1

Turning Copyright on its Head?

Mass digitization is commonly conceived of as the conversion of copyright works in digital format on an industrial scale. It has recently become prominent as the activity by which books, journals, photographs, sound recordings, and films are digitized in bulk to feature in the collections of online archives, repositories, digital libraries, search engines, and data aggregators. Examples include private commercial programmes like Google Books; not-for-profit ventures, such as the Internet Archive and its affiliate, the Open Library; large-scale collaborative projects, such as the HathiTrust; publicly funded projects like Europeana; as well as a number of initiatives launched by archives, libraries, publishers, and private companies for both commercial and non-commercial goals. While the joint effect of all these endeavours promises to enhance access to culture and knowledge, the lawfulness of the activities underlying mass digitization remains largely uncertain. Envisioned as the gateway to Europe’s cultural heritage, the Europeana project has so far progressed conservatively, most notably for reasons of copyright clearance. Google Books has been challenged by lawsuits in Europe and the United States, where a class action has been pending since 2005. Even though Google has concluded agreements with publishers to settle lawsuits or to agree on the digitization of their catalogues, legal uncertainty over the permissibility of the activities underlying Google Books remains. The ambiguity as to the lawfulness of mass digital activities has attracted a wide scholarly debate and judicial attention. In its decision of 10 October 2012, the District Court of

4 <http://www.europeana.eu>.  
5 For a comprehensive account of small-scale digitization projects, cf Kwong Bor Ng, Digitization in the Real World (New York, NY: Metropolitan New York Library Council, 2010).  
7 Authors Guild Inc v Google Inc, 05 Civ 8136 (DC), 37. See also American Society of Media Photographers Inc v Google Inc, No 10 Civ 2977, filed 7 April 2010, and Editions du Seuil et autres v Google Inc et France, Paris District Court, 3rd Chamber, 2nd Section, 79 PTCJ 226, 18 December 2009 (decision appealed; the case was dismissed in June 2012 based on an agreement between the French Publishers Association (SNE) and Google; see <http://googlepressfr.blogspot.it/2012/06/le-syndicat-national-de-edition-sne-et.html>).
New York found that all the activities carried out in the context of the HathiTrust digitization project are permitted ‘fair use’ of the defendants’ copyright protected works. Marking a significant victory for libraries engaged in digitization of their collections, this decision addresses some key issues of mass digitization while leaving others unanswered.

From a copyright perspective, mass digitization is just the visible side of a recent phenomenon of the digital age, whereby bulk copying and processing of copyright content has become the core business of many services as well as research-oriented activities. The various digitization projects share the vision of creating a ‘universal library’ that could include virtually all the world’s knowledge. Being as old as the internet, this vision started materializing in the mid-1990s, when libraries and not-for-profit organizations initiated the digitization of their collections. These projects, however, are largely diverse in terms of dimension, function, and objectives, and their diversity is only indicative of the fact that mass digitization involves activities that exceed the purpose of building a digital library. In this respect, the legal issues arising from mass digitization projects have common features with those emerging from other activities of the web economy, such as the conversion of newspaper articles in digital format to offer news services, the systematic extraction and reutilization of the contents of various databases to make price comparisons, or the aggregation of digital content into web results. What brings all these activities together is the ‘mass’ effect, namely the fact that copying in large quantities is the prerequisite for making use of the information contained in copyright works for various business purposes, without the work even being displayed to the public. These technologies have been described as ‘copy-reliant’, since they operate on the basis of routine, automatic, and indiscriminate copying of works en masse, for which no authorization can be realistically sought. The mass effect is manifested by a shift that seems to turn copyright on its head: while copyright is a system of *ex ante* permissions, mass digitization comes with a compelling demand to revert copyright into an opt-out regime. Clearing rights for individual uses is not feasible or may be prohibitively costly. As we will show later, the use of so-called orphan works—ie works still in copyright for which no rightholder has been identified or located after a diligent search—as well as the Google Books approach to digitization, are indicative of this shift.

Even though the outcome of the activities involved in mass digitization promises to be extremely valuable to the public at large, is this potential public utility enough to overcome established copyright norms, or does copyright represent an impediment to a future of universal access to knowledge?

---

8 *Authors Guild Inc v HathiTrust*, No 11 Civ 6351 (HB), 2012 US Dist. The case has been appealed.
Mass digitization projects aspire to build a twenty-first-century universal library that will ideally be accessible by everyone and will last forever. To achieve this mission, these projects engage in transplanting the cultural heritage of humankind, as deposited in books and other physical carriers, into the digital networked environment.

The first venture to envision the creation of a digital copy of ‘all the world’s books’, as well as of other content in the public domain (such as sound recordings, images, and films) was the Internet Archive. Founded in 1996, this not-for-profit digital library runs several projects, such as the Wayback Machine, a service that allows browsing into archived copies of web pages, and the Open Library, which hosts public domain books in fully readable and downloadable formats. The Carnegie Mellon Million Book project, which was launched in 2001 and joined its efforts with the Internet Archive, declared the ‘long-term objective’ of ‘capturing all books in digital format’. To date, the Internet Archive offers permanent access to over three and a half million books and other texts, to over one million audio recordings, and to half a million moving images, all of which are in the public domain. The Wayback Machine makes available over 150 billion archived web pages.

