataloguers try to uniquely identify authors based on name + birth date + death date).
(3) Consolidate items" to works (i.e. associate multiple items – distinct catalogue entries – to a single work).

Take the first step, which is the simplest. Implementing this requires several steps. First we need to parse the MARC itself for which we use an existing library (pymarc). Next we need to determine which fields we require and do some processing to consolidate them (for example, additional authors are often placed in the MARC 700 field in addition

[21]All of the code used in this work is open-source and available from the following mercurial repository: `http://knowledgeforge.net/pdw/hg/`.

to the main author field, field 100). For this, after experimenting with our own, we were able to adapt the existing work of the Open Library project. Next we need to parse and normalize values such as dates and author names (while author names tend to be standardized within a given catalogue we often need to match both across catalogues and to other datasets). Dates, in particular, are obviously crucial to our project and are often not given in a easily 'machine readable' format. For example, we had to cope, not only with different dates levels, e.g. 1865, March 1865, etc, but also ambiguity, e.g. '15 or 17 March 1865', BC dates e.g. '213 BC', or latin dates, e.g. 'MCLVI'.[22] Thus, even in the very simplest stage of the process, there was much to be done of this kind of mundane but essential work.

It is also important to realize that working with large datasets creates special challenges because speed and space really matter. Consider again the loading process just discussed. On a 1 million record load, at the first step (simple load into the DB), we averaged (depending on hardware, DB backend etc) between 8s and 25s per thousand records with speed fairly constant throughout. This means that a full load required between 2.5 and 7.5 hours.

The second and third stages then took another few hours each. All in all, this may not seem too bad: around half a day to load and process the whole catalogue. However, it must be remembered that these steps must be repeated each time a bug is discovered or a schema change or algorithm improvement necessitates a reload. Moreover, such improvements – and the need for them – often cannot be discerned until *after* you have run them on a reasonably large dataset.[23]

Finally, we should note that all of the code used in our work has been made open-source and is available directly from this mercurial repository: `http://knowledgeforge.net/pdw/hg`.

[22]Many programming languages and databases do not even support BC dates in their standard date formats. For example, python, the programming language we were primarily using, will not store (as date objects) dates before 0 AD, and postgresql (a major opensource database) will not store date before 4317BC as date objects.

[23]To take yet another very minor example: in computing public domain status based on author death dates we are required to conduct a join of two large tables (items to persons). The preferred way of doing this involved an outer join. However, running on one of our database backends (sqlite) this outer join would seemingly never end (after leaving this for 12h+ overnight we terminated the process). There were two ways to solve this: use an inner join (which ran very quickly) at the cost of some work arounds elsewhere or switching to an alternative backend (requiring another 8h+ load!). After some checking that revealed that the alternative backend (postgres) did not have the problem with the outer join we switched.

5. Results

From a practical point of view the task of computing PD size has two distinct aspects:

(1) Estimate the number of items produced (of a given type and vintage/publication date)
(2) Compute their public domain status (or proportions)

For the first step, as already discussed, we shall be relying on library catalogues to furnish us with our estimates. For the second part, as already discussed in the section above, there are have two options:

(1) (Direct/Precise) Compute the public domain status of individual items based on applicable copyright law – this will usually require author death date information
(2) (Indirect/Approximate) Apply a simple public domain 'weighting/proportion' based on *vintage/publication date*

Clearly the first of these approaches is preferable. However, limitations of available data mean that the second approach will often be the only one possible. In particular we cannot directly compute public domain status of individual items when either a) information on individual items is not available – e.g. when we only have summary data or b) the individual items do not contain sufficient information (e.g. no author death date). Nevertheless, we shall begin with an attempt at the direct approach. Not only will this show some of the difficulties involved but it also will provide the basis for the weightings to use in the indirect case.

5.1. **Direct Calculations Using Raw Data.**

As is clear from the preceding discussion direct calculations are only possible with raw data since it is only raw data which has information at the level of individual items, and, furthermore, information relevant to public domain status: a) a publication date b) unambiguously identified author(s) with, perhaps, a birth date(s) and, less frequently, a death date.

