The 20th Century Black Hole: How does this show up on Europeana?

As cultural heritage institutions across Europe digitize more and more of their collections and make them available online, an alarming pattern is starting to emerge. Collections that consist of works dating from the 20th century or that contain large proportions of works from that period are available online to a much lesser degree than collections from the periods before or after the 20th century. This effect has been called ‘the 20th century black hole’ and can be attributed to the way copyright interacts with the digitization of cultural heritage collections.

Copyright law requires that anyone who reproduces and/or makes available copyright-protected works needs to obtain permission from rights-holders to do so. In the context of mass digitization projects, this means that cultural heritage institutions need to obtain permission from all rights-holders before they can digitize and make available works that they have in their collections (in the majority of cases, archives, libraries and museums own physical copies of copyright-protected works, but they do not own the copyrights which tend to rest with the creators or publishers of the works).

Obtaining permission from rights-holders for historical collections can be very time consuming as it requires the institutions to identify rights-holders and then obtain permission from them. This is complicated by the fact that many rights-holders do not actively manage their works any more. Rights clearance tends to be one of the most expensive elements of digitization projects and as a result institutions often limit digitization projects to collections that are in public domain (which can be used without permission) or newer collections (for which copyrights were cleared when works were acquired). The result is a marked lack of online availability of 20th century collections - the 20th century black hole.

Showing the 20th century black hole in Europeana

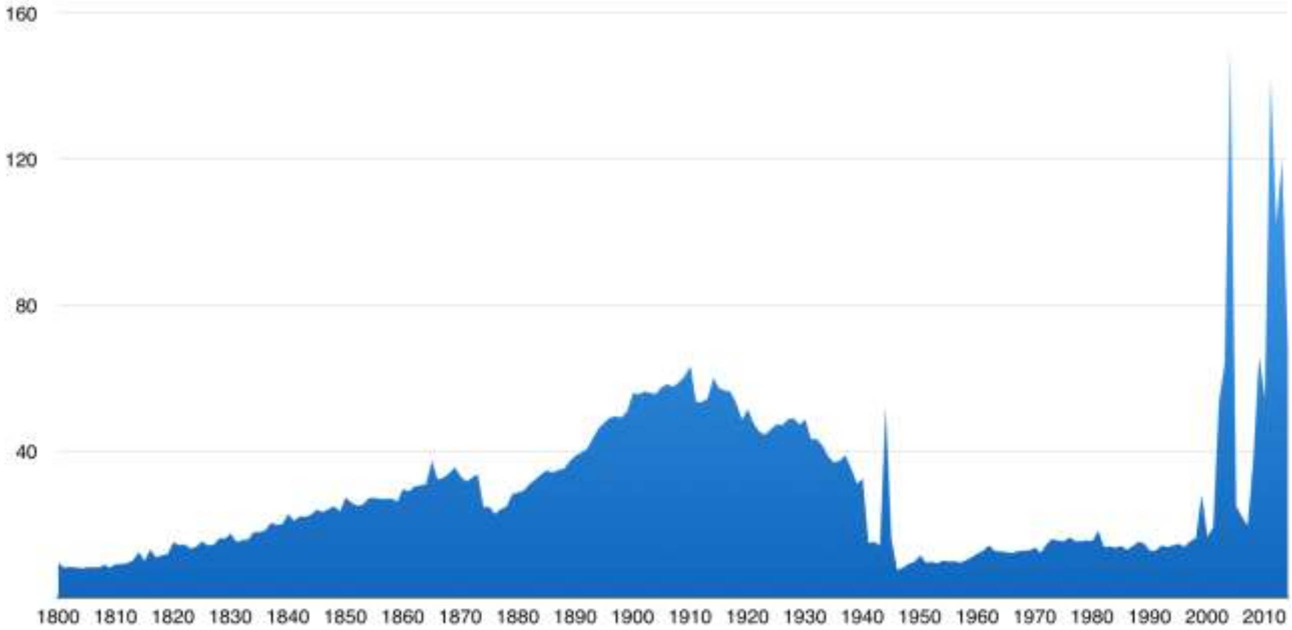
We have analysed the Europeana dataset to explore this claim and our findings show that there is a clear gap in availability of digitized material from the 20th century. The following visuals represent an analysis of Europeana’s dataset[1] of roughly 45 million objects. Of the 45 million, we selected approximately 7,300,000 objects that contain the most reliable