With Google entering the business of digitization of books in 2004, digitization was given impressive and—so far—unrivalled power. Google Books, earlier known as Google Print and Google Book Search, is pursuing the company’s corporate mission to ‘organize the world’s information and make it universally accessible and useful’. While books are certainly part of ‘the world’s information’, the ‘useful’ features that Google aims at unveiling have little to do with the conventional utility that books embody. Featuring in Google’s search engine results alongside any other ‘resource’, books become searchable on a word-by-word basis and are increasingly integrated with other services and projects developed by the company. Books in the public domain are available to read and download, and books whose rights have been cleared are available depending on the agreements with the relevant rightholders. All other books are displayed either in short extracts, the so-called snippets, or are made available in a metadata-only view. In February 2012, the whole corpus allegedly contained over 20 million books, and snippet view, in response to search

---

10 See eg Matt Simon, ‘Alexandria 2.0: One Millionaire’s Quest to Build the Biggest Library on Earth’, Wired, 20 August 2012.
11 <http://www.archive.org>. The project to create ‘one web page for every book ever published’ was launched by Internet Archive in 2006 under the name Open Library, see <http://openlibrary.org>.
Turning Copyright on its Head?

queries, was available for over four million books in English. One might argue that Google has simply engaged more efficiently on a route paved by others. Yet, what is different with Google Books is not just the scale of the project, but also the quality of ‘mass’ in the digitization process. First, Google does not aspire to create specific collections of digital copies: it aims at digitizing everything—namely, every book ever printed in every language. Human intervention in selection and arrangement of the material, as well as in the quality control of the digital copies and their indexing, is reduced to a minimum. Second, Google digitizes every book, without running any selective process, and without clearing rights on an individual basis: it has entered into licensing agreements with some of the biggest libraries of the world to scan their corpuses irrespective of copyright status. Legally speaking, this is the most disruptive and unique feature of the project, whose modus operandi has been rightly described as ‘scanning first and asking questions later’. Rightholders are merely given the opportunity to opt out from having pages or snippets of their works publicly displayed. Third, while working in partnership with the major libraries of the world, Google Books is and remains an entirely for-profit initiative. Even though there have been other privately owned initiatives in the past, Google is to date the only private player in mass digitization.

In September 2005, the Authors Guild of America, together with five publisher members of the Association of American Publishers (AAP), sued Google for copyright infringement. Google offered to settle the lawsuit for US$125 million and

15 As reported by Google's Engineering Director Daniel Clancy in its Declaration in support of Google Inc's opposition to plaintiff's motion for class certification, Authors Guild Inc v Google Inc, No 1:05-CV-08136, Document 1004, filed 8 February 2012, 2.
17 Dirk Smillie, ‘Has Google Already Won The Book War?’, Forbes, 26 August 2009. In-copyright books are digitized but are not displayed to the public, apart from short excerpts in response to search queries ‘even if the search term appears many times in the book’ (Petition for Permission to Appeal Class Certification, Authors Guild Inc v Google Inc, No 1:05-CV-08136, 14 June 2012, 5). Google also takes measures to prevent users from viewing several contiguous snippets, and conducts ‘human review of every individual book placed in snippet view to determine whether the book is appropriate for snippet view’ (Petition for Permission to Appeal Class Certification, 5). Reference works such as dictionaries and cookbooks are excluded from snippet view.
18 In the FAQ of Google Books, these are the instructions given to authors: ‘You can exclude your books from the Library Project. If you’re a partner, you can do this from within your account. Otherwise, you’ll need to provide us with a small amount of information about yourself, identify yourself as the rights owner and let us know which books or serials to exclude.’ See <http://support.google.com/books/bin/answer.py?hl=en-GB&answer=43756>, last accessed September 2012.
19 A project launched by Microsoft in December 2006, called Live Search Books, was dismissed two years later after having digitized 750,000 books.
on October 2008 the parties announced a Settlement, which was later amended\textsuperscript{21} and eventually rejected by the court in March 2011.\textsuperscript{22} Summarizing the reasons for this rejection, Chin J held that ‘[w]hile the digitization of books and the creation of a universal library would benefit many, the [Amended Settlement Agreement] would simply go too far.’\textsuperscript{23} To be precise, by virtue of the Agreement, where licensing would not be possible for practical reasons, there would be no need to clear rights. Hence, an opt-out rule was meant to be introduced, which would reverse the established requirement to obtain licences before engaging in a restricted act. This shift would result in a situation where authors and rightholders would have the burden of removing their works from the corpus, if they did not wish to permit this use, and where a plethora of orphan works would have to remain in the corpus under Google’s ‘custody’. The judge urged that the Settlement be revised from ‘opt-out’ to ‘opt-in’, and set a status conference for discussion of the next steps. Even though a series of status conferences were held throughout 2011, delegates did not reach consensus as to an amended ‘opt-in’ Settlement.\textsuperscript{24}

Most of the controversy has to do with the fact that mass digitization seems to have ‘turned copyright on its head’. This point has been very frequently made in the fairness hearing of the Google Books case in several contexts. Professor David Nimmer, for instance, acting on behalf of Amazon, was the one to initially raise this point with regard to the fact that the Settlement Agreement envisaged a situation whereby no authorial consent would be required to proceed with a series of activities that normally require authorial permission—first of all copying the work as a whole.\textsuperscript{25} The Deputy Assistant of the Attorney General of the US Department of Justice, William Cavanaugh, has also repeated this view, by arguing that the proposed Agreement would turn copyright on its head on similar grounds, namely that it would eviscerate the requirement of prior authorial consent, a requirement that traditionally pertains to copyright as a system of permissions.\textsuperscript{26}

