Queen Mary Journal of Intellectual Property, Vol. 1 No. 1, April 2011, pp. 21–52

Non-display uses of copyright works: Google Books and beyond

Maurizio Borghi and Stavroula Karapapa*
Law Lecturers, Brunel University Law School

With the advent of mass digitization projects, such as the Google Book Search, a peculiar shift has occurred in the way that copyright works are dealt with. Contrary to what has so far been the case, works are turned into machine-readable data to be automatically processed for various purposes without the expression of works being displayed to the public. In the Google Book Settlement Agreement, this new kind of usage is referred to as 'non-display uses' of digital works. The legitimacy of these uses has not yet been tested by Courts and does not comfortably fit in the current copyright doctrine, plainly because the works are not used as works but as something else, namely as data. Since non-display uses may prove to be a very lucrative market in the near future, with the potential to affect the way people use copyright works, we examine non-display uses under the prism of copyright principles to determine the boundaries of their legitimacy. Through this examination, we provide a categorization of the activities carried out under the heading of 'non-display uses', we examine their lawfulness under the current copyright doctrine and approach the phenomenon from the spectrum of data protection law that could apply, by analogy, to the use of copyright works as processable data.

Keywords: *digital copyright, digitization, digital libraries, non-display uses, Google Books Settlement Agreement, transformative uses, data processing*

INTRODUCTION

The concept of non-display use of digital works is newly born in the copyright world. It made its first appearance about two years ago, in the proposed agreement to settle the class action brought by the Author's Guild of the United States against Google for the activities carried out within the Google Books project.[1] According to the purely negative definition put forward in the Agreement, non-display uses of digital works are those carried out without displaying the expression of digital copies of works to the public.[2] In practice, they include any sort of automated processing of works for purposes of data mining, computational analysis on texts and automatic extraction of information. All these forms of processing may in turn enable the development of

* The authors wish to thank Michael Birnhack, Abraham Drassinower, Paul Edward Geller, Peter Jaszi and David Nimmer for helpful comments, critiques and suggestions. An earlier version of this Article has been presented at the Annual Workshop of ISHTIP (the International Society for the History and Theory of Intellectual Property) at American University, Washington DC, in September 2010.

1. Settlement Agreement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (28 October 2008), later amended (13 November 2009) (hereinafter: *Amended Settlement Agreement*).
2. *Amended Settlement Agreement* § 1.94.

Journal compilation © 2011 Edward Elgar Publishing Ltd
The Lypiatts, 15 Lansdown Road, Cheltenham, Glos GL50 2JA, UK
and The William Pratt House, 9 Dewey Court, Northampton MA 01060-3815, USA

tools such as automated translation and targeted advertising software. Although these uses prove to be extremely profitable, and are even likely to be the most valuable part of the Agreement,[3] their legitimacy under copyright law and principles has not yet been subject to scrutiny. Are these uses just copyright-irrelevant on the ground that they do not entail the display of the work to the public? Or should they, alternatively, amount to copyright infringement on the mere ground that they presuppose the repeated making of copies of the work?

In this article, we suggest that non-display uses tend to escape the traditional contours of copyright. This begs for a regulatory framework based on broader legal principles. In particular, we maintain that non-display uses of digital works are better captured under the heading of 'automated data processing', and as such they may be more conveniently examined through the lens of data protection law rather than that of copyright law *per se*. In fact, the analogy between uses of copyright works *as* data-containers and uses of personal data may enable development of a consistent regulatory framework for automated processing of copyright works.

Yet, at the same time, non-display uses offer an interesting insight on an emerging – and so far almost unnoticed – phenomenon of the so-called digital era. While digital technology has dramatically altered the method by which works are *reproduced* and *distributed*, it seems that the time has come when digital technology purely affects the way that works are *used*. Works are no longer massively reproduced and distributed to be appropriated by human beings only, as has so far been the case! The core meaning of non-display uses is clearly captured in the 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.'[4]

In this light, we feel that mass digitization projects, such as Google Books, mark a conceptual shift in the way that digital works are used and dealt with. Whereas the advent of digital technologies leads to the dematerialization of works by skipping the necessity for a tangible carrier, with *mass* digitization works are somewhat de-intellectualized, as they are detached from the very condition of their being, namely their *intelligibility* to humans. Similarly to what occurs when a source code of a computer program is turned into an object code, digital copies are dealt with as mere containers of data, from which information may be automatically extracted and used without being exposed to human eyes and human intelligence. Is this use compatible with the purpose for which works are created and disseminated? Does it fit the legitimate understanding of what a use of a work properly is? Or does it rather involve, somehow, a violation of the intimate nature of a work?

In this article, we examine the legitimacy of non-display uses carried out in the light of mass digitization, acknowledging that the legal issues arising in this context go beyond the Google Books case and entail a more fundamental question on what are the legitimate uses of data which virtually include the heritage of all humankind.[5] Our

3. F von Lohmann, 'Google Book Search Settlement: A Reader's Guide', 31 October 2008, available at <http://www.eff.org/deeplinks/2008/10/google-books-settlement-readers-guide>.

4. Reported in G Dyson 'Turing's Cathedral. A Visit to Google on the Occasion of the 60th Anniversary of John von Neumann's Proposal for a Digital Computer', 24 October 2005, available at <http://www.edge.org/3rd_culture/dyson05/dyson05_index.html>.

5. The project of digitizing all the world's books is a crucial step to achieve Google's corporate mission to 'organize the world's information and make it universally accessible and useful' (<http://www.google.com/corporate>, last accessed September 2010. For a critical analysis see S Vaidhyanathan, *The Googlization of Everything and Why we should Worry* (University of California Press, 2011) 149–73.

purpose is to provide some guidance in defining the legal boundaries of these uses, by reference to the principles of copyright law. The main question that we shall therefore address is the extent to which large-scale digitization projects are entitled to make use of copyright works for automated processing.

This article is divided into two parts. In Part One, we examine what non-display uses are in the context of the Google Book Settlement Agreement, we review the judiciary arguments in the US and Europe on the lawfulness of such uses as well as the rationales supporting and opposing the view that non-display uses should be put under the rightholders' control. In Part Two, we discuss a regulatory framework for non-display uses based on a purpose-driven analysis of the activities that are carried out under this heading; guidance will be sought eventually from the analogy between uses of works as data-containers and the automated processing of personal data.

PART ONE UNDERSTANDING NON-DISPLAY USES

Digitizing a literary work is an act implying two different operations. First, the book (or any other physical carrier) is 'scanned', namely it is copied page by page and converted into a series of digital images. Second, the digital images are processed by optical character recognition (OCR) software, which converts the image into a text file.[6] The first operation creates a *copy* of the work, as with other reproduction technologies such as photography and reprography. The second operation, i.e. OCR-ing, is qualitatively different from mere copying. Its purpose is not just to 'reproduce' the work, namely to multiply copies to be perceived by human intelligence,[7] but to modify the format of the work so as to make the copy readable by computers.[8] Thereby, an object which can be read (only) by humans, such as a book's page or a picture thereof, is turned into a machine-readable entity. As a result, the work's text can be not only *reproduced*, but also *processed* by other applications. For example, the text may become searchable on a word-by-word basis, and information for indexing the work may be automatically extracted.

Mass digitization of books entails both these operations – scanning and OCR-ing – on a large scale. The aggregate result of this ongoing digitization is that a number of resources which were available only in print form, or on other physical carriers, are turned into a huge corpus of machine-readable data. In this part, we examine some of the effects that this alteration to the format of books triggers. We will show that while isolated acts of digitization may have the sole purpose of indexing books and making them searchable by users, the cumulative effect of those single acts on a massive scale creates the conditions for new, unprecedented uses of works – or rather, as we will see later on in this article, for uses *on* works.[9] Below we examine what these 'non-display uses' are in the context of Google Books and of the ongoing case of *Author's Guild v*

6. See *Infopaq International A/S v Danske Dagblades Forening*, Case C-5/08 [2009] ECDR 16, § 18–19.
7. The US Copyright Act defines copies as 'material objects [...] from which the work can be *perceived,* reproduced or otherwise communicated', USC 17 § 101 (emphasis added).
8. The question whether this act of modification is a 'transformative' use under US copyright law will be discussed below (nn 56–65 and accompanying text).
9. See below Part Two, § 2b (nn 112–25 and accompanying text).

Google.[10] Our focus will be on the relevant provisions of the Agreement that have been proposed by the parties to settle the case.[11]

1 Non-display uses in the Google Books Settlement Agreement

In 2004, when Google announced its ambitious project of making all the world's books traceable and searchable online, the attention of users and observers mainly focused on the possibility of locating information quickly and at no cost in an unprecedented corpus of books; this was the *public* side of the project, namely the fact that millions of books were about to be made available online for searching and reading. Google itself presented its project as a comprehensive search engine for full text search of books, rather than as a library in the classic sense of the term.[12] The policy of the project has remained unaltered since its beginning: scanning books, making them fully searchable on the web, then seeking permission from copyright holders to display their books by offering them a place in a partnership programme.[13] In case the rightholders reject this offer, the book remains non-displayed to users, with the exception of short excerpts ('snippets') made available in response to users' search queries. While the project's policy has been described by David Nimmer as 'turn[ing] copyright law on its head'[14] – since copyright is normally about seeking permission *before* engaging in a restricted act, and not *after*! – Google's practice may be justifiable on other grounds.[15] For instance, clearing rights in mass digitization projects may be a very arduous task, sometimes even impossible, as it would raise unprecedented transaction costs.[16] In this light, the exclusivity afforded to the reproduction right seems to be an insurmountable obstacle for the making of a universal searching tool, which is a socially valuable goal.

10. *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136, filed 20 September 2005.
11. Any discussion about the likely outcome of the lawsuit or about the fairness of the proposed Agreement as such remains beyond the scope of this article. For a comprehensive discussion of the Google Books case see P Samuelson, 'Google Book Search and the Future of Books in Cyberspace' (2010) 94(5) Minnesota Law Review; see also P Samuelson, 'Is the Proposed Google Book Settlement Fair?' (2010) 2. *AMI: Tijdschrift voor Auteurs, Media & Informatierecht 50* 60, available at <http://www.ischool.berkeley. edu/~pam>. For a bibliography of academic articles and papers on the *Authors Guild v Google* case see <http://thepublicindex.org>.
12. See eg L Vincent, 'Google Book Search: Document Understanding on a Massive Scale', International Conference on Document Analysis and Recognition, ICDAR 2007, Curitiba, Brazil, September 2007, available at <http://www.vincent-net.com/luc/papers/07icdar_googlebooks.pdf> (describing the main technical features of Google Book Search project). Also see P Samuelson, 'Google Books is Not a Library', *The Huffington Post*, 13 October 2009, available at <http://www.huffingtonpost.com/pamela-samuelson/google-books-is-not-a-lib_b_317518.html> (discussing the differences between commercial ventures like Google Books and traditional libraries).
13. The partnership programme provides a range of schemes, under which a book can be displayed in whole or in part (see Vincent 'Google Book Search', (n 12) 2).
14. Fairness Hearing Transcript, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (18 February 2010) 46.
15. As Neil Netanel puts it, 'Google's scanning and public display of mere snippets of in-copyright books [...] should be fair use' (Netanel, *Copyright's Paradox* (OUP 2008) 212).
16. This is eg the case with the copyright orphans. See in this respect RC Picker, 'The Google Book Search Settlement: A New Orphan Works Monopoly?', (2009) 5(3) Journal of Competition Law & Economics, 383–409, and HR Varian, 'Copyright Term Extension and Orphan Works' (2006) 15(6) Industrial and Corporate Change, 965–80, at 976.