[1] Europeana’s dataset is categorised in images, texts, video, 3D objects and sounds.

information with regards to the date of creation of the work (see the annex for more information on the methodology using the *dcterms: issued* values in the Europeana dataset).

The results give an approximate breakdown of the numbers of items represented digitally in Europeana that were created in each year since 1800 (dates submitted for the digital object are known to not always be accurate):

Chronological distribution of *dcterms:issued* values in Europeana dataset (1800-today)



Distribution of digital object across historical periods

Total Values	6,223,992	100%
19th century	2,403,810	38.6%
20th century	2,859,345	45.9%
21st century	960,737	15.43%
1st half of 19th century (1800-1849)	747,741	12.01%
2nd half of 19th century (1850-1899)	1,656,069	26.60%
1st half of 20th century (1900-1949)	2,179,631	35.01%
2nd half of 20th century (1949-1999)	680,084	10.92%

It is evident that the amount of cultural heritage made available online increases steadily from the 1800s to the second half of the 20th century (the unusually high number of works dating from 1944 is the result of a single large dataset).

From the 1950s onwards, the amount of material that is made available online falls dramatically. While the first half of the 20th century represents 35% of the sample, the second half is only around 11%. These findings reinforce our earlier research (from 2012) and illustrate once more that cultural heritage institutions are hampered in their ability to make collections from the 20th century available online.

Conclusion:

As we have shown above, the 20th century black hole can be clearly illustrated by the Europeana dataset. While we cannot show a causal relationship between this and the way copyright law interacts with digitization efforts by cultural heritage institutions, we have received numerous reports from cultural heritage institutions indicating that the complicated copyright status of 20th century collections hinders digitization efforts. Cultural heritage institutions need the legal space to do their work without always having to negotiate with copyright holders who are often impossible or prohibitively expensive to find.

This means that in order to promote the online availability of cultural heritage from the 20th century, it is necessary to reduce the burden of rights clearance for these institutions. This can be done by updating the existing exceptions to copyright that apply to cultural heritage institutions. Making available works from their collections that are not in commercial circulation or otherwise actively managed by their rights-holders should not require permission from rights-holders (see our position paper for more details on how updated exceptions can help cultural heritage institutions to make their collections available online without harming the legitimate interests of creators and other rights-holders).

Annex 1: Methodology for analysing Europeana data

The Europeana Data Model allows data providers to enter the date of a work in a number of ways.The following table outlines the different options (the properties column) available to data providers and what each property means.

Property	Description	Amount
dc:coverage	The spatial or temporal topic of the CHO, ie. <dc:coverage>1996-1997<dc:coverage>, or <dc:coverage>Berlin<dc:coverage>	2,324,324
dc:date	Use for a significant date in the life of the CHO. For example <dc:date> Early 20th century<dc:date>	20,384,768
dcterms:created	The date of creation of the CHO. For example <dcterms:created>Mid XVIth century<dcterms:created> , or <dcterms:created>1584<dcterms:created>	10,408,765
dcterms:issued	Date of the formal issuance of the CHO. For example <dcterms:issued>1993<dcterms:issued>	7,302,190
dcterms:temporal	Temporal characteristics of the CHO. i.e what i.e what the CHO is about or depicts in terms of time. <dcterms:temporal> Roman Empire<dcterms:temporal>	4,530,677
edm:year	A point of time associated with an event in the life of the original analog or born digital object.This property is slightly different from the others. From July 2015 the four digit year (YYYY) is created by Europeana during the ingestion process from all the properties listed above.	15,462,075

Depending on their characteristics, and the number of records populated, using each property has both advantages and disadvantages to depict the 20th century black hole.

