KNOWLEDGE, INFORMATION AND VALUES IN THE AGE OF MASS DIGITISATION

Maurizio Borghi

Man's final conquest has proved to be
the abolition of Man
(C.S. Lewis)

One of the most common assertions of our times is that the networked information technologies change the way in which the humankind accesses, shares and "produces" knowledge. There are in fact three, plainly observable facts that distinguish the present technology from the past. These facts, as well as their actual or potential consequences, are variously articulated in most of today's narrative of the so called "cyberspace". The first fact is, the increased speed in the flow of information and easiness in access. The second is, the expansion of the quantity of information which is available to people. The third is, the qualitatively different way of making use of information. This translates in the enhancement of participatory culture, peer-production and promotion of democratic values.[1]

Increases in speed, quantity and quality are facts, and as such they strike most of the time our attention and call upon our capacity to think and to act. What keeps them together is apparently nothing but a pure incremental operator "plus": *more* speed (and easiness), *more* quantity, *better* quality. This incremental operator, this "plus"-trait, is the common ground of the three observable facts, as well as of all their possible combinations. It does not, however, equally capture our attention. On the contrary, it remains most of the time unnoticed as such, or it is set aside as purely self-evident. This is not accidental. Since the "plus"-trait provides the background upon which increases in speed, quality and quantity become conspicuous as striking facts of our age, it remains invisible in its provenance. To our common understanding, information simply *happens*

[1] See e.g. Yochai Benkler: *The Wealth of Networks. How Social Production Transforms Markets and Freedom* (New Haven: Yale University Press 2005); Lawrence Lessig: *Code: And Other Laws of Cyberspace*, Version 2.0 (New York: Basic Books 2006); James Boyle: *The Public Domain. Enclosing the Commons of the Mind* (New Haven: Yale University Press 2008).

to flow the more and more speedily and easily, and, as a consequence, it *happens to* increase in quantity while actually, or hopefully, the quality of its use *does* improve. The "plus"-trait encounters the man of our age with the disarming cogency of a compelling fatality.

A proper insight on how this cogent "plus"-trait hiddenly orients all instances of our age, and determines the character of what we call "the cyberspace" in particular, is not yet within reach. As a matter of fact, very few thinkers in our times have dared to engage into clarifying this trait in its provenance and meaning.[2] For the time being, I will focus on one aspect of the information age which interprets the "plus"-trait in a peculiar and to some extent unique manner. This is the advent of the digital format and the digitisation, namely "the migration of all we know in the universal form of digital bits".[3] More specifically, I will focus on a relatively recent development of the said migration. This is the conversion of all past "printed culture" – books, papers and every recorded document – in digital format on an industrial scale. It is an enterprise which mobilises enormous energies from public and private companies, citizens and governments. It is commonly referred to as *mass digitisation*.

What is characteristic with this enterprise, is that it comes with the promise to gather "all human knowledge" – past and present – in one single place and to make it quickly and easily available to everyone. Face to this unprecedented promise, it seems that even elemental questioning – for instance, on legal and regulatory issues of the said enterprise – is fated to appear as unnecessary distraction.[4] Being represented as the spectacular realisation of a venerable old dream, namely that of building the "universal library of all human knowledge", mass digitisation is perceived as the inexorable departure from what we have so far called "books", "libraries" and "readers" – briefly the sources and pillars of our knowledged coa-

---

[2] Martin Heidegger's attempt marks undoubtedly the most advanced post in this clarification. See Martin Heidegger: *Gesamtausgabe* (Frankfurt a.M.: Vittorio Klostermann 1975-). As it has been correctly observed, "the dominant philosophical fact of the past thirty years has been and is the appearing – with a frequency of two volumes per year – of Martin Heidegger's *Gesamtausgabe*". Ivo De Gennaro: "Why Being Itself and Not Just Being?", in *The New Yearbook for Phenomenology and Phenomenological Philosophy Vol. VII* (2008), p. 159.

[3] Kevin Kelly: "Scan This Book!", in *The New York Times Magazine* (14 May 2006), p. 2.