This is not the only case where mass digitization changes copyright from a system of giving or refusing permissions into a system where the authorial entitlement takes effect \textit{ex post}, by seeking the removal or exclusion of a work from unwelcome uses. Another example comes from the litigation between the Authors Guild of America and the HathiTrust. Founded in 2008, the HathiTrust is a partnership of over seventy US research institutions and libraries working together to create a

\begin{itemize}
\item \textsuperscript{21} Settlement Agreement, \textit{Authors Guild Inc v Google Inc}, No 1:05-CV-08136 (28 October 2008. The Amended Settlement Agreement was filed on 13 November 2009).
\item \textsuperscript{22} \textit{Authors Guild Inc v Google Inc}, 770 F Supp 2d 666 (SNDY 2011).
\item \textsuperscript{23} \textit{Authors Guild}, at 669.
\item \textsuperscript{24} An ‘opt-in’ agreement was reached between Google and the French publishers in June 2012; cf \texttt{<http://googlepressfr.blogspot.it/2012/06/le-syndicat-national-de-ledition-sne-et.html>}.\textsuperscript{24}
\item \textsuperscript{25} David Nimmer, Deputy on behalf of Amazon.com, at the Fairness Hearing for the Google Book Settlement, \textit{The Authors Guild et al v Google Inc}, No 1:05-CV-08136, 18 February 2010, 46, 47.
\item \textsuperscript{26} William Cavanaugh, Deputy Assistant Attorney General, US Department of Justice, at the Fairness Hearing for the Google Book Settlement, \textit{The Authors Guild et al v Google Inc}, No 1:05-CV-08136, 18 February 2010, 124. The point that the Agreement turns copyright on its head was also made during the fairness hearing by Cynthia Arato from the New Zealand Society of Authors (at 81 with regards to the works of foreign authors), and Bruce Keller, attorney of the plaintiffs (at 161, 163).
\end{itemize}
large-scale collaborative repository of digital content, including content digitized by Google, the Internet Archive, and Microsoft, and also works that are locally digitized by an initiative of partner libraries. As of November 2012, the HathiTrust comprised over ten million volumes, of which over 3.2 million are in the public domain. The HathiTrust allegedly hosted orphan works too. The Authors Guild, along with two foreign authors groups and a number of individual authors, sued HathiTrust. They challenged HathiTrust’s digitization efforts to enable full-text searching, preservation, and access for people with print disabilities, and their plan to make orphan works available to faculty, students, and library patrons. The case was decided in favour of the defendants in October 2012. The District Court found that the HathiTrust’s activities were fair use, and that the orphan work issue in the complaint was not ripe for adjudication.

Institutional users like libraries and archives are also involved in mass digitization, but do not engage in copying of in-copyright works. In 2006, the European Commission announced the decision to promote a European counterpart of Google Books, by joining the efforts of all European cultural institutions in a single framework. The portal of the European ‘digital heritage’, Europeana, was launched in 2008. This is an entirely publicly funded initiative that shares only some of Google’s ‘mass’ attributes. It aims at hosting all of Europe’s cultural heritage by creating a universal platform of interoperable digital objects. Unlike Google Books and the HathiTrust, however, it operates a selection of the material through its partner institutions; and, most importantly, it does not digitize in-copyright works without permission. What is subject to digitization are mainly works in the public domain, works whose use has been licensed, or the metadata over works that are still under protection. As of April 2012, Europeana offered access to over 23 million objects digitized by partner institutions, which comprise more than 2,200 cultural and scientific organizations in Europe. In spite of these impressive figures, the project has progressed conservatively so far, especially with regard to Europe’s recent cultural heritage, which includes in-copyright works. This is mainly due to the complexities involved in defining the copyright status of works and in clearing rights.

27 <http://www.hathitrust.org/>

28 The HathiTrust consortium—led by the library of University of Michigan—announced on May 2011 the ‘Orphan Works Project’. The aim was to identify the status of books digitized by Google or the Internet Archive (‘MLibrary launches project to identify orphan works’, Press release, 11 May 2011 <http://www.lib.umich.edu/orphan-works/>). The issue with orphan works is discussed extensively in Ch 4.

29 Authors Guild v HathiTrust, No 11 Civ 6351(HB), 2012 US Dist. See discussion of this case in Chs 2 and 4. The court did not reach the merits of the copyright claims concerning orphan works because the HathiTrust had already suspended the Orphan Works Project shortly after the complaint was filed (‘U-M Library statement on the Orphan Works Project’, Press release, 16 September 2011 <http://www.lib.umich.edu/news/u-m-library-statement-orphan-works-project>).


31 <http://www.europeana.eu>.
for works that are presumably in-copyright. As acknowledged in the European Commission’s Digital Agenda for Europe, ‘[f]ragmentation and complexity in the current licensing system also hinders the digitisation of a large part of Europe’s recent cultural heritage. Rights clearance must be improved, and Europeana…should be strengthened.’ One such improvement was the enactment of the Directive on Orphan Works in October 2012. The Directive will enable European cultural institutions—but not commercial enterprises—to digitise and make available a large number of in-copyright works belonging to their collections.

More recent partnership projects include the Digital Public Library of America that was launched in 2010 with the aim of unifying the collections of the main public libraries in the United States, including the Library of Congress. Drawing inspiration from Europeana, this project aspires to link resources from the collections of its partner institutions and, besides access, to enable ‘deep research’ in existing global digital library infrastructure.