The judiciary history is well known: in September 2005, the Author's Guild Association of America, together with five publishers, sued Google for copyright infringement.[17] In October 2008, the parties announced a settlement, which was later amended and it is still awaiting judicial approval.[18] Under the Agreement authors will have the possibility of contracting with Google only over the display uses of their work. As regards non-display uses, however, their value and potential impact has been, and still is, underestimated both by Google's counterparts in the lawsuits and by observers. Probably this has to do with the fact that most of these uses are still in their infancy, and one can only speculate what their value might be in the future.[19] It was only in one of the objections to the agreement that the issue of non-display uses has been brought up, according to which the plaintiffs 'have ignored both the critical issue of control over Non-Display Uses of their works and the potentially enormous value of such uses'.[20]

This mainly has to do with the definition given to non-display uses in the Agreement, which has been kept deliberately broad and vague. Non-display uses are defined as 'uses that do not display Expression from Digital Copies of Books or Inserts to the public'.[21] Among others, these uses include 'display of bibliographic information, full-text indexing without display of Expression (such as listing the number or location of search matches), geographic indexing of Books, algorithmic listings of key terms for chapters of Books, and internal research and development using Digital Copies'.[22] Under the Agreement, Google is authorized to make non-display uses of all books contained in the corpus[23] and copyright holders have a 'right' to have their books 'removed' from the corpus.[24] Yet, such 'right' is not unlimited in scope and, most importantly, it is limited in terms of time.[25] After a given

17. *Author's Guild, Inc. v Google, Inc.*, No 1 05-CV-08136, filed 20 September 2005.
18. Settlement Agreement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (28 October 2008), available at <http://thepublicindex.org/docs/settlement/settlement.pdf>. The Amended Settlement Agreement was filed on 13 November 2009 (available at <http://www.googlebook settlement.com>. For a view comparing the two versions see <http://thepublicindex.org/ settlement>. The final fairness hearing was held on 18 February 2010 and one year later the agreement was still awaiting approval.) Since the provisions we are referring to in this article remained unchanged from the original to the amended version of the Agreement, in the following we will make reference solely to the Amended Settlement Agreement.
19. 'Imagine the kinds of things that data mining all the world's books might let Google's engineers build: automated translation, optical character recognition, voice recognition algorithms. And those are just the things we can think of today'. See von Lohman (n 3).
20. Objections of Arlo Guthrie, Julia Wright, Catherine Ryan Hyde, and Eugene Linden to Proposed Class Action Settlement Agreement, *Author's Guild, Inc. v Google, Inc*., No 1:05-CV-08136 (2 September 2009) [*Guthrie's Objection*], 12. The arguments put forward in this objection will be discussed in this Part, § 3a.
21. *Amended Settlement Agreement*, § 1.94.
22. Ibid.
23. Ibid, § 2.2
24. Ibid, § 3.4(a) ('Right to remove').
25. 'The right to Remove [...] is limited to requests on or before April 5, 2011 for Removal as described in Section 1.126(a)' [stating that Digital Copies of the Book 'are not accessible to Google [...] other than on *back-up tapes or other electronic back-up storage media*'] *or after April 5, 2011 but on or before March 9, 2012, for Removal as described in Section 1.126(b)* [stating the same as in Section 1.126(a), but with further limitations on the benefit of Participating Libraries and Host Sites]. Thereafter, requests will be honored *only to the extent that the Books have not yet been Digitized* as of the date the request is made; if the Books at issue have already been Digitized, the Rightsholder *may request exclusion from particular*

deadline, copyright holders are entitled to have their books 'excluded' only from some or from all the *display* uses.[26] Nonetheless, a right to exclude is different from a right to remove, in that the author can have her books only inaccessible to users, but not eliminated from the corpus. As Pamela Samuelson explains: 'even if the author asks for her book to be removed from [Google Books], this does not mean that Google will actually purge them from its servers; these books will just be less accessible than if the author hadn't asked for them to be removed'.[27] Practically, this means that authors do not have any means of control over *non-display* uses.

Most of the activities related to non-display uses fall under the umbrella term of '[Google's] internal research and development using Digital Copies'.[28] Broadly speaking, research and development include four ranges of activities involving data and text mining.[29]

First, there are activities of data analysis to improve search algorithms. As a Google engineer has observed, 'the very worst [search] algorithm at 10 million words is better than the very best algorithm at 1 million words'. In this light, Google Books is a formidable increase in the quantity of data that Google's search engine can process, in order to improve the 'core business' of Google, namely its search engine.[30] Note that – at least to date – books forming the corpus of Google books are accessible only from Google's search engine; they do not feature in the search results of other search engines like Yahoo! or Bing.

A second group of activities relate to more sophisticated text analysis. For instance, the content of a book, or of the whole production of a particular author, may be analysed to extract information, which in turn may be sold 'to third-party behavioral advertisers so that they could direct ads to people running searches for that [book or] author's work'.[31]

Further activities relate to analysis of meta-data on the use of digital copies. This data analysis is carried out on the search patterns of users so as to create databases of user profiles. By running users' queries against the database of book content, Google 'could compile dossiers on individual users', which 'would allow Google to personalize

Display Uses [...] *but not Removal*.' Ibid, § 3.5(a)(iii) ('Limitations to the right to exclude'; emphasis added).

26. Ibid, § 3.5(b)(i) ('Right to exclude').

27. Letter of Pamela Samuelson to Judge Chin on behalf of academic authors, in Opposition to the Settlement Agreement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (3 September 2009) 8. See also *Guthrie's Objection* (n 20) 18: 'Removal would not prevent Google from generating additional electronic records relating to Books [...], and using such records to make Non-Display Uses of those works'.

28. *Amended Settlement Agreement*, § 1.94.

29. Data mining is defined in the computer sciences as the 'extraction of implicit, previously unknown, and potentially useful information from data', with the aim of looking for 'patterns in data' (IH Witten and E Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Elsevier, San Francisco 2005) 9). Likewise, text mining is 'the process of analyzing text to extract information that is useful for particular purposes' (ibid, 351). As applied to literary works, text mining is about finding 'structural patterns' in texts, extracting information out of these patterns and combining them with data on the use of work (such as data on works search and access).

30. Objection of Yahoo! Inc. to Settlement Agreement, *Authors Guild, Inc. v Google, Inc.*, No. 1:05-CV-08136 (8 September 2009) 25.

31. *Guthrie's Objection* (n 20) 19.

advertisements to or aim products at specific users'.[32] Information about how many times a work is being searched, by whom it is searched, for how long it is browsed, or about the online communities that flourish around it, may be processed for numerous purposes, including behavioural analysis and personalized advertising.[33] Analysis of books' content, combined with processing of meta-data on the books' uses, may prove to be a powerful means for targeted marketing purposes.[34]

Fourth, non-display uses include a number of activities which are not directly functional in nature. In the language of the proposed Agreement, these latter uses fall under the definition of 'Non-Consumptive Research'. This is defined as 'research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book'.[35] Contrary to what is commonly believed, the main research value of a corpus of millions digitized books does not consist only, and not primarily, in the ease of locating sources and information to be accessed by readers. As it has been straightforwardly pointed out: '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'.[36] As opposed to the traditional research uses of books, this kind of uses which is based on computational analysis does not aim at either displaying the book or appropriating its content. It includes image analysis and text extraction,[37] textual analysis and information extraction,[38] linguistic analysis,[39] automatic translation,[40] and indexing and search.[41]

Under the proposed Agreement, only 'qualified users' will have access to the corpus of Google Books to conduct non-consumptive research,[42] under a detailed

32. Memorandum of Amicus Curiae The Internet Archive in Opposition to Amended Settlement Agreement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (27 January 2010) 7.

33. These uses raise specific privacy issues, which remain outside the scope of this article. For a summary of privacy-related issues, see Privacy Authors and Publishers' Objection to Proposes Settlement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (8 September 2009).

34. See H Barry, 'New Technological Uses', slides presentation, available at <http://www.law.berkeley.edu/institutes/bclt/statuteofanne/pdf/Barry.pdf>.

35. *Amended Settlement Agreement*, § 1.93 ('Non-Consumptive Research').

36. The Stanford University Libraries Amicus Letter in Support of the Settlement Agreement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (8 September 2009) 4.

37. 'Computational analysis of the Digitized image artifact to either improve the image (eg, de-skewing) or extracting textual or structural information from the image (eg, OCR).' *Amended Settlement Agreement*, § 1.93(a).

38. 'Automated techniques designed to extract information to understand or develop relationships among or within Books or, more generally, in the body of literature contained within the Research Corpus. This category includes tasks such as concordance development, collocation extraction, citation extraction, automated classification, entity extraction, and natural language processing', ibid, § 1.93(b).

39. 'Research that performs linguistic analysis over the Research Corpus to understand language, linguistic use, semantics and syntax as they evolve over time and across different genres or other classifications of Books', ibid, § 1.93(c).

40. 'Research on techniques for translating works from one language to another', ibid, § 1.93(d).

41. 'Research on different techniques for indexing and search of textual content' ibid, § 1.93(e).

42. Ibid, § 7.2(b)(vi) ('Use for Non-Consumptive Research'). 'Qualified users' can only be

series of limitations.[43] For instance, these users are allowed to publish their results or to exploit algorithms developed when performing research; they are, however, forbidden from using data extracted from the corpus of Google Books to develop commercial products or services that would compete with those offered by Google.[44] As Pamela Samuelson rightly observes, the limitations over non-consumptive research 'will preclude [researchers] from becoming next-generation entrepreneurs capable of developing radically new information services arising from their non-consumptive uses of the [Google Books] corpus'.[45] In our view, the Agreement designs a troubling deal between Google and the academic and research world: researchers are given access to a hugely valuable stock of data and Google has a monopoly over the incorporation of their research results into commercial products or services.

Computational analysis on a corpus of millions of books opens up possibilities that scholars can now only imagine. Inter-linguistical analysis may help refining techniques of automated translation. 'Associative' and 'taxonomic searching tools' may provide 'access to ideas more or less independent of the exact expressions of ideas across many texts'.[46] Researchers may track and quantify 'influences', 'cultural exchanges' and the absorption of 'cultural values' from different areas of the world and can quantify phenomenon such as 'the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology'.[47] They may also compare and contrast 'contemporary views of events' and 'philosophies from different national and linguistic context'.[48] In a way, works are no longer used *as* works, but as containers of data from which information is automatically extracted.

These techniques of automated text analysis as applied to a corpus containing virtually all the world's information are still in their infancy, and the possibilities that they will open up in the future are largely unforeseeable. To a distinguished classicist, 'the Google collection can be for us comparable to historical projects such as the Human Genome and the Sloan Sky Survey'.[49] And as is the case with these groundbreaking projects, the digitization of the world's books will certainly lead to unpredictable consequences for the method by which humankind will make use of works of literature and science. Similarly to what occurs with the Human Genome, here the only legal point is whether such a project, and the resulting uses, ought to be regulated or whether it should remain – as currently – the wild 'land of pioneers and of the long hunters'.[50]

non-profit researchers affiliated to a US university, research organization or governmental agency, or otherwise demonstrate 'that he or she [...] has the necessary capability and resources to conduct Non-Consumptive Research', ibid, § 1.123(a–d).