In this section we shall look at using this kind of information to compute public domain status for a large set of records. Of the three sets of raw catalogue data available we shall focus on that provided by Cambridge University Library (CUL). This is primarily for simplicity of exposition but we would note that CUL's data was as detailed (if note

more so) than either of other two raw datasets and was reasonably large (over 1 million records up to 1960).

Publ. Date	Total	No Author	Any Date	Death Date
1870-1880	50564	6634 (13%)	23016 (45%)	21876 (43%)
1880-1890	66857	8225 (12%)	31135 (46%)	28570 (42%)
1890-1900	66883	8733 (13%)	32169 (48%)	28971 (43%)
1900-1910	70360	8594 (12%)	35401 (50%)	29922 (42%)
1910-1920	60489	7722 (12%)	31336 (51%)	24608 (40%)
1920-1930	78670	9023 (11%)	44219 (56%)	32658 (41%)
1930-1940	90576	11004 (12%)	46849 (51%)	29372 (32%)
1940-1950	72692	7638 (10%)	36495 (50%)	22155 (30%)

TABLE 2. PD relevant information in CUL catalogue. 'Any Date' indicates records with either a birth *or* a death date for the author.

We begin by indicating the basic data problems, even with raw information, when trying to compute public domain status directly. Table 2 presents a summary of how much relevant information is available for items (books) of particular vintages in the CUL catalogue – we only show data from 1870 onwards on the presumption that (almost) all pre-1870 publications are PD (their authors would have had to live for more than 70 years post-publication for this not to be the case).

As the table shows, at best only just over 40% of items have a recorded authorial death date and extending to include birth dates only raises this proportion to, at best, the mid mid-to-low fifties. Taking account of items which lack any associated author, raises these figures somewhat further to around 60%, though we should note that the reason for the lack of an associated author is not clear – is it because they are genuinely anonymous or simply because the information has not been recorded?[24] Thus, even for the earliest items listed a large proportion of items (50% or more) lack the necessary information for direct computation of public domain status.

At the same time, we can take some heart, and some interesting facts, from this table. First, a reasonable proportion, amounting to many thousands of items, did have associated death dates. Second, at least for older items, the majority of items with any date had a

[24]This 'No Author' figure, which is remarkably constant over time, has independent interest, most notably for the 'orphan works' debate. For these figures give a precise measure of the number and proportion of items without any identifiable author and which are therefore in the strictest sense 'orphan' (a broader definition would also include those works for which an author can be identified but where than author can be located without significant difficulty).

death date (95% for 1870-1880 and still at over 70% for 1920-1930).[25] Third, and this is a more general observation, proportions were surprisingly constant over time. For example, the proportion of 'anonymous' items lies in a narrow band between 10% and 13% for the entire periods. Similarly the proportion of items with any date information ranged only from 45% to 56%. At the same time, and reassuringly, though the proportion with death dates is relatively constant for the oldest periods, in the more recent ones it falls substantially; as one would expect given that some of the authors from those more recent eras are still alive.

Publ. Date	**Total**	**PD**	**Not PD**	**?**	**Prop 1**	**Prop 2**
1870-1880	50565	22157 (43%)	68 (0%)	28340 (56%)	99%	96%
1880-1890	66858	28325 (42%)	649 (0%)	37884 (56%)	97%	90%
1890-1900	66884	26723 (39%)	2418 (3%)	37743 (56%)	91%	83%
1900-1910	70362	24032 (34%)	5838 (8%)	40492 (57%)	80%	67%
1910-1920	60491	16200 (26%)	8306 (13%)	35985 (59%)	66%	51%
1920-1930	78671	16127 (20%)	16351 (20%)	46193 (58%)	49%	36%
1930-1940	90583	8973 (9%)	20835 (23%)	60775 (67%)	30%	19%
1940-1950	72696	5000 (6%)	19316 (26%)	48380 (66%)	20%	13%

TABLE 3. PD status based on algorithm described in the text. '?' indicates items where PD status could not be computed. Prop(ortion) 1 equals total PD divided by total for which status could be computed (sum of total PD and Not PD). Prop(ortion) 2 equals total PD divided by number of items for which any author date was known ('Any Date' in previous table).