From a copyright perspective, mass digitization operates in four directions. The first includes the digitization of public domain works, of works on which copyright does not subsist, and of the metadata of both in-copyright and public domain works. The second covers the use of works as authorized by the relevant rightholders, either through direct licensing arrangements or forms of open licensing. The third direction comprises uses of in-copyright works that are either subject to collective management, as is the Nordic system of extended collective licensing, or covered by special legislative provisions, such as those concerning orphan works and out-of-distribution works. To date, this kind of use is not generalized, although it promises to become prominent in the near future. Finally, there is the use of in-copyright works that is not subject to any kind of authorization from the rightholders or collecting societies acting on their behalf. Projects engaged in this latter kind of use will have to rely on copyright exceptions and limitations, such as the ‘fair use’ defence in the United States. Mass digitization projects like the Internet Archive and Europeana operate in the first two directions. Depending on the availability of legislative solutions in the relative jurisdictions, the activities of these projects may also expand toward the third direction, namely the digitization of in-copyright but out-of-distribution works. Google Books is the only project

---

32 Automatic systems to help establishing the public domain status of works have been developed, although they cannot replace individual searches. See eg the Public Domain Calculator of Europeana: <http://outofcopyright.eu/>.


37 See Ch 4.
that, along with the first two directions, has decidedly embarked along the fourth path too.

The case of Google Books is exemplary of the potential clash between copyright and mass digitization. In many respects, Google’s activity can be seen as carrying out for books what it is commonly done for any other resources on the web, namely ‘making copies and indexing the text to make it searchable’. Yet, copyright on the internet works in peculiar ways, which do not squarely correspond to the norms applicable in the offline environment, where wholesale copying for commercial purposes requires prior permission or authorization. As rightly observed, ‘through its scanning program, Google had hoped to impose the copyright norms of the digital world onto the analogue world’. The core question arising from the encounter of copyright and mass digitization is whether this imposition adheres to legal norms and principles.

Digitization as ‘Moral Imperative’

Whereas traditionally the reproduction right is considered to be the ‘core of copyright’, mass digitization projects engage in bulk copying of copyright works. This involves scanning of works that have not been digitally born, and their processing by optical character recognition (OCR) software.

Does the goal of enabling universal access to knowledge and enhancing social welfare exempt these activities from infringement of the reproduction right? This would not serve as a viable defence under European copyright law, since this kind of copying cannot convincingly fall within any of the specific activities that are statutorily excluded from infringement, despite the fact that there is a growing body of case law that holds in favour of its permissibility. Nonetheless, mass digital projects may find refuge under the fair use defence of US copyright law. The first factor of fair use, often referred to as the ‘transformative’ factor, has been recently expanded to cover technologically empowered acts of verbatim

---

39 Testimony of David Drummond, Senior Vice President of Corporate Development and Chief Legal Officer, Google Inc, Before the House Committee on the Judiciary, Hearing on ‘Competition and Commerce in Digital Books’, 10 September 2009. 2. David Drummond also adds: ‘[t]oday it is understood that the act of copying the web to index it is a fair use under our nation’s copyright laws. Fair use is the very reason search engines exist’ (‘Competition and Commerce in Digital Books’). However, in his opinion rejecting the Google Books Settlement Agreement, Judge Chin does not appear to see the proclaimed analogy between books and web resources. Citing the opposition of Microsoft’s Counsel, Thomas Rubin, he opines that ‘[w]hile its competitors went through the “painstaking” and “costly” process of obtaining permissions before scanning copyrighted books, “Google by comparison took a shortcut by copying anything and everything regardless of copyright status.” Authors Guild Inc v Google Inc, 770 F Supp 2d 666, 679 (SNDY 2011).

40 Vaidhyanathan, The Googlization of Everything (and Why We Should Worry), 167.


43 These cases are discussed in detail in Ch 2.

44 USC, § 107.
reproduction of works, insofar as these activities are able to give the secondary use a new purpose or new function, and to the extent that they serve the public interest.\(^{45}\)

Mass digitization promises to be socially beneficial by reviving our cultural heritage and making it accessible to anyone in the world. Brewster Khale, the founder of the Internet Archive, puts this promise of universal access in evocative words: ‘[w]e can provide all the works of humankind to all people of the world. It will be an achievement remembered for all time, like putting a man on the moon’.\(^{46}\) While this is certainly a goal worth pursuing, the ways of interpreting and fulfilling it are anything but straightforward. The vision of Brewster Khale differs in several ways from that of Google, for instance, despite the fact that concepts and metaphors almost coincide.\(^{37}\) The notion of ‘access’ itself is ambiguous, since the enhancement of access through digitization entails, at the same time, a change in the way that the content is used. When transplanted in the digital environment, texts, images, and sounds are given a new meaning that goes beyond their renewed accessibility. Moreover, a simple glance at the digitization projects that we mentioned earlier indicates that ‘access’ is not a neutral and self-evident concept. With respect to books, for instance, there are projects where the search is based on metadata, as in libraries; others that allow search-inside on single books; and finally others that go further and link the search with other web services, and so on. A digitized book differs substantially from its physical counterpart, with the possibilities enabled by digital processing having just begun to materialize.