43. Ibid, § 7.2(d) ('Limitations to Non-Consumptive Research').

44. Ibid, § 7.2(d)(2)(vii–x).

45. Samuelson (n 11) 1353.

46. The Stanford University Libraries Amicus Letter (n 36) 5. The technique of extracting 'key ideas' from books is explained in BN Schilit and O Kolak. 'Exploring a Digital Library through Key Ideas', *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (Pittsburgh, Pennsylvania, USA 16–20 June 2008), available at <http://sites.google.com/site/schilit2/fp035-schilit.pdf>.

47. J-B Michel et al., 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Sciencexpress,* 10 December 2010 <www.sciencexpress.org>.

48. The Stanford University Libraries Amicus Letter (n 36) 5.

49. Letter from Gregory Crane to Judge Chin in Support of the Settlement Agreement, *Author's Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (7 August 2009) 1.

50. TJ Skeyhill *Sergeant York: Last of the Long Hunters*, quoted in JE Smith, *Reconstructing*

In the following, we discuss the legitimacy of non-display uses in the light of the current US and European jurisprudence.

2 How US and European jurisprudence (would) deal with non-display uses

The legitimacy of non-display uses has as yet been neither properly tested by courts nor reviewed by scholars. In fact, many scholars have commented on whether the activities carried out within the Google Books project are covered by the fair use defence under US law.[51] In their majority, scholars argue that scanning (and OCR-ing) for purposes of indexing and search is fair use, and this is also considered to be the most probable judicial ruling, should the case *Authors Guild v Google* ever come to a conclusion in court.[52]

One might also argue that the making of non-display uses on copyright-protected works is inevitable in the age of the internet. For instance David Drummond, Chief Legal Officer of Google, describes Google's activity as doing 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'. He 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'.[53] Although in Drummond's testimony there is no mention of non-display uses, his argument may find support both in the jurisprudence and in the common practice. After a work is being scanned and OCR-ed, it is difficult, if not impossible, to draw a clear line between what is necessary for making the work accessible through searching and what exceeds this purpose. As a consequence, it might be argued that any such uses should be *prima facie* permitted, unless they cause harm to an author's economic or moral rights.

Yet, the point is that 'indexing and search' is neither the only nor perhaps the most relevant purpose for which works are scanned and OCR-ed. Most of the activities that fall under the umbrella term of non-display uses do not relate to indexing and search. What does data mining and computational text analysis have to do with the purpose of indexing a work and making it searchable to users? Are *all* such activities covered by the fair use defence?

To Pamela Samuelson, who first drew attention to the relevance and legal implications of non-display uses,[54] Google could have a plausible fair use defence for these uses too. This is because 'non-display uses of the [Google Books] corpus would likely result in advancing knowledge and/or in the creation of new non-infringing works of authorship, such as new tools to aid in the translation of texts from one language to

the American Historical Cinema: From Cimarron to Citizen Kane (The University Press of Kentucky, 2006) 240.

51. USC 17, § 107.

52. Such a ruling will pave the way, at least in the US, to mass digitization projects, which base their activity on scanning and OCR-ing non-digital documents.

Pamela Samuelson writes: 'In February 2006, I hosted a workshop of about 15 copyright professors to discuss Google's fair use defense in the *Authors Guild* case. The general consensus at that meeting was that this fair use defense was likely to succeed'. Samuelson (n 11) 1314 (footnote 33).

53. 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.

54. See papers and presentations on the Google Books case at <http://www.ischool.berkeley.edu/~pam>.

another'.[55] These uses would be 'transformative' under the conditions established by the American jurisprudence,[56] and as such 'are unlikely to bring about any harm or potential harm to the market for the underlying work'.[57]

Although non-display uses in mass digitization projects have not yet been brought to, and examined by, American courts, the finding of fair use might be supported by a line of cases on internet search engines. In *Kelly v Arriba Soft Corporation*,[58] the appeal court found that the creation of thumbnails of in-copyright works passed the fair use test and qualified as a permissible transformative use. In determining whether Arriba's use of the images was transformative, the court found that the use had merely to supersede the object of the original pictures or add a further purpose or different character to the pictures. Following this test, the court found that Arriba's use of Kelly's images as thumbnails was transformative.[59] A similar holding was made in two other cases on the reproduction of images by a search engine for purposes of searching and indexing.[60] Yet, in none of these cases have uses exceeding the purpose of indexing and searching been examined, as is the case with most of the non-display uses carried out by Google.

However, a case which has many similarities with Google's non-display uses has been recently discussed in the US. In *iParadigms*, the storage and processing of students' courseworks by an online service for detecting plagiarism was found to be a fair use.[61] In particular, the appeals court upheld the district court's decision by taking the stance that archiving the student works for the automated evaluation of originality of other student's works did not infringe copyright. This applied irrespective of the fact that iParadigms was carrying out a commercial operation.[62] Rather, the court found the use in question to be transformative on the basis that the use of the papers was made 'for an entirely different purpose [from the original], namely, to prevent plagiarism and protect the students' written works from plagiarism'.[63] This use of the

55. Samuelson (n 11) 1363 (footnote 280).
56. *Campbell v Acuff-Rose Music, Inc.,* 510 US 469 (1994).
57. Samuelson (n 11) 1363 (footnote 280).
58. 336 F 3d 811(CA9 2003).
59. The Court affirmed the position earlier adopted in *Campbell v Acuff-Rose Music, Inc.*, 510 US 569, 579, 114 S Ct 1164, 127 L Ed 2d 500 (1994). However, see *contra: Infinity Broad. Corp. v Kirkwood*, 150 F 3d 104, 108 (2d Cir 1998) (the retransmission of a radio broadcast over telephone lines is not transformative); *UMG Recordings, Inc. v MP3.com, Inc.,* 92 F Supp 2d 349, 351 (SDNY 2000) (the reproduction of an audio CD into a computer MP3 format does not transform the work); *Los Angeles News Serv*., 149 F 3d at 993 (reproducing news footage without editing the footage was not found to constitute a transformative use).
60. These are *Field v Google* 412 F. Supp 2d 1106 (D Nev 2006) and *Perfect 10 v Google* 416 F Supp 2d 828 (CDCa 2006) (*affirmed in part, reversed in part sub nom Perfect 10 v Amazon.com,* 487 F 3d 701 (9 Cir 2007)). For a critical discussion of these cases see R Jeweler, 'Internet Search Engines: Copyright's "Fair Use" in Reproduction and Public Display Rights', CSR Report for Congress, 12 July 2007. On the applicability of this line of cases to the Google Books case see H Travis, 'Google Book Search and Fair Use: iTunes for Authors, or Napster for Books?' (2006) 61(601) University of Miami Law Review 628, and M Sag, 'The Google Book Settlement and the Fair Use Counterfactual' (2010–11) 55(1) NYL Sch L Rev. Even though a finding of transformativeness will outweigh the fact that a use is commercial, uses that are exploitative in nature will not be considrerd fair; *Kelly* instructs that a use made to directly promote the services provided by the secondary user will be exploitative. See Kelly (n 58) 818.
61. *AV et al. v iParadigms*, LLC, 562 Federal Reporter, 3d Series [2009], 630–47.
62. Ibid, 638–39.
63. Ibid, 638 citing the district court decision.

student works 'did not impair the market value' – if any – 'for high school term papers and other such student works'.[64] The fact that there was no substantial alteration to the original works was not found as a barrier of establishing transformativeness in the use.[65]

If such broad interpretation of transformativeness is adopted, it is easy to predict that many non-display uses will benefit from a fair-use defence in American courts. Certainly, given the international dimension of the Google Books case, considering this case only through the lens of the US copyright doctrine is too restrictive. A project like Google Books is by definition transnational, not only in the sense that a digital library on the internet is accessible from everywhere in the world, but also in the sense that books in the corpus come from different languages and traditions and cover virtually all the world's cultural heritage. As a matter of fact, the activities underlying the Google Books project are less likely to be permitted in Europe,[66] and they have actually been found infringing in France.[67] Although the French decision did not discuss non-display uses, it seems that the activities ranged under this umbrella term have little hope of finding a plausible defence in that jurisdiction.[68] This is because, under European jurisdictions, unauthorized reproduction of the whole work has no *passe partout* defence comparable to that of American fair use. Defences such as that for private copying or for reproduction for educational purposes apply neither to public commercial ventures, such as Google Books, nor probably to the making of digital libraries in general.[69] In the UK, the defence of fair dealing is too narrow in scope to cover massive reproduction of copyright works for purposes of automatic search and indexing.[70] Non-display uses of digital copies inevitably require the reproduction of the whole work – not only as obvious prerequisite (a 'copy' must initially be made), but as condition for these uses as such (permanent and/or temporary copies of the work are repeatedly made in the course of the automated processing). Therefore, if the initial copy made is the result of unlawful reproduction, the legitimacy of every other subsequent act made in respect of that copy is tainted by the unlawfulness of the first act and may be infringing – according to the principle flowing from the Roman maxim: *fraus omnia corrumpit*.

64. Ibid, 636.
65. Ibid, 639.
66. See P Ganley, 'Google Book Search: Fair Use, Fair Dealing and the Case for Intermediary Copying', Working Paper, 13 January 2006, available at SSRN: <http://ssrn.com/abstract=875384>, 17–21.
67. *Editions du Seuil et autres c Google Inc et France*, Tribunal de grande instance de Paris 3ème chambre, 2ème section Jugement du 18 décembre 2009.
68. Needless to say that any decision in a jurisdiction other than the US will not affect non-display uses, as far as they are carried out on digital copies stored on servers located in the US. In other words: no non-US court can prevent Google from making non-display uses of the copies stored in their servers – and probably not even from continuing scanning and OCR-ing non-US books within the US territory.
69. See Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society, OJ L 167/10, 22.6.2001, Arts 5(2)(b) (reproduction for private use) and 5(2)(c) (reproduction made by publicly accessible libraries, educational establishments or museums). For an overview of the copyright limitations available in the context of digital libraries see M Ricolfi 'Digital Libraries in the Current Legal and Educational Environment: a European Perspective', in L Bently, U Suthersanen and P Torremans (eds), *Global Copyright: Three Hundred Years since the Statute of Anne, from 1709 to Cyberspace* (Elgar 2010), 216–29.
70. Copyright, Designs and Patents Act 1988, ss 29–30. See Ganley (n 56) 21 ('none of the fair dealing categories fit: using a searchable electronic index of works for "informational" purposes is simply not an exempted category of use').

3 Copyright-irrelevant uses? Views from the copyright field

The current jurisprudence seems to design a legal paradox: while invasive and economically significant activities on copyright works, such as those of data mining, are likely to be exempt from infringement, an act of reproduction that might be potentially innocent is, at least in Europe, straightforwardly infringing, irrespective of its aim and purpose. In our view, both sides of this paradox lead to legally uncertain conclusions. This is because the fact that works are not displayed to the public cannot sufficiently establish that all activities falling under the umbrella term of non-display uses are copyright irrelevant. At the same time, determining the lawfulness of an act of reproduction requires an examination of the *purpose* for which the copy was made. Indicative in this respect is the fact that, even though reproduction is an exclusive right under every jurisdiction, acts of reproduction may escape liability on the basis of their purpose.[71]

To determine the legitimacy of non-display uses, it is essential to review the arguments that may support the view that non-display uses should be put under rightholders' control, as well as those opposing this view.