Methodology for analysing the data

We have decided to use *dcterms:issued* for several reasons:

- It is less ambiguous than other properties. It is used only for the purpose of indicating the date of issuance of an object. Other properties are not as reliable. *dc:date* for example could be used to indicate any important date in the life of the object, including creation, acquisition, donation etc. This could create a big inaccuracy when dating the object.
- The data present in *dcterms:issued* is more homogeneous than in other categories so it is easier to identify the different data patterns and capture them.

Nevertheless using *dcterms:issued* also has some disadvantages:

- The year of issuance does not necessarily match the year of creation and there could be a big discrepancy between them.
- Some data represent fragments of works. For example, the pages of a book could be described individually and not as a single record, therefore the same work could be counted more than once.
- This property is not the most frequent in the data. It is only populated in 7,302,190 records out of 44,725,949.

In order to get the data we have queried the Europeana API. This is the query that we have used:

http://www.europeana.eu/api/v2/search.json?query=*:*&profile=facets&facet=proxy_dcterms_issued&rows=0&f.proxy_dcterms_issued.facet.limit=350000&wskey=api2demo

Not all the data is valid for our purposes so it is necessary to filter it out. We have done this based on the following:

We have only taken into account the 19th and the 20th century and the first years of the 21st century up to 2014.

- We have used only data that indicates a year precisely, so approximations like *[ca. 1850] 1890s, [2008 ?]*, or periods like *1914-1918* have been ignored.
- We have tried to capture all the year representations and standard date formats present in the data, including *dd/mm/yy, dd.mm.yy, dd-mm-yy* etc (see patterns below).
- After examining the data we have included other patterns that are difficult to define beforehand like: *1823 [publication]* or *2013-05-30T19:40:27Z*.

As a result of this filter we have discarded approximately 20% of the results.

The following table lists the data patterns that were applied to capture the diversity of dates that are present in the data:

Regular expression[2] (below *YEAR* is replaced with the exact year being measured, which ranged between 1800 and 2015)	Example
Patterns for specific dates	

`^\s*[\[(]YEAR[\])][^?\w]*$`	`[1980]`
	`(1977)`

`^\s*YEAR(\s*[-/.]\d{1,2}){0,2}[^?\w]*$`	`1980`
	`1980/12`
	`1980.1.01`

`^(\s*\d{1,2}\s*[-=/.]){1,2}\s*YEAR[^?\w]*$`	`12-1980`
	`01-12-1980`

Patterns for annotated dates	
`^\s*YEAR([-/.]\d{1,2}){0,2}\s*\[[Pp]ublication\][^?\w]*$`	`1980/10 [Publication]`
`^\s*YEAR([-/.]\d{1,2}){0,2}\s*\(first performance\)[^?\w]*$`	`1980.2 (first performance)`
Patterns for date ranges within a single year	

`^\s*YEAR\s*[-=/.]\s*YEAR[^?\w]*$`	`1980-1980`
	`1927/1927`

`^\s*YEAR[-/.]\d{1,2}\s*[-=/]\s*YEAR[-/.]\d{1,2}[^?\w]*$`	`1980.1/1980.12`
	`1912-05/1912-05`

`^\s*YEAR[-/.]\d{1,2}[-/.]\d{1,2}\s*[-=/]\s*YEAR[-/.]\d{1,2}[-/.]\d{1,2}[^?\w]*$`	`1885-03-06/1885-08-01`
	`1948-05-26 - 1948-05-26`

Patterns for timestamps	

`^\s*YEAR-\d{2}-\d{2}[T\s]\d{2}:\d{2}:\d{2}Z?[^?\w]*$`	`2008-10-03 13:17:16`
	`2011-11-22T22:00:50Z`

[2] see: https://developer.mozilla.org/en/docs/Web/JavaScript/Guide/Regular_Expressions