By eliminating the boundaries that exist between tangible objects, digitization’s ultimate goal is to create—as techno-enthusiast Kevin Kelly has put it—‘a single liquid fabric of interconnected words and ideas’.\(^{48}\) In such a ‘fabric’, all books and all pieces of information will allegedly become a single book—namely ‘one very, very, very large single text: the world’s only book’.\(^{49}\) Whereas in the analogue world, books are just ‘isolated items, independent from one another’ lying on library shelves,\(^{50}\) in the universal interconnected environment it is expected that ‘no book will be an island’.\(^{51}\) This is because:

each word in each book is cross-linked, clustered, cited, extracted, indexed, analyzed, annotated, remixed, reassembled and woven deeper into the culture than ever before. In the new world of books, every bit informs another; every page reads all the other pages.\(^{52}\)

Consider, for instance, Dan Brown’s mystery-detective novel, the *Da Vinci Code*. The story starts at the Louvre museum, where the dead body of its curator is found in the pose of the Vitruvian Man with a series of symbols and codes, like

45 The issue of technological transformative use is discussed in Ch 2.
48 Kelly, ‘Scan This Book!’, 5.
49 Kelly, ‘Scan This Book!’, 5.
50 Kelly, ‘Scan This Book!’, 3.
51 Kelly, ‘Scan This Book!’, 3.
52 Kelly, ‘Scan This Book!’, 3.
a Fibonacci number sequence. In an empowered reading environment, readers may be able to unveil the plot of the book by being virtually—and not only mentally—directed inside the Louvre, with active links to the museum’s website as well as to maths websites with detailed analysis on what a Fibonacci sequence is. Readers can seek the automated translation of the text while virtually ‘walking’ in the neighbourhoods where symbology expert Robert Langdon is carrying out his investigations (through active maps or satellite-enabled street views). The Da Vinci paintings could be studied in great detail through links to the online art collections and more information could be found by referencing to the World Wide Web. Associations can also be made to websites selling relevant ‘products’ beyond the other works of Dan Brown, such as the 2006 Columbia Pictures film that is based on the book, the 2005 parodic version *The Va Dinci Code* by Adam Roberts, and perhaps to thematically relevant books, such as Baigent and Leigh’s *The Holy Blood and the Holy Grail*.

This environment can be generated by the cumulative effect of users engaged in searching, reading, and sharing content online, while the platform hosting the works may track down reading habits and preferences, and generate personalized reading environments.

This unprecedented empowering of the reading experience with the accumulation of all the world’s knowledge in electronic format, and with all possible associations that can be made by exploring the computational potential of this knowledge, is what mass digitization promises to achieve. This achievement is frequently put forward as conclusive evidence of public utility. In this vein, having all books and other copyright content searchable online appears to be the necessary condition to make them part of cyber-space, with all its connotations as a breeding ground for democracy and progress in the information society. It is no surprise, therefore, that a purely technical operation such as scanning books has even been seen as a ‘moral imperative’ and a ‘moral obligation’.

However, this endeavour raises legal and policy concerns that exceed the microsphere of the interests of the owners of copyright works. Even though existing copyright norms seem to hinder the fast development of mass digital projects, copyright was never meant to function as an impediment to the free access to knowledge. Both from a principle-based perspective and by reference to its historical evolution, copyright is a juridical order aimed at promoting ‘the advancement of learning’, and this is achieved by protecting the dignity and the interests of the authors while securing the right of the public to read and appropriate what the authors present.

---

53 Note that there has been a case for copyright infringement in the UK regarding the *Da Vinci Code* and this book, where Smith J found that the *Da Vinci Code* was not infringing the copyright of the plaintiffs. See *Baigent v Random House* [2007] EWCA Civ 247.


55 Peter Branley, director of technology for the California Digital Library, quoted in Kelly, ‘Scan This Book’, 8.


57 An Act for the Encouragement of Learning, by Vesting the Copies of Printed Books in the Authors or Purchasers of such Copies, During the Times therein mentioned, 1710, 8 Anne, c 19.
Briefly, copyright is a legal recognition of the communication between authors and their public as a paramount expression of human coalescence.\textsuperscript{58} How is this function of copyright to be understand in the mass-digital age?

Preservation, Access, Computation

To achieve the objective of a universal library—or a universal bookstore\textsuperscript{59}—with all the social benefits this entails, mass digitization carries out a threefold purpose, namely it aims at the preservation of, access to, and computation with works. What seems to be peculiar with mass digitization, however, is that it comes with a sense of urgency and compulsiveness. As already indicated, the modus operandi of Google, for instance, has provoked a hastening of all digitization programmes worldwide, and a rush to achieve the goal of transplanting ‘all the world’s books’ and all documents of the world’s cultural heritage into digital format as quickly as possible.\textsuperscript{60} Digitization proclaims to be a—if not the—compelling trend of the information age, and its ‘mass’ quality comes as an inevitable consequence. But which are the reasons for this urgency?

The first argument that is generally put forward to explain the urgency to digitize printed material is the necessity of preserving cultural heritage. As a matter of fact, public libraries and archives have been digitizing their collections for preservation purposes for over twenty years already, even before the advent of the World Wide Web and the promise of the fully interconnected online world. Acknowledging the significance of this activity, most copyright laws provide specific exemptions for preservation and archival purposes. Most notably, this is done to overcome the risk that libraries or single collections of unique materials may be damaged or even destroyed by natural calamities or human negligence.\textsuperscript{61} In this light, digitizing all the world’s knowledge does not differ from what has been always done in the past to protect works from the risk of getting lost—namely reproducing them in many copies and storing these copies in safe places.

The European Copyright Directive permits copying for preservation purposes of all works, irrespective of their status and without a requirement of prior authorial

\textsuperscript{58} Cf Maurizio Borghi, ‘Copyright and Truth’, \textit{Theoretical Inquiries in Law} 12/1 (2011) 1.

\textsuperscript{59} The most advanced post in the rush to digitize all the world’s books, namely Google Books, is a fully-fledged commercial venture and not a ‘library’ in any conventional meaning. Cf Pamela Samuelson, ‘Google Books is Not a Library’, \textit{The Huffington Post}, 13 October 2009.