(a) Work as object of property (and counter-arguments)

Some authors objecting the Google Books Settlement Agreement[72] view non-display uses as a very lucrative market which is still in its infancy, or, otherwise put, as 'the very core of Google's powerful revenue engine'.[73] To them, this market will determine 'both how authors' creative works will be exploited and how authors will be compensated'.[74] In this light, granting Google an irrevocable licence to exploit such uses is just 'grossly unfair to authors'.[75] Certainly, these new uses generate a 'value' over which many parties have a legitimate claim. It is not clear on which grounds all the value should be appropriated by one party only, namely the 'user'! Moreover, the pro-authors' argument finds an orthodox justification in the copyright discourse, according to which rights should extend to all economically valuable uses of the work. This prescription is considered to be consistent with the very purpose and logic of copyright, which is that of enabling authors' works to reach the widest possible audience. As Paul Goldstein classically puts it: 'the best prescription for connecting authors to their audience is to extend rights into every corner where consumers derive value from literary and artistic works'.[76] On this basis, it can be argued that authors have a right in all uses of their works from which value can be derived, and these are not necessarily uses which display the content to the public. For instance, in relation to uses of data about books' uses – who reads them, where they are read, what keywords are searched for, and in what context – it has been argued that the author may have a legitimate entitlement to participate in the economic value that is derived from their exploitation. This is because, as the former CEO of Napster, Hank Barry, has put it, increasingly a work of authorship is not only an 'object' but a *shared* object, that is 'a powerful social connector' which is, or has the potential to be, a 'node' or a

71. See for instance the list of exceptions to the reproduction right enumerated under Article 5 of Directive 2001/29/EC.
72. See *Guthrie's Objection* (n 20).
73. Ibid, 12
74. Ibid, 20.
75. Ibid.
76. P Goldstein *Copyright's Highway: from Gutenberg to the Celestial Jukebox*, rev edn (Stanford University Press 2003) 216.

locus for online communities, the latter being 'powerful generators of money'. If this is the case, then an author should have a right to 'participate in the economic benefit of a community formed around [her] work'.[77]

This 'expansive' approach to copyright has become a source of criticism.[78] As far as non-display uses are concerned, one might argue for instance that the beneficial effects of having works easily retrievable and accessible outweigh the alleged unfairness to authors. This argument could find a public interest basis, but also be supported from the narrower viewpoint of authors and rightholders. Indexing and search in large databases could bring back to life works that lay forgotten on libraries' shelves and were out of print. Algorithms may enable the meeting of works with potential readership. Personalized advertising and behavioural analysis are essential ingredients in order to exploit the potential of the web as a distribution system. This model promises to 'match geographically dispersed buyers to a product of their choice efficiently', in all the cases where 'the limited demand for each individual title would not have proved to be sufficient under the old distribution model'.[79] In this light, one might conclude that non-display uses are as necessary in the web-based distribution as packaging and shipping was in the old distribution model. In the prospect of an unprecedented availability of works to the benefit of both the authors and the public, it would be more efficient to leave, as packaging and shipping were and still are non-display uses outside the author's control.

However illuminating might be the comparison of the new and old worlds, the utilitarian approaches do not help much in finding a sound regulatory principle for non-display uses. As a matter of fact, it is difficult to see how the chimerical 'right balance' between authors and users could be substantiated by sheer reference to works as values and to works' uses as appropriation of values.[80] Obviously, by not being regulated, the value generated by non-display uses is entirely appropriated by users, or rather by the only user which has currently both the know-how and the resources to exploit them, namely Google. Is this a welcome outcome of the counter-proprietary argument? In spite of the declared aim of 'striking the right balance' between authors and users, a value-based approach may be unfit to provide guidance in determining the scope of, respectively, authors' and users' rights over non-display uses.[81]

(b) Work as communicative act

Under a more consistent approach, a work of authorship is primarily an act of communication, in the sense that it bears a speech addressed by an author to the public by

77. Barry (n 34).
78. See generally Netanel (n 15) 154–68 (criticizing the 'proprietarian' argument).
79. Memorandum of Amicus Curiae Open Book Alliance in Opposition to the Proposed Settlement Agreement, *Authors Guild, Inc. v Google, Inc.*, No 1:05-CV-08136 (4 September 2009) 4.
80. For a critique of a value-based view of copyright, see A Drassinower 'From Distribution to Dialogue: Remarks on the Concept of Balance in Copyright Law', (2009) 34(991) The Journal of Corporation Law 999–1002.
81. Ibid, 1001: 'To be sure, the struggle for value between authors and users is a sociological fact, and an obvious one at that. The problem, however, is that once copyright law is understood as a balance preoccupied with sheer and mere value, it cannot account for itself as a practice providing the basis for a specifically juridical resolution of this struggle. Framed as a balance, copyright law becomes nothing more than a distributive mechanism, on the one hand designed to achieve a balance between authors and users, yet on the other unable to make the *qualitative distinctions* necessary to get the entire balancing process going in the first place'.

means of an intermediary.[82] From this perspective, the author–user relationship is not a struggle in view of appropriating value but rather an ongoing dialogue between collaborative parties.[83] The use of a work may be restricted not on the basis that it results in a misappropriation of value, but on the ground that it encroaches on an act of communication between the author and the public. As a consequence, an act is restricted either when it forces an author to speak against her will, as is the case with unauthorized publication,[84] or when it otherwise interferes with the modalities chosen by the author to speak to the public.[85] To what extent do non-display uses encroach on the communicative act in which the work consists?

It is worth noting that the right to communicate the work to the public (via publication, reproduction, distribution or otherwise) is necessarily an *exclusive* right under this approach. This is because the act of communication implies that the intermediary (e.g. the publisher) speaks to the public in the name of the author, and this can only be done through an explicit mandate of the author. Otherwise put, an author does not merely have a right to be compensated for the use that the intermediary makes of her work but, more importantly, an exclusive right to choose *who* is authorized to speak *in her own name*. This means that no intermediary can speak in an author's name on the sole ground that he is not causing any harm – or that he is even bringing about advantages – to the author.

So, whereas some may see non-display uses as not harmful to authors, these uses could have some unprecedented impact upon the way the author's word is delivered to the public. Despite the fact that the work as such is not displayed, it may be indexed or appear as the result of search queries in an environment becoming more and more shaped by search algorithms. In this light, non-display uses will increasingly determine the *context* in which works will be communicated to the public. The absence of any control over non-display uses might result, for example, 'in advertisements being positioned next to or associated with authors' work in ways that are objectionable, offensive, or harmful to the author'.[86] Automatic associations may be generated as effect of 'groups attempting to create a false appearance of endorsement'.[87] More subtly, behaviours of groups of readers which use an author's work for own purposes, for instance to support own political views, may influence algorithms so that the author will be automatically associated with those views.

While the practice of misleadingly appropriating authors and works for political or ideological aims – a practice which we may term 'author squatting' – is as old as

82. This approach is based on Kant's argument on unlawful book reprinting. See ibid, 1002. See also A Drassinower, 'Authorship as Public Address: On the Specificity of Copyright vis-à-vis Patent and Trade-Mark' (2008) 1 Michigan State Law Review 1999; M Borghi, 'Copyright and Truth' (2011) 12 Theoretical Inquiries in Law 1.
83. As shown by Drassimower (n 80).
84. See Kant's argument (the unauthorized reprinter speaks in the author's name without the author's consent) and Drassinower (copyright infringement as 'compelled speech').
85. The so-called moral right of integrity flows from this understanding of the work as communicative act. In the wording of the Berne Convention, the right of integrity is defined as 'the right [...] to object any distortion, mutilation or other modification of, or other derogatory action in relation to, the [...] work, which would be prejudicial to [the author's] honour or reputation' (Berne Convention for the Protection of Literary and Artistic Works, 1886 as amended, Art 6 bis).
86. *Guthrie's Objection* (n 20) 19.
87. Ibid.

humankind,[88] its use as a source of information to produce automatically generated associations is brand new. In this respect, computational analysis of language, coupled with aggregated data mining, may associate an author with other authors, products, or even 'ideas' *independently from their expression*. In a 'reading environment' determined by non-display uses, for instance, a book might not be available for viewing and reading, but users might catch its 'key idea' as extracted via automatic text processing; they might even be readdressed to authors which allegedly share 'similarities' or even 'similar ideas' based on associative and taxonomic data analysis.[89] A work which has been searched may not be available for viewing, or may be available only upon payment, but the search engine may still provide users with any sort of proxies – spanning from snippets, quotes, cross-referencing, or even 'key ideas'. This means that, irrespective of whether the work as such is displayed by the intermediary, or is only made searchable by it, computer-generated associations may deeply interfere with the modalities chosen by the author to communicate her work to the public.

(c) Work as manifestation of personality – and the use thereof

Apart from the impact on the way in which the work is communicated, an author may not be content with certain uses of her work. For instance, an author may be uncomfortable with the automated processing of her work for the extraction of information over the work's access and use, or for the development of commercial products. She might also find that automated textual analysis applied to her work is detrimental to her dignity as author, or, simply, that this kind of automated processing of human language is contrary to her principles as author. A poet might find the use of her words as 'semantic units' in the context of computational linguistic analysis disturbing, simply because she believes that words are not to be 'used' in this way. Similarly, a philosopher might object to the automatic extraction of ideas from her books, on the very simple ground that, as Charles Dickens masterfully expressed, 'an idea, like a ghost, must be spoken to a little before it will explain itself'.

The point here is whether an author has a legitimate claim to oppose such uses on the ground of own judgement and beliefs. Certainly, a claim of this kind can be substantiated only on the assumption that an indissoluble link exists between the author and the work, so that the first is – to a certain extent – entitled to oppose certain uses of the latter once it has been released to the public. The personality right theory provides support to this argument.

The assumption of author's rights under the umbrella of personality rights is commonly referred to Hegel.[90] In fact, Hegel explicitly treats the legal meaning of the 'products of the intellect' under the heading of 'goods, or rather substantive determinations [...], which constitute my own person and the universal essence of my self-consciousness in the deepest sense'.[91] Contrary to what is sometimes superficially reported, this does not mean that the work should be seen as an 'expression' or even as an 'emanation' of 'the author's personality', as would be – legally speaking – the 'smell of landfill gas and litter [...] emanating from domestic waste'.[92] The author is

88. On the use and abuse of classical authors by totalitarian regimes see H Arendt, *The Origins of Totalitarianism* (New York, Harcourt Brace & Company 1951).
89. As illustrated in Schilit and Kolak (n 46).
90. See GWF Hegel, *Grundlinien der Philosophie des Rechts*, § 43 and 68–9. See also GWF Hegel, *Die Philosophie des Rechts. Vorlesung von 1821/22*, § 68–9.
91. Ibid, § 66.
92. *Blackburn v ARC Ltd* [1998] Env LR 469.

not the 'owner' of her 'emanations' in the way that people are holders of any smells emanating from their own properties. For Hegel, 'personality' is not an empirical concept describing a sort of 'internal sphere' as opposed to external manifestations thereof; it is a purely abstract concept which defines the human being in its capacity to *possess* rights, and not only to be subject of legal relationships.[93]

Accordingly, when it comes to 'things' which have an existence as such but, at the same time, are substantial determinations of one's own personality, the question arises as to whether, and to what extent, these 'things' maintain a link with the person once they are separated from it and, so to speak, attain a life on their own. So, for instance, when an author alienates the work (or a copy thereof) to a buyer, the latter comes into possession of both 'the possibility [...] to appropriate the thoughts that are communicated through it' and of 'the universal way and manner to express oneself in the same way'.[94] However, while the buyer acquires whole possession of the first possibility, as it constitutes 'the sole destination and the value' of an object such as a literary work, she has no right to make use of the second element, namely the power to 'produce the work in the same way'. This latter is a special kind of use which is 'different and separate from the use to which the object was directly appointed'.[95] The purpose of publication is therefore to enable other human beings to appropriate the author's thoughts; it is not that of facilitating the reproduction of the work in the same way.[96]

While this argument has so far provided the justification for the reproduction right, it may also form the justificatory ground for restrictions on other uses of the work which are unrelated to the use to which the work was directly appointed. Otherwise put, it could justify a right of the author to have her work used in the context of content appropriation but not for purposes disconnected from this one, since other purposes could come in conflict with the inherent purpose of the work. In this light, the fact that reproduction has been so far the only meaningful case of a use not related to the intended function of the work does not mean that this should remain the *only* restricted use of an author's work.