Table 3 reports the results of direct computation of PD status based on the information available. Note that, in doing these computations, we have augmented the basic life plus 70 rule with the additional assumptions that a) all items published in 1870 or before are PD b) no-one author is older than 100 (so if a birth date is more 170 years ago the item is PD) c) every author lives at least until 30 (so that any work published by an author born less than a 100 years ago is automatically *not* PD).

As is to be expected, for the majority of the periods, availability of PD status (either PD or Not PD) closely tracks the availability of death date information – it is only in the last period 1940-1950 that the birth date appears to make any contribution (in all other periods the sum of PD and Not PD approximately equals the number of items with death date information). More interesting, is how the number PD and Not PD vary over time,

[25]This was something we were slightly surprised at given our prior understanding of cataloguing procedure.

especially relative to each other (and as a proportion of the records for which any date date is available).

These two proportions/ratios are recorded in the last two columns which record, respectively: 1) the PD total relative to the number of items for which any status could be computed (i.e. the sum of PD and Not PD) 2) the PD total relative to the total number of items for which any date information is available. These ratios change dramatically over the periods shown: starting in the 1870-1880 period in the high 90%s by the 1940s they are down to 20% or below.

Pub. Date	**% PD**
0000-1870	100
1870-1880	95
1880-1890	90
1890-1900	85
1900-1910	65
1910-1920	40
1920-1930	15
1930-1940	5
1940-1950	1
1950-Now	0

TABLE 4. PD Proportions

The key question for us is how to extrapolate these PD proportions to the full set of records – i.e. from the set of records for which there is the necessary birth/death date information to that where there is not. The simplest, and most obvious, approach is to assume that the proportions are identical and therefore that the PD proportions calculated on the partial dataset apply to the whole. However, there are some obvious deficiencies in this approach. In particular, our ability to compute a PD status is largely linked to the existence of a death date and it is likely that the presence of this information is itself correlated with authorial age – after all a death date can only exist once that person has died! This correlation, and the bias it gives rise to, is probably small in the early periods – the authors of any pre 1930 work are almost certainly no longer alive today. However, for the later periods, the bias may be more substantial – it is in these last two periods (1930-1940 and 1940-1950) that there is a significant reduction in the number of records with a death date and (relatedly) a significant increase in the number of records for whom the PD status is unknown.

Thus, in converting the partial PD proportions to full PD proportions it seems sensible to revise down somewhat the partial figures with the revision being greater in later periods. Moreover, we have a lower bound for any downwards revision provided by the total PD as a proportion of all records which, even in the 1940-1950 period, was 6%.[26] In light of these considerations Table 4 gives suggested values for PD proportions that we can use for the rest of the paper. We would emphasize that we have elected to be conservative in our assessment and to choose figures, particularly in later, periods at the lower end of the possible band (or even below).

5.2. **PD Size.** We now combine the PD proportions derived in the previous section with estimates of total production to obtain an overall figure for the total number of PD items of a given type in a given jurisdiction.

We begin by presenting results based on analysis of 'raw' catalogue data provided by three libraries: Cambridge University Library, the British Library and the Slovakian National Library. Both Cambridge University Library and the British Library are UK "copyright libraries", that is, they have a right to obtain, though not an obligation to hold, one copy of every book published in the UK. However, it should be made clear that this right is not always exercised (and libraries also acquire books published abroad). Thus it is difficult to know how well catalogues, even for copyright or national libraries, accurately represent past production levels. Nevertheless, since duplication is minimal, these sorts of figures do, at the very least, indicate lower bounds. Furthermore, in the UK we have the advantage of two different sources which will allow us to do some double-checking.

The results of the simple 'proportional' approach to estimating PD size are shown in figures 1, 2, 3 and table 5 (we only show the one summary table – for CUL – for reasons of space). These show publications per year up until 1960 (when the datasets end) based on the publication date recorded in the catalogue. Interestingly the basic pattern shown by the CUL and BL catalogues is very similar. However, the BL catalogue records 4-5

[26]In many ways this is a surprisingly high number: the output of any author alive in 1940 are still in copyright! Thus, one presumes that much of this figure corresponds to new publications of existing material which had already previously published (the only other alternative is posthumous publications). In the nomenclature developed previously these would be considered new *items* but not new *works*. This then is one area where a size measure based on works and one based on items will clearly diverge.