\textsuperscript{60} The concept of ‘all the world’s books’, which is frequently used in the context of mass digitization projects, is unsettled. At the outset of the Google Books project, it used to correspond to about 30 million books. The number soon increased to 50 million. Now the latest Google’s account speaks of 129,864,880 unique items. See Leonid Taycher, ‘Books of the World, Stand Up and Be Counted!’, in \textit{Google Blogspot}, 5 August 2010. For a critique of Google’s estimates see Nick Pole ‘The Cost of Digitising Europe’s Cultural Heritage’, A Report for the Comité des Sages of the European Commission, the Collections Trust, November 2010, 4.

Turning Copyright on its Head?

consent. Article 5(2)(c) of the Copyright Directive provides for an exception in favour of archives or publicly accessible libraries, educational institutions, or museums to carry out specific acts of reproduction for non-commercial purposes.\(^{62}\) This Article, however, does not impose a mandatory provision, and Member States have been at liberty to decide whether or not to transpose it into national law. Some Member States have not implemented this exception, and others have interpreted it in a very narrow fashion.\(^{63}\) For instance, the UK Copyright Act does not allow copying of sound recordings, broadcasts, or films for preservation purposes,\(^{64}\) and, for these categories of works, authorial consent is required to make lawful copies.\(^{65}\) In the United States, there is a special exemption for the limited circumstances under which libraries are permitted to reproduce and distribute copyright works for purposes of preservation, replacement copies, and the fulfilment of patron requests. Following its revision in 1998, when the Digital Millennium Copyright Act was passed, the US Copyright Act, in section 108, permits libraries to create copies of works in ‘machine-readable’ or digital format for preservation purposes and as replacements for published works. There are two limitations to this exception, however: there can be no further distribution of the digital format; and the digital copy cannot be used ‘outside the premises of the library or archives’. What is more, section 108 requires that the reproduction is ‘made without any purpose of direct or indirect commercial advantage’. The *HathiTrust* decision, however, has made clear that section 108 does not preclude the availability of the fair use defence of section 107.\(^{66}\)

The second argument which is commonly mentioned when explaining the ‘imperative’ character of mass digitization is that the availability of content in digital format will eventually enhance access to information and knowledge, especially in the context of the creation of indexes and search tools. This argument, however, does not find unequivocal support in copyright laws.\(^{67}\) As copyright currently stands on a worldwide basis, the lawfulness of copying with a view to provide—or otherwise enable—access is dubious, because it involves an active engagement with

\(^{62}\) Directive 2001/29/EC, Art 5(2)(c): ‘Member States may provide for exceptions or limitations to the reproduction right provided for in Article 2…in respect of specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage.’ See Ch 3, 57–58.

\(^{63}\) IViR, ‘Study on the Implementation and Effect in Member States’ Law of Directive 2001/29/EC on the Harmonisation of certain aspects of copyright and related rights in the Information Society’, Part 1, 46–47, February 2007. See, however, the broad scope that has been given to Art 5(2)(c) in Norway, where it covers the making of copies for non-commercial computational analysis (‘text mining’). See Ch 3.

\(^{64}\) See Copyright, Designs and Patents Act 1988 (CDPA), s 42, which applies only to literary, dramatic, or musical works, illustrations that accompany such works, and typographical arrangements of printed editions.

\(^{65}\) It should be borne in mind that there are multiple copyrights that need to be cleared. Finding the respective rightholders and obtaining permission is not only time-consuming but could also be unfruitful. As the Duke Centre has observed, the issue of orphan works slows down the preservation initiatives but does not affect the decomposition of these works. Duke Center for the Study of the Public Domain, ‘Access to Orphan Films’ (March 2005) 3.

\(^{66}\) *Authors Guild Inc v HathiTrust*, 12–13.

\(^{67}\) Note, however, that in the context of orphan works, digitization for enhancing access will be permitted under certain conditions laid down in the Directive 2012/28/EU. See Ch 4.
a series of restricted activities, namely reproduction, display of works, and making those works available, in full or in part. Unless a work is in the public domain, there is a need to clear rights for providing access to it. However, mass digitization enhances access to knowledge in a broad sense, which does not refer solely to the act of making works available to the public. Here, an important distinction has to be made between activities that are directed at providing access and activities that are merely supportive or ancillary to access. No permission is required, for instance, to make indexes or metadata that enable or facilitate the retrieval of works—even copyright-protected works—and of the information contained therein.68

The conundrum with digital technology, however, is that most of these access-ancillary activities require the making of copies. This conundrum is exacerbated by mass digitization. To enable the retrieval of information contained in works, a digital project must engage in wholesale copying—an act that is prima facie restricted to the right-holder. Moreover, the fact of making small portions of works publicly available on a systematic basis, as part of the process of information retrieval, may exceed the boundaries of a mere access-ancillary activity. Google Books is so far the only venture that provides this service for in-copyright literary works. Because in mass digitization the distinction between access provision and access-supportive activities is blurred, there is legal uncertainty as to the boundaries of permissible uses. Although Google maintains that copying for purposes of indexing and search is permissible under both US law and the laws of European Member States, it has also, at the same time, prudentially entered into content-usage agreements with book publishers on both sides of the Atlantic.69