The non-display use of digital works is a new technologically empowered activity that does not seem to relate to the inherent purpose and function of works, namely content appropriation. At the same time, it does not feature in the list of restricted acts as reproduction is, probably due to the fact that copyright doctrine has not so far encountered anything analogous to this kind of use. In the next part we discuss these issues with a view to defining a regulatory principle for non-display uses.

93. *Grundlinien*, § 36.
94. Ibid, § 68.
95. Ibid, § 69.
96. This argument resounds with the idea-expression dichotomy, as firstly discussed by Fichte in his booklet on *Proof of the Illegality of Reprinting* (1793): 'By purchasing the book, we acquire the possibility of appropriating the author's ideas […] What, on the other hand, can absolutely never be appropriated by anyone else, is the form of the ideas, the combination in which, and the signs through which they are presented' (JG Fichte, 'Proof of the Unlawfulness of Reprinting (1793)', transl M Woodmansee, in L Bently and M Kretschmer (eds), *Primary Sources on Copyright (1450–1900)* <http://www.copyrighthistory.org>. For a discussion see M Borghi, 'Owning Form, Sharing Content: Natural-Right Copyright and Digital Environment', in F MacMillan (ed.), *New Directions in Copyright Law*, vol. 5 (Elgar 2007), 197, 205–211. Compare also A Drassinower, 'A Right-Based View of the Idea/Expression Dichotomy in Copyright Law' (2003) 16 Can. J.L. & Jurisprudence 3.

PART TWO REGULATING NON-DISPLAY USES

Non-display uses do not involve the dissemination of the work to the public, but may deeply interfere with its dissemination. Neither are non-display uses pure uses of the work, since they do not aim at performing the only reasonable use for which a book is published, namely reading and learning from its content; the ideas expressed in books are not 'spoken to' so as to explain themselves and be perceptible! Since the life of copyright has been shaped by the test of time and subversive technologies, it is important to see how non-display uses of works may be regulated from a copyright perspective.

To be sure, the current lack of regulation over non-display uses triggers the emergence of private ordering mechanisms. Contracts preventing text mining are already used in the context of digitization projects, such as that of British Library,[97] and these contractual restrictions are fated to be more frequent in future contracts between authors and digital libraries, or between libraries and digitization projects.[98] However, the contractual route cannot be a viable solution as regards mass digitization projects, which mainly include works in the public domain or works whose rightholder cannot be located.[99] Should contracts over non-display uses be permissible, this would generate legal uncertainty, since the corpus of a digital library would be fragmented between several legal conditions of use. This calls for a comprehensive regulation of non-display uses that is applicable *ex ante* to copyright works, and cannot possibly be overridden by contract.

It should be borne in mind that the term 'non-display uses' covers a range of acts, spanning from those necessary to make the work searchable by users, to more sophisticated processes. Not all of those activities can be justified on the mere ground that they are *necessary* to make a digital library or a search engine. In our view, non-display uses essentially involve two copyright-significant activities. First, they require the copying of the works for the purposes of indexing and search. Copying in this context entails both the creation of temporary copies, and the scanning and OCR-ing of works. Secondly, the copies that are created are automatically processed. Automatic processing may be carried out for computational analysis, i.e. non-consumptive research, or for purposes of data mining. Below, we examine these activities with a view to determining a regulatory framework for non-display uses.

1 Copying for indexing and search

Mass digitization projects are premised upon the creation of copies of works. While the case of temporary copies is settled by copyright law, in terms that the creation of such copies is exempt from infringement under certain conditions,[100] the situation of scanning and OCR-ing may not be equally straightforward. To examine whether these latter activities are reproductions within the meaning of copyright law, we first need

97. See D LJ Brindlay, 'Phoenix in the Internet Era – the Changing Role of Libraries', in Bently, Suthersanen and Torremans (n 68) 184.
98. However, it should be borne in mind that, should the Amended Settlement Agreement be approved by the court, authors and libraries would have no capacity for contracting out of Google's non-display uses (see above nn 17–20 and accompanying text).
99. See above (n 16 and accompanying text).
100. See, for instance, Article 5(1) of Directive 2001/29/EC.

to make a distinction between the terms 'expression', 'form' and 'format'. By use of the term 'expression', we refer to the way by which the author has fixed her work; for instance, the way that she has selected and ordered the words in her book. It is when the 'expression' of the work is copied that the reproduction right is infringed. The term 'form' is used to refer to the way in which the 'expression' is recorded from a technical perspective. For instance, a book may be recorded in analogue or digital 'form', while in both cases carrying the same 'expression'. 'Format' is the technological specification of the 'form'. For instance, a book in digital form may be recorded in a text-file or an image-file 'format'.

Scanning consists in the process of transposing the information contained in a work from analogue into digital 'form' – from symbols perceptible and understandable by the *human* senses into a series of ones and zeros. This activity is highly likely to qualify as an act of reproduction.[101] This is because the passage from the typographical character to the numerical token does not change the 'expression' underlying the work; the 'expression' is used as a basis and the work is translated from one 'form' to another.

OCR-ing, on the other hand, modifies the 'format' of the work by converting the image file that has been created via scanning into a text file.[102] Even though the 'copy' of the work created through OCR-ing is not meant to be perceived by human intelligence but to be readable by computers, nevertheless it is not the 'expression' of the work but the 'format' into which the work is recorded that is still subject to modification. Whereas OCR-ing deeply impacts on the reproduction right, since the 'expression' comprising the work is copied and remains intact as such, under some jurisdictions it could qualify as a transformative use, insofar as the use made appoints the copy a new function or purpose. And whereas in Europe this kind of use is highly unlikely to escape from infringing liability on the grounds of transformativeness, it may be found transformative and hence permissible under US copyright law.[103] As indicated earlier, the test for evaluating whether a use is transformative in the US requires that the work is employed in a different manner or for a different purpose from that of the original.[104] Even if OCR-ing passes the scrutiny of this test under the merits of a specific case, most jurisdictions would be slow to accept that it escapes infringement for being transformative. In Europe, for instance, this process would most likely fit within the description of an act of reproduction and would be held infringing once carried out without authorial consent.[105]

101. Compare M Ficsor, 'Collective Management of Copyright and Related Rights in the Digital, Networked Environment: Voluntary, Presumption-Based, Extended, Mandatory, Possible, Inevitable?' in D Gervais (ed), *Collective Management of Copyright and Related Rights* (Kluwer Law International 2006) 37–83, at 67. A radio station digitizing musical works for radio transmissions was held to infringe the reproduction right in Austria. Austrian Supreme Court, 26.1.1999, MMP 1999, 352: *Radio Melody II,* file no. 4 ob 345/98h. For an analysis, see A Haller, 'Digitalisation/Storage for Broadcasting Purposes Constitutes Reproduction' (1999) 4(6/8) *IRIS* 1–2.

102. The process is discussed in *Infopaq International A/S v Danske Dagblades Forening*, Case C-5/08 [2009] ECDR 16, § 18–19.

103. See above (nn 53–64 and accompanying text).

104. Ibid.

105. As upheld in *Editions du Seuil et autres c Google Inc et France*, Tribunal de grande instance de Paris 3ème chambre, 2ème section Jugement du 18 décembre 2009. See above (nn 65–9 and accompanying text).

Nonetheless, the purpose for which scanning and OCR-ing are carried out could be a valid determinant for their lawfulness. Copying for the sole purpose of indexing and search enhances the dissemination of information about the work, while it does not impinge upon its communication to the public. In this respect it should be exempt from infringement, even in cases where the reproduction is not authorized by the relevant rightholder or is addressed to the public.

However, the legitimacy of the activities that follow copying in the context of mass digitization remains questionable. Below, we examine the status of automated processing of copyright works from the viewpoint of the *purpose* of processing. We will attempt to make some distinctions based on copyright principles.

2 Copyright and the automated processing of works

As has always been the case in copyright, copying is merely a prerequisite for other activities to take place.[106] Traditionally, the most common activities following copying were, broadly speaking, the *dissemination* and the *use* of the work. Copying carried out with an exploitative purpose, such as that of distribution on commercial scale, is reserved by copyright, whereas dealing with the copies in a purely consumptive way is normally exempt from infringement.[107] This finds ground in the theoretical argument that the use of works is meant to enhance the understanding of the *content* of works, while the dissemination of copies to the public is associated with the exploitation of the *expression* embodied in works. In this light, the way in which the copies are dealt with is very important in determining the lawfulness of copying, despite the fact that unauthorized reproductions are infringing as such, by forming the basis of a broadly defined exclusive right in every copyright system.[108]

In the context of mass digitization, once the works are copied, they are automatically processed for a series of different purposes, from plain search and indexing to more complex data mining and text analysis. This form of dealing with copies, namely *automated processing*, bears *prima facie* no resemblance to any currently known form of dealing with copies, namely: *dissemination* and *use*. It appears to be a new kind of dealing with works that follows copying. First, it is almost self-evident that it is not a public dissemination since the copies are not displayed to humans. Is this conclusive evidence that there is no infringement? Secondly, it is contestable whether non-display uses are what the term refers to, at least in a copyright sense: 'uses'. We address these issues in the following subsections.

(a) Dissemination to the public and dissemination to machines

The creation of copies with a view to disseminate them to the public has traditionally been viewed, and still is considered, as implicit within the meaning of the reproduction

106. National Academy of Sciences, *The Digital Dilemma: Intellectual Property in the Information Age* (National Academy Press, Washington DC 2000) 140. From its very beginnings, copyright entails a right to control certain acts of 'copying'. The verb 'to copy' was probably first used, in its current legal meaning, in the Engravers' Copyright Act of 1735. See R Deazley, 'Commentary on the Engravers' Act (1735)', in L Bently and M Kretschmer (eds), *Primary Sources on Copyright (1450–1900)* (2008) <http://www.copyrighthistory.org>.

107. See above (nn 90–93 and accompanying text) for the discussion of this dichotomy based on personality rights argument.

108. See, for instance, Article 2 of Directive 2001/29/EC.

right.[109] This argument is supported by copyright history. Reproduction as the process of multiplying works was never meant to refer to 'copying as such', i.e. the isolated act of making copies of a work for purposes other than public diffusion. Rather, reproduction implicitly encompassed 'copying *and* selling' or 'copying *for purposes* of publication'.[110] So, unless the copies were disseminated to the public, copying did not matter *per se*. Scholars have convincingly supported this view, namely that reproduction does not matter *per se* until and unless copies are meant to be disseminated to the public. One of the arguments is that, in the digital environment, *any* use of the work involves an act of reproduction; hence, the making of copies is no longer 'a good predictor of whether there will be distribution to the public', and consequently of an 'intent to infringe',[111] as was the case in the analogue world. Copies can be, and normally are, made for purposes that cannot be realistically considered to be an infringement.