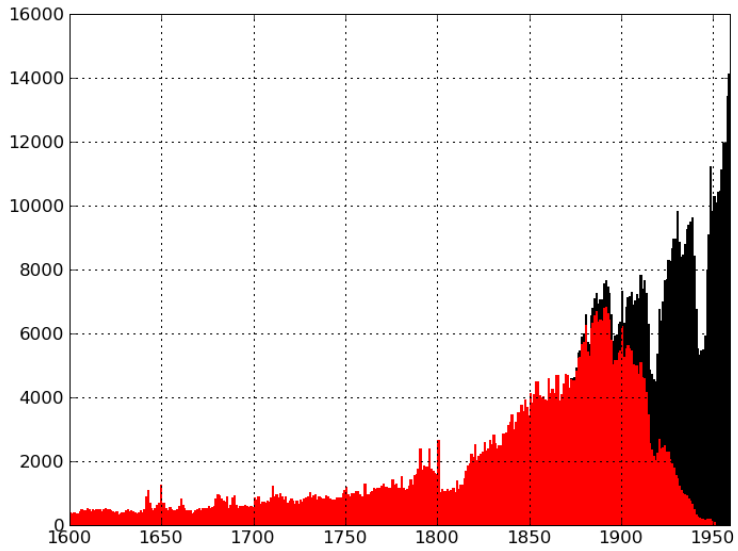

FIGURE 1. Total (Black) and PD (Red) Items based on the CUL Catalogue

times as many publications per year. The exact reason for this sizable discrepancy is not entirely clear.[27]

Pub. Date	Items	% PD	Number PD
1400-1870	389291	100	389291
1870-1880	50564	95	48035
1880-1890	66857	90	60171
1890-1900	66883	85	56850
1900-1910	70360	65	45734
1910-1920	60489	40	24195
1920-1930	78670	15	11800
1930-1940	90576	5	4528
1940-1950	72692	1	726
Total	946382	67	641330

TABLE 5. Estimated Number of PD Items based on CUL Catalogue

Thus, based on our assumptions of PD proportions and the CUL catalogue, there are somewhat over **600,000** textual items in the Public Domain. Just over half of this 600k

[27]Several possibilities were suggested. First, CUL have not fully digitized their catalogue data while the BL have done so. This means that CUL figures understate the true number of published items – the very approximate estimate was that there may be a few hundred thousand catalogue records not yet digitized for our period. Second, the British Library has acquired over time sizable collections from elsewhere (for example, from the British Museum and the National Lending Library), and this may have led to significant duplication in its holdings – i.e. having multiple copies of the same time – though this need not necessarily have resulted in duplicates in the catalogue depending on how thorough the process of consolidation had been.