The same legal uncertainty surrounds the third ground for mass digitization, namely digitization for computation purposes, which tends to become a standard practice in both not-for-profit and commercially oriented ventures. Mass digitization is not primarily—and certainly not only—about preservation or access enhancement: in contrast to conventional libraries, works are digitized to be further used in a different way. Copying is just a preliminary condition for securing works in their full computability.70 Institutional users may also be interested in exploiting the so-called ‘computational potential’ of digital archives.71 By this we refer to the possibility of using the content of works for extracting information through computational analysis and data mining. In the Da Vinci Code example we gave earlier, the empowered reading environment is partly built by means of

---

69 See the Agreement with French Publishers Association announced on 11 June 2012 (http://googlepressfr.blogspot.it/2012/06/le-syndicat-national-de-ledition-sne-et.html) and the Settlement Agreement with the Association of American Publishers announced on 4 October 2012 (http://www.publishers.org/press/85); see also, in the same vein, the Agreement with the Belgian news publishers to display snippets of newspaper articles in Google News (‘Partnering with Belgian news publishers’, Google Europe Blog, 12 December 2012 (http://googlepolicyeurope.blogspot.gr/2012/12/partnering-with-belgian-news-publishers.html)).
the automated processing of the content of the work, such as linguistic analytics and text mining. For instance, a link to relevant products, such as books and films, can be generated either as an effect of users’ interaction, or through an automated process of text mining, or by a combination of both. However, computation on digital works does not by necessity entail their display to the public and may be unrelated to access enhancement. As has been straightforwardly pointed out in the context of Google Books:

different from our current understanding of a library, this corpus of works would not be made available for the purpose of reading the works. Instead, this group of works is intended to be made available for computational analysis on works.72

As a matter of fact, the whole corpus ‘cannot be read by a human’, 73 and this ‘impossibility’ opens up a variety of new fields of research at the intersection between information technology and social sciences, aimed at replacing humans in the very task of reading and understanding books. In this context, computers discover information that human intelligence can never extract, such as quantitative and qualitative data on ‘cultural trends’ over centuries and across languages, migration of ‘ideas’ from one place to another, evolutions of linguistic phenomena, influences of one author on another, and vice versa.74 In turn, this information is linked back to the individual item—book, image, sound recording, video—in order to create an empowered reading environment, from which data about reading habits and, in general, behaviours associated with the experience of content are extracted and processed. It is no surprise to read the following reported words of an anonymous Google engineer: ‘we’re not scanning all those books to be read by people. We’re scanning them to be read by Artificial Intelligence’.75 The use of the verb ‘to read’ in this last sentence is not just geek slang. Books are actually ‘read’ by artificial intelligence. This does not only take place in the sense that viewing books in digital format is necessarily mediated by computer programs and applications. Computers ‘read’ books in the straightforward sense that, by supplying a surplus of computational power, they are capable of extracting information and value that cannot be obtained manually.76 This feature of mass digitization is the one that presents the greater challenge to the established copyright norms. It is also

73 Jean-Baptiste Michel et al, ‘Quantitative Analysis of Culture Using Millions of Digitized Books’, Science Express (16 December 2010) 1: ‘If you tried to read only the entries from the year 2000 alone, at the reasonable pace of 200 words/minute, without interruptions for food or sleep, it would take eighty years. The sequence of letters is one thousand times longer than the human genome: if you wrote it out in a straight line, it would reach to the moon and back 10 times over’.
76 Computation is carried out by computers, and it includes information retrieval and value extraction. Retrieval and extraction are the two sides of the same coin, whose organizing principle is the search engine as supreme validating instance of all computations. See Borghi, ‘Knowledge, Information and Values in the Age of Mass Digitisation’, 428–429.
the most prominent and the one that guides the other purposes as well, namely preservation and enhancement of access.

While digitization allows the fast and easy reproduction and distribution of works, the aggregation of digital works in large repositories creates the conditions for the exploitation of their computational value. It is hence not just the scale of mass digital projects that distinguishes them from other activities carried out in the digital world; it is also the fact that the processes involved differ qualitatively. From the standpoint of copyright law and principles, this represents a paradigm shift that is manifested through three main changes. We refer to these changes as the ‘traits’ of mass digitization.

The ‘Head-Turning’ Traits of Mass Digitization

From works to data

For over three hundred years, copyright has covered human interactions, however technology-mediated these may be. Since, however, in mass digitization there is a high degree of automation at all stages—in the sense that human intervention in selection, transformation, display, and even use of content is reduced to a minimum—a shift is marked from machine-mediated to machine-driven uses of works. This means that computers not only function as intermediaries, but they are also employed to carry out uses of works, some of which may have an exploitative nature, either directly or indirectly. In this respect, digital repositories are not just libraries but also silos of data that are available for data mining, ie for automated processing by artificial intelligence. This kind of processing may serve purposes that are not directly related to the display of works to users, and that deviate from the original destination of the work, as conventionally understood. As already mentioned, digitized books may be used to feed search engine algorithms and to improve web services, including advertisement and content personalization. They may also be used as a ‘cultural genome’ to quantify cultural trends over centuries and across languages, as repositories of ‘key ideas’ to be extracted through data mining, as data containers to be mined in order to refine search engine algorithms, or statistical machine translation. While in both the analogue and the digital world works have been used primarily as works—namely as expressions addressed by the author to the public—in the mass digital environment works are primarily used as data.

77 Reportedly ‘[m]uch of the company’s public focus lately has been not on mass digitization but on how to use individuals’ data to create more focused advertising and online browsing’ (Jennifer Howard, ‘Google Begins to Scale Back Its Scanning of Books from University Libraries’, The Chronicle of Higher Education, 9 March 2012).