When it comes to the automated processing of works in non-display uses, it is almost obvious that it does not fit within the concept of dissemination *to the public*.[112] This is because, under our current understanding of copyright doctrine, dissemination inherently entails the concept of a public to which the copies are diffused, even in cases where the public is only abstractly present, namely where there is only a possibility that members of the public shall access the work. Yet, this concept of the public, however broadly defined, is missing from activities related to automated processing.[113] Not only is the public excluded from this type of dealing with copies, but it also becomes unprecedentedly irrelevant! In a way, the copyright-doctrinal dipole between public and private activities seems like an outdated distinction, valid only in the display world of dissemination. Since dissemination as a form of dealing with copies is inextricably linked with the presence of the public, entailing even a mere possibility of public access, the automated processing of works does not constitute an act of dissemination in this sense.

109. See eg *Tonson v Collins, (1762),* 1 Black W. 321, at 325: the purchaser of a single book may make any use he pleases of it; but no man, without leave from the author, has the right of making new books, by multiplying copies of the old.'

110. See eg the Engraver's Copyright Act (1735) which defines the scope of infringement as follows: 'engrave, etch, or work [...] or in any other manner *copy and sell*, or cause to be engraved, etched, or *copied and sold*, in the whole or in part, by varying, adding to, or diminishing from, the main design, ...'; also see Article 6 of the Fine Art Copyright Act of 1862: 'If the Author of any Painting, Drawing, or Photograph in which there shall be subsisting Copyright, after having sold or disposed of such Copyright, or if any other Person, not being the Proprietor [...] shall, without the consent of such Proprietor, repeat, copy, colourably imitate, *or otherwise multiply for Sale, Hire, Exhibition, or Distribution*, ...' (emphasis added).

111. E Miller and J Feigenbaum, 'Taking the Copy Out of Copyright', in T Sander (ed), *Security and Privacy in Digital Rights Management*, Lecture Notes in Computer Science (Springer Berlin/Heidelberg, 2002) vol 2320, 233–44, at 233 *et seq*.

112. Under the WIPO Guide and Glossary, the public includes 'persons in general, that is, not restricted to specific individuals belonging to a private group.' Albeit arguably broad, this definition interprets the concept of the public in a conventional sense: it refers to a *human* public. See WIPO, *Glossary of the Terms of the Law of Copyright and Neighbouring Rights*, WIPO, Geneva 1980, ISBN 92-805-0016-3, at 11*bis*12.

113. There is, however, the possibility of unintentional disclosure to the public. For example, Google's security systems preventing users from downloading full copies of books can be circumvented by means of an application called 'Google Books Downloader'. See <http://www.downloadsquad.com/2009/09/02/google-book-downloader> (last accessed September 2010).

Nevertheless, non-display uses of digital copies entail massive activities associated with the dissemination of data *to machines*. For instance, in non-consumptive research, as defined in the Agreement, copyright works in the form of data are transferred to 'qualified users' in order to perform computational text analysis.[114] More generally, dealing with digital copies may entail transfer of data to third parties to process these data on a non-display basis. Data may be aggregated with other data and processed by different parties for different purposes. In this light, the dissemination of works *from machine to machine* is likely to be the most common operation within the spectrum of non-display uses!

There are two main questions arising in this context. The first is whether this form of dissemination is innocent on the mere ground that it does not touch upon the public. This question has never been tested under copyright law, which is and remains a law of *public* places.[115] Copyright essentially conceives the activities connected with a work as aimed at *either* public dissemination *or* use. 'Dissemination' in the copyright sense does not cover the dealing with works *qua* machine-processable data. This brings about the second question as to whether non-display uses of digital copies of works are what their title refers to: 'uses'. Subject to examination, below is our current understanding of what a copyright use is with a view to ascertaining whether non-display uses fall under this kind of use.

(b) Uses of works and uses on works

The concept of 'use' in the context of copyright protection has not been defined by statutes, despite the fact that it is a benchmark concept underpinning copyright. The use of copyright works, however, has always remained outside copyright's exclusivity. As opposed to other disciplines of intellectual property, such as patent law, copyright has stopped short of granting exclusive *use* rights.[116] As long as works are communicated to others through publication, anyone is allowed to make free use of them, and authors have no right to exercise control over the use that is made. This is because, through publication, the author grants an *ex ante* implied licence to the public to carry out a certain range of uses. The scope of this implied licence may be subject to variations according to the technology used in the publication of the work.[117] Practically, this means that the ways of appropriating the content of a work through use are unpredictable and not subject to authorial control. So, a copyright use may be purely consumptive, in terms of appropriating the content of works, but it could also extend, for example, to the study or research of various aspects of works. Those aspects could include language, style and structure. In this respect, a use may lead to

114. *Amended Settlement Agreement*, § 1.93. See discussion (nn 35–41 and accompanying text).

115. This means, for example, that reproductions and performances of works for private use can be made without authorial consent (H Wistrand, *Les exceptions apportées aux droits de l'auteur sur ses œuvres* (Editions Montechrestien, Paris 1968) at 313). See also P Goldstein, *Copyright's Highway* (n 76) at 201; S Dusollier, 'Technology as an Imperative for Regulating Copyright: From the Public Exploitation to the Private Use of the Work' (2005) 27(6) EIPR 201–04, at 201.

116. PB Hugenholtz, 'Fierce Creatures, Copyright Exemptions: Towards Extinction?', keynote speech, *IFLA/IMPRIMATUR Conference, Rights, Limitations and Exceptions: Striking a Proper Balance*, Amsterdam, 30–31 October 1997, 6; A Latman, *Fair Use of Copyrighted Works*, Arthur Fisher Memorial Editions, Study No 14 (1955) at 5.

117. See BGH (German Supreme Court) Case I ZR 69/08, 29 April 2010 (making pictures available on the internet implies a 'silent licence' to reproduce the work in thumbnails).

associating a work with other works. However broad in scope, copyright use is the most welcome outcome of the bequest of works through publication and there are no legal restrictions on how this use is going to be made, insofar as it does not affect the sphere of exclusive rights granted by copyright laws.[118]

Going through the applicable copyright exceptions and limitations, one can track down four main types of copyright use. Those include consumptive,[119] educational,[120] informational[121] and creative[122] uses. While these uses differ as to the purpose for which they are carried out, they have one thing in common: they are made with a view to appropriate, or enable the appropriation of, the content of works by human users. For instance, consumptive uses involve the pure personal enjoyment of a work, via reading, viewing or listening to it.[123] Educational and informational uses are premised upon fundamental freedoms and as such they are exempt from infringement irrespective of the fact that the copies of works may be disseminated to others to appropriate.[124] In creative uses too, the content of a work needs to be appropriated so as to become part of a creative process and result in the creation of new expression that is original in its own right, while still embodying elements of the original work.[125] All aforementioned categories of uses are in essence uses *of* the work since they principally involve the appropriation *of* its content.

Despite the probably misguiding employment of the term 'use' in the context of non-display uses, the copies made to be further automatically processed do not comfortably fit in the currently known types of copyright use. This is because they are not made in the context of appropriating the content of works – an activity which is supposed to be practised by human beings only. Rather, they are carried out with a view to automatically processing the expression of works. They are in essence uses *on* the work. This is affirmed by the very definition appointed, for instance, to 'Non-Consumptive Research' in the Google Books Settlement Agreement. As we saw earlier, non-consumptive research is defined to refer to 'research in which computa-

118. This principle flows from both the understandings of work as 'communicative act' (see text accompanying nn 78–81) and as a manifestation of the author's personality (nn 85–91 and accompanying text).

119. Exceptions for consumptive use may be either made by the end user alone (eg Articles 5(2)(b) on private copying and 5(3)(n) on research and private study of Directive 2001/29/EC) or through intermediates enabling such use to be made (eg 5(2)(c) on library uses and 5(2)(e) for hospitals and prisons). Also see s 70 on the time-shifting of broadcasts. Copyright, Designs, and Patents Act (CDPA) 1988, chapter 48.

120. See Articles 5(3)(a) on use for the purpose of teaching and 5(3)(n) on research and private study. Also see s 29 of the CDPA on research and private study and s 32 *et seq* on copying for educational use.

121. See Articles 5(3)(c) on news reporting and Article 5(2)(d) on the archival of broadcasts of Directive 2001/29/EC.

122. See for instance Articles 5(3)(d) on criticism and review, 5(3)(j) on incidental inclusion and 5(3)(k) on caricature and pastiche of Directive 2001/29/EC. Also see s 30 CDPA 1988 on criticism review and news reporting.

123. See 1st Draft, Memorandum of Justice Stevens, *Sony v Universal Studios*, No. 81-1687, circulated 13 June 1983, at 17–18.

124. Note, however, that not every educational use is permissible: *Princeton Univ. Press v Michigan Document Services*, 99 F 3d 1381 (6th Cir 1999).

125. This kind of uses benefits society since it serves 'intergenerational equity'. See M Senftleben, *Copyright, Limitations and the Three-Step Test, An Analysis of the Three-Step Test in International and EC Copyright Law* (Kluwer Law International 2004) 204 *et seq*.

tional analysis is performed *on* one or more Books, but not research in which a researcher reads or displays substantial portions *of* a Book to understand the intellectual content presented within the Book'.[126]

Certainly, uses *on* works may facilitate uses *of* works. By this, we mean that automated processing may enhance the accessibility of works to users so that the latter appropriate the content of works. This includes automated processing that is carried out for the purposes of indexing and search. When a work which was only available on physical carrier is scanned and OCR-ed, and is then made available for searching on a publicly accessible network, the user is undoubtedly facilitated in having access to it. Regardless of whether the work as such is displayed or not, or if it is displayed only in part, the user is facilitated in operations that are preliminary or ancillary to the work's use properly so-called, such as: locating the work itself, obtaining basic information over it, identifying interesting parts in it. Automated processing for purposes of indexing and search is, therefore, a use *on* the work which may be considered as *subsidiary* to uses *of* works. As far as this kind of automated processing is limited and proportionate to use-subsidiarity, it may be considered permissible under copyright law and principles, insofar as it does not have impact upon any of the exclusive rights of dissemination[127].

There is, however, yet another category of uses *on* works that facilitate uses *of* works. These include automated processing that is carried out for purposes of computational analysis, i.e. for non-consumptive research. As we have discussed previously, most of these uses do in fact aim at facilitating the understanding of the work's content, for instance by means of automatic text analysis, or by supplying 'user-friendly' tools such as the automatic translation or automatic extraction of associations and of key ideas.[128] These uses *on* works, however, do not seem to be merely 'subsidiary' to uses *of* works. Obviously, their true status can only be ascertained on a case-by-case basis, but to a large extent they are not just *subsidiary*; they are, rather, *substitutive* to the use *of* the work. Once a user has been 'facilitated' in the use of a book to an extent that reading the book becomes superfluous or even immaterial to understanding the book itself, we might safely conclude that such use *on* the book has substituted the use there*of*. In the light of copyright principles, the legitimacy of these 'uses on' is not so straightforward as it may be in the previous case.[129]

Finally, there are uses *on* works that do not aim at facilitating uses *of* works but are carried out for other purposes. These include automated processing for data mining. Analysis on the search patterns of users for the purposes of personalized advertisement, for instance, seems to be totally irrelevant with the function and nature of copyright use. One might argue that these uses are part of a distribution model which is aimed at making more books available to the public, and as such they indirectly facilitate the use *of* books. Yet, precisely in this light, these uses are more similar to acts of exploitation than those of use, and as such they are more likely to fall outside the copyright understanding of 'use'.