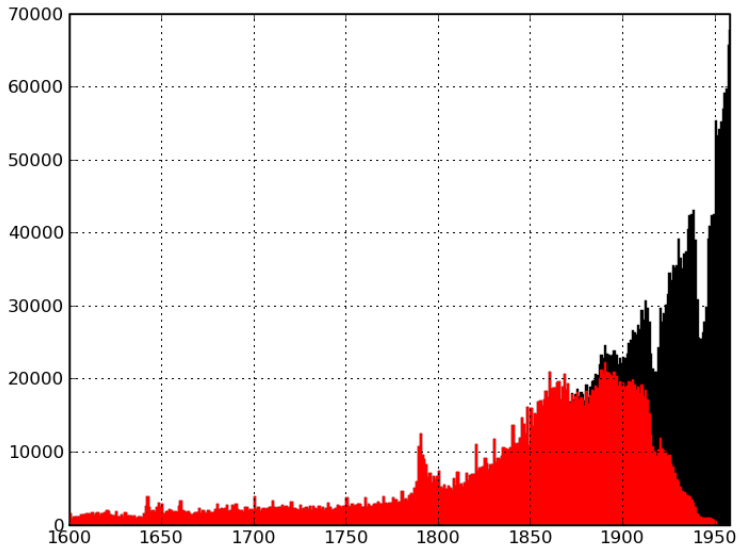

FIGURE 2. Total (Black) and PD (Red) Items based on the BL Catalogue

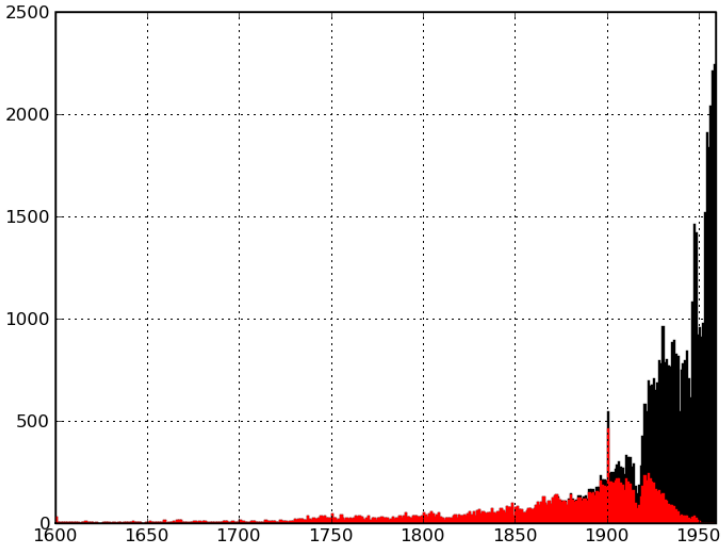

FIGURE 3. Total (Black) and PD (Red) Items based on the Slovak National Library Catalogue

(approx 390k) date from before 1870. For the BL dataset the same calculations yield approximately **2.3 million** Public Domain items. For Slovakia, based on the holdings of the Slovak National Library, the corresponding figures approximately 16 thousand items (out of a total of approx 42 thousand).

Lastly we present the analogous results in Table 6 for the other European countries for which we had data – though in summary not raw form (i.e. only simple counts by decade of publication were available). As the table shows, even in the lowest cases, the public domain is substantial, in the tens of thousands of items, and for the largest cases, for example Italy, Poland and Ireland, it runs well into the hundreds of thousands of items.

6. Sound Recordings

For recordings the public domain situation is both simpler and more complex. Sound recordings contain two distinct 'rights'. An authorial copyright in the composition (music and lyrics) and a 'recording' copyright (or neighbouring right) in the recording itself. The rights in the authorial copyright are treated like any other copyright and receive the standard life plus 70 years. However, the recording right runs for a simple 50 years from publication. Thus, if we restrict our meaning of public domain to the expiry of the recording copyright then it is sufficient to know the date of release. However, if we wish to know that the recording is 'fully' public domain with both authorial and recording copyright expired, then we shall need to have detailed information on the composer of the work.

The majority of the data provided was of a summary nature, only providing counts based on decade of release. Of this type, we had information from three libraries in three countries: Poland, Ireland and the UK. Focusing on pre-1960 recordings, i.e. those recordings whose recording copyright has, or is nearly, expired, these three reported, respectively, total holdings of approximately fourteen thousand, four thousand **290 thousand** recordings. It is clear from this that the UK holdings (at the British Library) are by far the largest. Furthermore, though without more detailed data, we cannot calculate their full PD status for certain, the prominence of classical music in early recording makes it likely that a very substantial number of these recordings, probably the majority, are fully public domain.

To add to this, we did manage to do some calculations based on 'raw' data. There were two sources: the BBC CAIRNS catalogue which lists recordings held in the BBC archives, and the catalogues collected by the CHARM project at King's College London. Unfortunately though the CAIRNS catalogue was large – running to the hundreds of