From an ontological perspective, whereas digitization dematerializes works by skipping the necessity for a tangible carrier, mass digitization de-intellectualizes works and detaches them from the very condition of their being—namely their intelligibility to humans. As when a source code of a computer program is turned into an object code, digital copies are treated as data containers, from which information may be automatically extracted prior to—and even without—being exposed to human eyes and human intelligence. This inevitably raises the question as to whether this shift bears any substantive meaning from a copyright perspective, since it impacts on the very concept of the ‘use’ of works. Is there, and should there be, a distinction between the use of works for purposes of experiencing their content and uses on works for purposes of indexing and search, computational analysis, and data mining?

**From unit to bulk**

As mass digitization projects have to do with ‘masses’ of works, copyright licensing necessarily takes the form of ‘bulk’ licensing. At the beginning of the digital age, the distribution of creative content was visualized through the metaphor of a ‘celestial jukebox’, allowing every user on earth to access any kind of content, and every rightholder to potentially charge for each and every use of the work. However, individual rights management is incompatible with mass digitization; clearing rights in each digitized work may be either economically unsustainable or practically unfeasible. Therefore, the activities underlying mass digital projects must rely either on copyright-permitted acts (eg fair use) or systems of collective rights management—or even both of these mechanisms. However, the inherent structure of copyright as a system of ex ante permissions and exclusive rights is at odds with the bulk effect of mass digitization.

Copyright in not-digitally born works—ie works first created and fixed in a medium other than a digital file—is about to turn, to a large extent, from a system of permissions and ex ante authorizations to an ‘opt-out’ system. To be sure, copyright does not always require that permission be sought from rightholders: certain uses of the work are premised upon collective licensing systems or other forms of ‘paying public domain’; and there are also other activities that are permitted without any requirement of authorial consent or prior authorization. These are typically the exceptions and limitations to copyright, which are afforded in certain special cases on the grounds of public policy objectives or fundamental freedoms. They cover activities that are fair and permitted without prior authorial consent. The rightholders can neither license these uses nor do they have a positive entitlement under copyright law to exclude others from carrying them out.

---

What seems to be the case with mass digitization, however, is that an ‘opt-out’ character has gradually started to be appointed to copyright with regard to activities that normally require prior authorization. Google Books has overcome the problem of seeking licences directly from the relevant rightholders by forming agreements with the world’s largest libraries instead, and leaving rightholders with the option to opt out. To an extent, this approach has recently started to feature in statutory language, too, such as in the French law on the digital exploitation of unavailable books. 83

From decentralized to centralized

Mass digitization comes with a renewed centrality of intermediaries. In the first years of the digital age, the landscape seemed to be dominated by decentralized interaction and the bypassing of intermediaries, for instance in the form of peer-to-peer file sharing or user-generated content. 84 In mass digitization, however, works are mainly accessed through intermediaries, such as digital libraries, online repositories, or publishers’ databases. Although small-scale projects also exist, and distributed digitization efforts play an important role, 85 intermediaries tend to be increasingly bigger in size and fewer in number. What is more, works are stored ‘in the cloud’ of the intermediaries’ servers: they are accessed by end users under determined terms and conditions, and their access in bulk is mostly technologically restricted. Digitization agreements are concluded between big players—mass digital projects on the one side and major publishing companies on the other—by increasingly stepping aside authors.

The Authors Guild v Google case is exemplary of this trend. In its Petition for Permission to Appeal Class Certification, Google made the argument that authors’ interests are not adequately represented as a class by the Authors Guild. This is because the majority of authors (58 per cent according to an internal ‘random survey’) agree to their books being scanned for indexing and search, and ‘benefit economically and in other ways from the Google Books project’. 86 In October 2012, the co-plaintiff publishers announced the conclusion of a separate settlement agreement with Google and their decision to abandon the lawsuit. 87 With

---

83 Cf Ch 4, 90–91.
86 Petition for Permission to Appeal Class Certification, Authors Guild Inc v Google Inc, No 1:05-CV-08136, 14 June 2012, 20.
87 ‘Publishers and Google Reach Settlement’, Press release, 4 October 2012 <http://www.publishers.org/press/85>. Commenting on this agreement, James Grimmelmann observed that ‘this does exacerbate the publisher–author tension’, since now ‘Google is going to increasingly use the consent of the publishers as an argument that the authors don’t even speak for copyright owners’ (reported in Timothy Lee, ‘Publishers abandon fight against Google book scanning’, Ars Technica, 4 October 2012 <http://arstechnica.com/tech-policy/2012/10/publishers-abandon-fight-against-google-book-scanning>).
this agreement, the role of intermediaries has gained a renewed centrality, fully encompassing authorial concerns or users interests.

Centralized control and access to cultural heritage raises issues of user freedom and cultural pluralism, and seemingly goes against the moral impetus of a truly democratic, non-elitist society, which inspires the whole vision of mass digitization. It is questionable both whether this is an inevitable side effect, and whether decentralized projects will be viable within the current copyright framework. There is a risk that monopolies with access to culture may be created, and these could have an impact on copyright via the creation of new, perhaps *sui generis*, entitlements.

***

The technological shift occurring with mass digitization impacts on the common copyright perception of the concept of *use*, on the meaning and function of the *work* as protected subject matter, on the grounds for entitlement to *rights*, and on the scope of the *exclusivity* of such rights. The bulk effect of the mass digital age penetrates into the core of copyright, unveiling its very nature and purpose. It tells us something new about copyright and its limits. Since technology enables uses of works that may be unrelated to their internal purpose and function as embodiments of the author–user coalescence, we need to reconsider the role of copyright as a regulatory system and a driver of innovation.