Uses *on* works are essentially different than uses *of* works. Yet, this distinction does not appear clearly in the light of copyright principles. At the same time, it inevitably

126. *Amended Settlement Agreement*, § 1.93 (emphasis added). See discussion above (nn 35–41 and accompanying text).
127. This is consistent with the finding of fair use in the US search engine cases discussed above (n 59).
128. See above (n 84 and accompanying text).
129. In the context of US copyright, this substitutive effect might cause the fourth fair use factor to weigh significantly against the defendant (USC 17, § 107).

raises the question as to whether this shift in the way that works are used means that the works cease to be used as works but are used as something else. This is because the subject matter of copyright is essentially the *work* – not the work as a container of data.

3 Works as works and works as data

While for many years copyright has been presumed to be regulating the 'intangibles', works have never been 'intangible' *ipso facto*. The practicalities of the analogue world required the embodiment of works in a physical carrier to meet the 'fixation' requirement[130]; this means that the intellectual creation, namely the work, had to be attached to a physical embodiment, i.e. a copy. With digitalization, things changed. The digital age came with a separation of the intellectual object from its physical embodiment. This marked a shift not only in the way that copyright is being regulated, through the introduction of new rights and the initiation of anti-circumvention provisions,[131] but also in copyright language. Interestingly, with dematerialization it is quite common to refer to works in digital form as copyright 'content'. Certainly, this linguistic shift does not feature in copyright statutes. A careful reading of the statutes, however, indicates that this shift has occurred. For instance, the 'material' distribution right provided under Article 4 of Directive 2001/29/EC[132] covers works and copies thereof, whereas the 'immaterial' distribution rights of Article 3 refer to 'works' only.[133] The difference in the wording used to frame the scope of the aforementioned rights seems to adhere to the new reality brought about by technological change, namely that works have been denuded of the necessity for a tangible carrier and can thereof flow as 'content' within online networks.

With mass digitization, there is yet another shift that has occurred. While digitiza-

130. In the US, a work is fixed 'when its embodiment in a copy or phonorecord […] is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than a transitory duration.' See Paragraph 101 of the United States Copyright Act. Also see s 3(2) of the UK See Copyright, Designs and Patents Act 1988 (CDPA) 1988 in conjunction with s 178 of the same Act.

131. WIPO Copyright Treaty (WCT) adopted by the Diplomatic Conference on 20 December 1996, CRNR/DC/94; WIPO Performances and Phonograms Treaty (WPPT) adopted by the Diplomatic Conference on 20 December 1996, CRNR/DC/95; Directive 2001/29/EC; Digital Millennium Copyright Act of 1998.

132. This Article reads that 'Member States shall provide for authors, in respect of the original of their works or of copies thereof, the exclusive right to authorise or prohibit any form of distribution to the public by sale or otherwise.'

133. This Article reads that: '1. Member States shall provide authors with the exclusive right to authorise or prohibit any communication to the public of their works, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them.'

2. 'Member States shall provide for the exclusive right to authorise or prohibit the making available to the public, by wire or wireless means, in such a way that members of the public may access them from a place and at a time individually chosen by them

(a) for performers, of fixations of their performances;

(b) for phonogram producers, of their phonograms;

(c) for the producers of the first fixations of films, of the original and copies of their films;

(d) for broadcasting organisations, of fixations of their broadcasts, whether these broadcasts are transmitted by wire or over the air, including by cable or satellite.'

tion turns works into 'de-materialized' 'content', mass digitization turns content into 'de-intellectualized' 'data'. This means that the expression of the idea embodied in the work is not primarily used to communicate the 'speech' of the author to the public but rather to form the basis of machine-workable algorithms. In this context, digital works may be purely treated and dealt with as containers of *data*; uses *on* them can be made along with uses *of* them.

At this point, one could perhaps argue that once the works are not used as works but as something else, no copyright infringement occurs plainly because this kind of use exceeds the scope of protection afforded by copyright. Examples pertaining to physical property provide support to this argument. For instance, using heavy books as door-stops has nothing to do with copyright; or placing CDs on cars' rear windscreens to dodge speed traps in Italy or hanging CDs on balconies in Greece as pigeon-deterrents can only contribute to some sort of *couleur locale* but not amount to copyright infringements. This is because copyright provides no right to the copyright holder to control the physical embodiments of their works once sold. This is the core meaning of the exhaustion principle. Yet, it should be borne in mind that the aforementioned examples, under which works are not used as works but as something else, are borrowed from physical property.

Finding similar examples in the context of the ideational component comprising works is rather difficult. A good one comes from trade mark law. Whereas the use of a trade mark in the course of trade is restricted to the proprietor of the mark, its use in a different context does not raise issues of infringing liability. This is because trade marks function as badges of origin, i.e. as indications that the products or services bearing the mark originate from the same manufacturer.[134] In this light, the use of a trade mark as a trade mark is infringing, but its use as something else is not, in the sense that the property right of the trade mark owner to be identified as such is not affected.

The question arises as to whether the same conclusion applies in respect of the use of copyright works not as works but as something else, for instance as machine-processable data. To answer this question, we need to understand the essential difference between copyright works and trade marks. This lies in the way that these objects of intellectual property protection convey the information contained in them to the public. Trade marks function in a direct way, that is the trade mark owner uses the mark to convey the message that it is he who uses the mark; in turn, when members of the public come across the mark they (have to) recognize that the product or service bearing the mark originates from that owner. This is not the case with copyright works. Their function is not to indicate that a specific work originates from a specific author. In fact, indications that might serve this function, such as the title of the book or meta-data extracted from it, can hardly qualify as original and hence be eligible for copyright protection.[135] As opposed to trade marks, the function of copyright works as works involves a circular process: the author, through a blend of originality and fixation, expresses his idea(s), namely he creates a copyright work. The users then have the possibility of getting back to the author's idea(s) via the appropriation of the content of work. So, starting off with the idea(s) an author communicates his or her

134. Opinion of Advocate General Cosmas in *Windsurfing Chiemsee Produktions- und Vertriebs GmbH v Boots- und Segelzubehör Walter Huber and Another*, ECR I-02779 (1999), point 27.

135. See, for instance, *Rose v Information Services* [1987] FSR 254; *Exxon Corp. v Exxon Insurance* [1982] RPC 69.

original expression to the public so that the public is carried back to that idea through content appropriation. Unless a user reads a book or listens to the music, the work only carries the capacity of functioning as a work; it is this possibility of appropriation of the ideational component of works that distinguishes copyright works from other objects of intellectual property and that forms the inherent nature and purpose of works.

It is because of the inherent circularity of this process that some see a work as an act rather than an object. Because of this powerful dialectic relationship that works generate between the author and the users it seems convenient to look at works as acts of communication rather than as objects of property.[136]

The question arises as to whether authors should have the capacity of objecting to uses that contravene, and interfere with, the inherent nature and purpose of works. Answering this question proves extremely difficult, since at no stage of copyright history has the use of works affected, or had an impact on, this inherent purpose and function. So, to see whether authors have a right to object to non-display uses as violations of the purpose and function of works, we need to revert to some guidance in copyright principles. As we have discussed in Part One, the author's entitlement over non-display uses may be justified under the different approaches to copyright.[137] However, it is from a personality theory point of view that this entitlement attains its clearest stance as regulatory principle. Under this approach, the possibility of content appropriation is, as we illustrated earlier, 'the sole destination and the value' of an object such as a literary work.[138] In this light, publication serves as a means of enhancing the possibility of content appropriation rather than of reproducing the work in the same way with the author, reproduction being a use that is 'different and separate from the use to which the object was directly appointed'.[139] The fact that reproduction has been so far the only meaningful case of a use that is not related to the intended function of the work does not mean that this should remain the *only* restricted use of an author's work. The personality theory could justify a right of the author to have her work used as work, namely in the context of content appropriation, but not as something else, since this could come into conflict with the inherent purpose of the work.

Whereas the issue of the authorial entitlement to object to uses of her work that are unrelated to the intended function of the work has never been discussed with reference to copyright works; it has been examined in the context of personal data. The use of personal data and information for purposes which are entirely unrelated to that for which they were initially released is normally considered to be a violation of the right of the data subject. Some argue that this holds true even when the personal data is anonymized, thereby excluding any possible harm to the data subject.[140] This is because, as Deryck Beyleveld has put it, 'the information has been obtained from the personal information [that the data subject] provided, and would not exist otherwise'.[141] Therefore, the individual's legal interest in personal information not being

136. See above (n 79 and accompanying text).
137. See Part One, § 3.
138. See above (nn 89–92 and accompanying text).
139. See above (n 91).
140. See G Laurie, *Genetic Privacy. A Challenge to Medico-Legal Norms* (CUP 2000) at 224–6, and D Beyleveld 'Law, Ethics and Genomics', *Business Briefing: PharmaTech* (2001) 30.

used in ways contrary to their judgement and beliefs is not exhausted with the first release of the information.

Should a similar approach be applied to copyright works, then the author might have a legal interest in that her work is used *as work*, and not for entirely unrelated purposes. Below, we examine how some principles of data protection law could apply, by analogy, in respect of the automated processing of copyright works.

4 Automated processing of works as data: a lesson from data protection law?

With mass digitization, works are dealt with as containers of data, and the copies of works are automatically processed. This bears an instantaneous analogy to the protection afforded to personal data, which in Europe is governed by Directive 95/46/EC.[142] Article 2(b) of this Directive defines the 'processing of personal data' as 'any operation or set of operations, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction'.[143] This definition of automated processing bears a strong resemblance to the automated processing of copyright works. The processing of works is an operation carried out by automated means and includes the collection of works, their recording via scanning and OCR-ing, their storage in a corpus of data, their consultation or use for non-consumptive research as well as their dissemination to machines. Deprived of intelligibility, the millions of books that comprise the corpus of a digital repository become the content of a massive database containing any potentially valuable information related to the works and the uses thereof. It was probably for this reason that some commentators have compared Google Books to the Human Genome in terms of historical significance.[144] But perhaps historical significance is not the only analogy that can be drawn from these two cases; it may be that an analogy of copyrighted works to personal data can be substantiated in terms of automated processing, since the structure and function of databases containing copyright works or personal data is astonishingly similar. Below, we examine the case of population genetic databases, as an example where the use of personal information for purposes different from that for which it was initially released – and for which informed consent were obtained – has been successfully regulated by law.

141. Ibid.
142. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the movement of such data, OJ L 24/1, 30.1.98.
143. An early definition of 'automated processing of data' is included in Article 2(c) of the Convention for the Protection of Individuals with regards to Automatic Processing of Personal Data Strasbourg, 28.1.1981, European Treaty Series, No 108. This Article reads that '"automatic processing" includes the following operations if carried out in whole or in part by automated means: storage of data, carrying out logical and/or arithmetical operations on those data, their alteration, erasure, retrieval or dissemination'.
144. Letter from Gregory Crane (n 49). In Pamela Samuelson's view, a project of mass digitization should ideally work as 'a kind of Human Genome Project-like initiative […] implemented by the major research libraries themselves working in cooperation with one another.' Samuelson (n 11) 1367.