[4] Commenting on the Google Books case (see *infra* note 11), Kevin Kelly claimed: "The courts may haggle forever as this complex issue works its way to the top. In the end, it won't matter; technology will resolve this discontinuity first". Kelly: "Scan This Book!", p. 8.

lescence.[5] In this context, the cogent "plus"-trait operates more compellingly than ever.

In the following, an attempt will be made to resist this cogency and to clarify some traits of this phenomenon. I will break up my attempt into six stages. First, I will present some historical facts on mass digitisation, and I will briefly outline the role of Google as the major, and to an extent *unique*, player in this enterprise. Second, I will discuss the rationales that are commonly put forward to justify mass digitisation. Far from representing just another reproduction technology, digitisation is rather the conversion of books into computable objects. The effects of this conversion on an industrial scale, as well as its implied meaning, will be discussed in the third and fourth stages. In the last two stages I will illustrate how mass digitisation alters the meaning of the *whole* in which books consist. I will show that, far from entailing a mere technological or cultural change in the way in which the humankind relates to books, this alteration corresponds to an enigmatic transformation in the relation of man to *truth*.

The Road to Mass Digitisation

The idea of creating collections of books in digital format, even on a large scale, is as old as the internet. Libraries and non-for profit organisations have in fact engaged in digitisation projects since the mid 1990's and digital libraries have been created by initiative of both public and private organisations. These libraries are very diverse in terms of dimension, quality and ambition. The ideal of creating a "universal library" that could virtually include every book ever printed has been in sight ever since. One of the earliest and most prominent projects, the Internet Archive, was the first to have the vision and the ambition of creating a digital copy of "all world's books", as well as of other items in the public domain such as sound recordings, images and films.[6] The Carnegie Mellon Million Book project, which started in 2001 and joined its efforts with the Internet

[5] "Whatever the future may be, it will be digital". Robert Darnton: *The Case for Books* (New York: Public Affairs 2009), XV.

[6] http://www.archive.org. The project to create "one web page for every book ever published" has been launched by Internet Archive in 2006 under the name Open Library (http://openlibrary.org), and it currently includes over one million books.

Archive, declared the "long-term objective" of "captur[ing] all books in digital format".[7]

The turning point occurred when Google entered the business of digitising of books in 2005. The Google Print project, which then became Google Book Search or simply Google Books, gave to digitisation an impressive and so far unrivalled bound of power. Digitisation of books was envisioned as part of Google's corporate mission to "organize the world's information and make it universally accessible and useful".[8] Books are part of "the world's information" – and probably not the less "useful" part of it. The stated goal of Google is the conversion of printed books into digital format, and the inclusion of the whole corpus of books into its database of searchable information. As a matter of fact, books now feature in Google's search engine results alongside any other "resource", and the whole corpus was alleged to contain over fifteen million books by the end of 2010.[9] One might argue that Google has simply engaged more efficiently on a route paved by others. What is different with Google Books, however, is not just the scale of the project, but also the quality of the "mass" effect.[10] First, Google does not aim at creating specific collections of digital copies: it simply aims at digitising *everything*, namely every book ever printed in every language. In this respect, human intervention in selection and arrangement of the material, as well as in the quality control of the digital copies and their indexing, is reduced to a minimum. Second, Google digitises every book, regardless of quality and value, and irrespective of their copyright status. From a legal perspective, this is the most disruptive and unique feature of the project.[11] Third, Google Books, while working in partnership with the major libraries of

---

[7] Quoted in Karen Coyle: "Mass Digitisation of Books", in *Journal of Academic Librarianship*, v. 32, n. 6 (2006).

[8] http://www.google.com/corporate.

[9] James Crawford: "On the Future of Books" (14 October 2010) at: http://booksearch.blogspot.com/2010/10/on-future-of-books.html.

[10] Coyle: "Mass Digitisation of Books". For a thorough discussion of Google's role in the global information society see Siva Vaidhyanathan: *The Googlization of Everything* (*And Why We Should Worry*) (Berkeley and Los Angeles: University of California Press 2011).