Period	**Bulgaria**		**Czech Republic**		**Poland (1,5)**		**Italy (2,3)**		**Ireland (4)**		**Finland**	
	Total	PD	Total	PD	Total	PD	Total	PD	Total	PD	Total	PD
Before 1850	226	226	7787	7787	60000	60000	0	0	0	0	20497	20497
1850-1870	1054	1054	7747	7747	80000	80000	0	0	0	0	1728	1728
1870-1880	900	855	6837	6495	40000	38000	0	0	15232	14470	1840	1748
1880-1890	2789	2510	8607	7746	50000	45000	45110	40599	22234	20010	4402	3961
1890-1900	6392	5433	13278	11286	50000	42500	103291	87797	32615	27722	9398	7988
1900-1910	8296	5392	16363	10635	70000	45500	67931	44155	49721	32318	16264	10571
1910-1920	8487	3394	16735	6694	80000	32000	91644	36657	50240	20096	21403	8561
1920-1930	16009	2401	59374	8906	80000	12000	58227	8734	200000	30000	29338	4400
1930-1940	18444	922	83555	4177	150000	7500	103504	5175	325000	16250	34882	1744
1940-1950	6921	69	65905	659	30000	300	75569	755	400000	4000	36386	363
1950-1960	71	0	146319	0	100000	0	103559	0	500000	0	39433	0
Totals	69589	22256	432507	72132	790000	362800	648835	223872	1595042	164866	215571	61561

TABLE 6. Number of Public Domain Items based on Summary Data from Various National European Libraries

thousands of records – the level of detail for most entries was very low with not only little author information but often not even a release date.[28]

The CHARM data was rather better though it was limited in that it was focused only on discographies of particular record labels or particular composers (e.g. Schubert) and had no aim to be comprehensive. Nevertheless, as a high quality resource with few, if any, duplicates and good information on composers and publication dates this dataset can at least provide us with a good lower bound.[29] Based on this data, we estimated that there are at least twenty thousand fully public domain recordings.

7. Film

In the EU copyright in films last for seventy years after the death of the last primary 'contributor' (director, script-writer, etc). Since film, as a artistic medium, has only been existence since the very end of the nineteenth century it is therefore likely that very little film is in the public domain – for example, Auguste Lumire, who, as one of the Lumire brothers, presented one of the first public cinema screenings in 1895, died in 1954.

8. Conclusion

In this paper we have provided the first quantitative assessments of the size of the European public domain – one of the first such assessments for any jurisdiction. Our results indicate that the public domain for books alone consists of hundred of thousands, and sometimes millions, of items, and that, taken as a whole, the European Public Domain must be measured in the millions, or even tens of millions. While a brief perusal of the relevant datasets indicates that much of this material may have only slight value today, nevertheless the scale and diversity of this vast public domain is indicative of significant value, cultural, social and commercial.[30]

[28]Where persons were identified performers were not distinguished from composers and unlike library catalogues persons were not necessarily uniquely identified and no dates were provided. Thus, any usage for public domain calculations would have required matching this database to a separate authorial database or the addition of dates by hand – something that was not really feasible given the scale of the catalogue.
[29]Also note that this data focuses on recordings released by the main record labels – unlike library catalogues which may include additional recorded material (e.g. interviews).
[30]A more precise quantification of this value, focused on price, usage and availability, is intended to be the subject of future papers.

In addition to the main size estimates, we have also produced results on several related points. Perhaps most important was the derivation of Public Domain 'proportions/weightings' which could be used based solely on the 'vintage' (publication date) of works. As discussed above, for the majority of cultural material, copyright is a function of authorial death date. Unfortunately, this information is frequently unavailable – in contrast, to the publication date which is almost always provided. By analyzing the raw catalogue data provided by some libraries we were able to link calculations of PD status based directly on authorial information to the publication dates thereby deriving PD 'proportions/weightings' which are usable in a much wider set of circumstances. Moreover, the very same analysis also gave valuable statistics both for the number of works for which direct calculation of PD status is not possible and for the number of works with no authorial information at all – important figures for any potential digitization project (as well as for the debate over orphan works).

References

David, P. A. and Rubin, J. (2008). Restricting Access to Books on the Internet: Some Unanticipated Effects of US Copyright Legislation. *Review of Economic Research on Copyright Issues*, (1):23–53.

Paul Torremans (2008). The 'public domain'. Produced as part of the Public Domain in Europe project.

Pollock, R. (2006). The Value of the Public Domain. Published by the Institute for Public Policy Research as part of a series on IP and the Public Sphere.