(a) Use of personal information for unrelated purposes: the case of population genetic databases

Population genetic databases are large collections of biological samples and personal details of individuals belonging to a given population, which aim at covering a whole population of a region or a significant sample of it.[145] One of the central legal questions about these collections is whether the consent of individuals must be sought before carrying out certain uses of the information. Central to this question is what form such consent should take. It is a well-established principle that informed consent must be given for personal information to be used for medical research, in the form of 'explicit and specific consent'.[146] This includes both consent in having personal information and biological material included in the database *and* consent in making use of information and material for a specific research purpose. However, obtaining new consent for uses that were not predictable at the time the information was released might be too burdensome or even impracticable.[147] In these circumstances, 'generic consent' may be acceptable. So, once an individual has consented for her data to be used in a given research, there is a reasonable presumption that she will give consent to other uses that fit in the same objective of promoting public health.[148] For instance, in the Icelandic Health Sector Database, one of the largest population genetic databases in the world,[149] material was included on an opt-out basis, even though it was initially collected without consent. This meant that individuals had the right to have their material excluded from the database without their consent being required to collect and make use of it.[150]

However broad in scope, the doctrine of 'general consent' is subject to limitations. First, the consent cannot be presumed to cover *any* uses of the information, but only those that are reasonably consistent with a goal of public utility. This is because, as Graeme Laurie observes, when personal information are released for a given purpose there is a 'legitimate expectation of use',[151] namely an expectation that the information is not used for entirely unrelated purposes. Second, even when the data are collected and used without specific consent, the individual has to be given the right to opt out. Although this right can be exercised unconditionally, there is also a right to object to specific uses of own information or genetic material on grounds of moral or personal beliefs.[152]

145. There are two purposes for creating a population genetic database: criminal law enforcement and medical research. See eg the United Kingdom National DNA Database, at <http://www.npia.police.uk/en/8934.htm>. Population genetic databases may be used in combination with other information to conduct, for instance, investigations on the interactions between genes, environment and lifestyle. JV McHale, 'Regulating Genetic Databases: Some Legal and Ethical Issues' (2004) 12 Med L Rev 70–96, at 72.

146. *Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects* (as revised, 2008), art 24–29.

147. McHale (n 145).

148. To the House of Lords Selected Committee on Genetic Databases, 'we believe that it can generally be presumed that individuals are content for data about them to be used for the common good, provided that their personal privacy is protected' (quoted in McHale (n 145) 82).

149. McHale (n 145) 72–3.

150. For a critique of the Icelandic approach see Laurie (n 140), 287–93.

151. Laurie (n 140) 226.

152. E Histed and D Beyleveld, 'Betrayal of Confidence in the Court of Appeal' (2000) 4(3&4) Med L Int 277–311, at 295. For example, 'Roman Catholics might object to new contraceptive methods being developed from information they have provided. Those who disapprove of the

Should we translate these principles in the case of digital libraries, we could argue that works can be safely included in the corpus, on a non-display basis, and on the ground of 'generic consent' from authors in having their works made searchable by the public. This would include consent in having the work automatically processed for purposes that are reasonably consistent with this goal. Copying for indexing and search may be safely considered to be consistent, in spite of the fact that it formally infringes the reproduction right. On the other hand, automated processing for data mining and text mining are certainly exceeding the scope of such implied consent! The more these uses are *unrelated* with uses of the work as work – and for which a 'generic consent' can be presumed to be given once the work is published – the more the uses are likely to be restricted.

(b) Processing data – processing works: purpose limitation and proportionality

Directive 95/46/EC lays down the framework under which the processing of personal data may be lawful.[153] In principle, personal data should *not* be processed, unless certain conditions are met. These conditions fall into three broad categories, namely informed consent, proportionality and legitimate purpose. Under which conditions may the same principles apply to the automated processing of *works*?

Works and personal data are both substantive determinations of one's personality.[154] This is empirically manifested by the fact that both works and personal data are disclosure-sensitive; they are meant to be kept private unless their rightful owner wishes to disclose them to the public. They differ, however, as to the *purpose* for which they are disclosed. As we have discussed earlier, works are published with the aim of being shared with other human beings, which, by means of publication, are put in the condition of appropriating the intellectual content of the work. Personal data, normally, are disclosed only for specific purposes, such as the delivery of a service or the completion of a process. Any use beyond the specific purpose is restricted, simply because, unlike works, personal data are not meant to be public *per se*. While works are acts of communication, i.e. words expressly addressed to the attention of the public, personal data are un-addressed signs which *may* be communicated to others in given circumstances.

In spite of this essential difference, works and personal data share at least one element in common: they are both disclosed *for a purpose*. Clearly, unlike with personal data, the 'specific purpose' for which a single work is disclosed cannot be taken into account in copyright, which is and remains a content-neutral law. Rather, in case of works, the *generic* purpose of allowing appropriation of the content – whatever the content is – by members of the public can and must be taken into account. If this is the case, when automatically processed, copyright works also raise an issue of 'legitimate expectation of use', namely an expectation that the work is not processed for purposes which are entirely unrelated to that of making the work's content appropriable by humans. Just as the anonymization of personal data does not 'purify' all subsequent uses of the data,[155] the non-display of copyright works does not indemnify each and every use of works. Only uses which the author may *legitimately expect*, in

policies of some pharmaceutical companies towards developing countries might object to these companies profiting from their information. The patenting of human sequence is integral to the process of the new drug development, but some consider this to be immoral' (ibid).

153. Directive 95/46/EC.

154. See discussion above, Part One, § 3c.

155. See above nn 140–41 and accompanying text.

order that work is *accessed* by the public, are indemnified by the non-displaying provision. Still, uses *exceeding* this purpose cannot be deemed to be lawful on the mere ground that they do not display the work to the public.

The concept of legitimate expectation is closely related to the 'purpose limitation principle' and to the principle of 'proportionality'.

Under the purpose limitation principle, the processing of personal data is lawful only when it is carried out for specified, explicit and legitimate purposes and it should not be done in a way that is incompatible with those purposes.[156] Practically, this means that the data controller should state a legitimate purpose for which the processing is done and should not exceed the scope of this purpose. As Kotschy rightly observes, the controller who wishes to process data should at the moment of their collection define a specific and lawful purpose for which the processing of the data is required.[157] Should the purpose limitation principle apply as regards the automated processing of copyrighted works, the digital copies of works should be processed only in the context of a specified legitimate purpose; processing that exceeds this purpose would be infringing. In mass digitization, and in the Google Books project in particular, this purpose should not exceed the legitimately expected use of a work as work, namely its use in the context of content appropriation by humans. In this light, automated processing for indexing and search can qualify as a legitimate purpose. Any further processing, such as computational analysis and data mining, may exceed the scope of this purpose and could be considered as infringing.

The proportionality principle applicable in respect of the processing of personal data is laid down in Article 6 of Directive 95/46/EC. Under this principle, personal data may be processed only insofar as the processing is adequate, relevant and not excessive in relation to the purpose for which the data are collected and/or further processed.[158] In this context, 'relevance' should be interpreted as strictly as possible.[159] Applied in the context of copyright, this principle could read: 'the work may be processed only insofar as this is adequate, relevant and not excessive in relation to the purpose for which it has been reproduced'. Indeed, the application of the proportionality principle would mean that copying for automated processing should be relevant and not excessive in relation to the purpose for which works are normally published; this includes the enhancement of accessibility through indexing and search. Yet, there is an implicit limit here: accessibility should not affect the author's exploitative rights, namely it should not go beyond the modes of exploitation that the author is legally entitled to employ.

CONCLUSION

Whereas, up to date, the use of copyright works was associated with content appropriation by humans, works can now be automatically processed for various purposes without their expression being displayed and without their content being appropriated

156. This flows from Article 6(b) of Directive 95/46/EC.
157. W Kotschy, in A Büllesbach, Y Poullet and C Prins (eds), *Concise European IT Law* (Kluwer Law International 2006) 44.
158. P Carey, *Data Protection: A Practical Guide to UK and EU Law*, 2nd edn (OUP, Oxford) 108; F Ferretti, *The Law and Consumer Credit Information in the European Community: The Regulation of Credit Information Systems* (Routledge-Cavendish, London & New York 2008) 168.
159. Kotschy (n 157) 45.

by humans. Certainly, the regulation of these activities and the definition of their legal boundaries cannot be left to private ordering mechanisms. The use of works on a non-display basis is massive and entails the potential of unprecedentedly changing the way we use our cultural heritage. So, while mass digitization promises accessibility to all world information, the legitimacy of its underlying activities remains questionable. Even though non-display uses may qualify as transformative and, hence, permitted in the US, this is not the case in Europe. From the perspective of copyright principles, too, the legitimacy of non-display uses is contestable. First, non-display uses represent a very lucrative market for the near future, whereas, at the same time, they take place without authorial consent and without any sort of remuneration to authors. Secondly, non-display uses could have some unprecedented impact upon the way in which an author chooses to 'speak' freely to the public. This is because the associations that non-display uses are in the capacity of generating may interfere with the function of a work as an act of communication. Finally, and most importantly, non-display uses may not fit the sole purpose and function for which works are created and disseminated, namely their appropriation by other human beings.

Under the copyright doctrine, there are two main ways of dealing with copies of works. Those are the *dissemination* and the *use* of copies. We have seen that automated processing – an activity which underlies every so-called 'non-display use' – does not fit comfortably in any of these kinds of dealing with copies. It entails dissemination, but not dissemination in the conventional sense; copies are distributed to machines and networks and not to human beings as has been so far the case. At the same time, automated processing cannot purely qualify as a 'use' of works, despite the probably misguiding employment of the term in the context of 'non-display uses'. We have noticed that while copyright has – up to date – been about uses *of* works in terms of appropriating the content, automated processing involves uses *on* works. In this context, we have argued that when uses *on* works facilitate uses *of* works, i.e. when they are subsidiary or ancillary to conventional uses, they should raise no infringement issue. This includes automated processing for indexing and search. Nonetheless, when uses *on* works become substitutive, or totally irrelevant, to uses *of* works, their legitimacy is not equally straightforward. These latter categories of uses *on* works include automated processing for computational analysis and for data mining.

This distinction between uses *on*, and *of*, works inevitably raised the question as to whether this shift in the way that works are used means that the works cease to be used as works but are used as something else. We have observed that while digitization turns works into 'de-materialized' 'content', mass digitization turns content into 'de-intellectualized' 'data'. The fact that works are not used as works but as data does not mean that authors are not in capacity of objecting to this kind of use. Quite on the contrary, some non-display uses contravene, and interfere with, the internal purpose and function of works, namely content appropriation, as the sole destination and the value of a work. In this light, this kind of use may be in conflict with authors' rights.

Given that some non-display uses may be infringing from the perspective of copyright principles and since the automated processing of works as data and that of personal data present striking similarities, we have sought some answers as to the regulation of non-display uses from data protection law. We have seen that in the context of genetic databases, the use of personal data should be in compliance at least with the generic consent of the data subject. The same ought to apply to the automated processing of digital works, whereby generic consent would include the use of works in a way that the author is legitimately presumed to expect, most notably for content

appropriation. At the same time, digital copies of works should be processed only in the context of a specified legitimate purpose, not exceeding the legitimately expected use of a work as work; copying for automated processing should be relevant and not excessive in relation to the purpose for which works are normally published, namely to be used in the context of content appropriation.

This new form of dealing with works promises to change for ever the way in which humankind will make use of information, knowledge and the whole of its common cultural heritage. Along with the promise, however, comes the challenge; almost inevitably, this new kind of copyright use begs for a redefinition of the contemporary copyright space: a space in which works will continue to flow though the visible and invisible channels opened up by technological change, and to feed human coalescence.