[11] In-copyright books are digitised but are not displayed to the public, apart from short excerpts in response to search queries. Soon after the project was launched, a class action for copyright infringement started in the U.S.A. on behalf of American authors and publishers (*Author's Guild, Inc. v Google, Inc.*, No. 1:05-CV-08136, filed 20 September 2005). Three years later the parties proposed an agreement to settle the case (Ibid., filed 28 October 2008), which was later amended (Ibid., filed 13 November 2009) and eventually rejected by the Court on 22 March 2011.

the planet, is and remains entirely a for-profit initiative. While it certainly does not represent the first and only privately owned digital repository, it is to date the only private player in *mass* digitisation.[12]

Since Google Books established itself as the major, and to a certain extent unique, mass digitisation project, other enterprises have followed its route on similar or alternative basis. In 2006 the European Commission announced the decision of promoting a European counterpart of Google Books, by joining the efforts of all European cultural institutions in a single framework.[13] The portal of the European "digital heritage", Europeana, was launched in 2008.[14] This is an entirely publicly funded initiative which shares only some of Google's "mass" attributes. It aims at covering all Europe's cultural heritage by creating a universal platform of "digital objects". Unlike Google, however, it operates a selection of the material through its partner institutions and, most importantly, it does not digitise in-copyright works.

What is characteristic with mass digitisation is that it comes with a peculiar sense of urgency and compulsiveness. To be sure, the modus operandi of Google – scanning first and asking questions later[15] – has provoked a hastening of all programmes, coupled with a peculiar loss of sense of proportions. One can observe, for instance, how formulae such "all world's books" or "all cultural heritage" are used as a sort of mantra, deprived of any experienceable content.[16] Having all books searchable online appears as the necessary condition to make them part of a "truly democratic", non-elitist, information society; it is as if humankind has been waiting for three millenniums for the moment in which all its

---

[12] A project launched by Microsoft in December 2006, called Live Search Books, was dismissed two years later after having digitised 750,000 books.

[13] Commission Recommendation of 24 August 2006 on the Digitisation and Online Accessibility of Cultural Material and Digital Preservation, OJ L 236, 31 August 2006, pp. 28-30.

[14] http://www.europeana.eu.

[15] Google Books is the only digitisation project where permission to display books is sought *after* the books have been scanned. This practice has been described by David Nimmer as "turning copyright law on its head" (Fairness Hearing Transcript, *Author's Guild, Inc. v Google, Inc.*, No. 1:05-CV-08136 (18 February 2010), p. 46).

[16] Even from a mere quantitative point of view, the concept of "all world's books" is very unsettled. At the outset of the Google Books project, it used to correspond to about 30 million books. The number increased soon to 50 millions. Now the latest Google's account speaks of 129,864,880 unique items. Cf. Leonid Taycher: "Books of the world, stand up and be counted!", in *Google Blogspot* (5 August 2010), at: http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html. Whatever the number may be, the point is to remind that all these books "cannot be read by a human".

knowledge could have been gathered and stored in one single place to be instantaneously accessed at no cost. In this light, scanning books has been even viewed as a "moral imperative"[17] and a "moral obligation".[18]

Human tendency to hyperbole has a major part in it, but it does not explain all. Digitisation is presented as a – if not *the* – compelling urgency of the information age, and its "mass" quality is the inevitable consequence of this pressure. But what is the purpose of mass digitisation? And why is it so important to entail no less than a moral obligation to our generation? These questions are addressed in the following.

Why Digitise in Mass?

The first argument which is put forward to justify the urgency to digitise printed material is the necessity of preserving cultural heritage. As a matter of fact, preservation is the reason why public libraries and archives have started digitising their collections even before the advent of the internet and of large-scale digitisation projects. Most copyright laws acknowledge the significance of this activity by providing specific exemptions for preservation and archival purposes.[19] It is true that libraries or single collections of unique materials may be ruined or even destroyed by natural calamities or human negligence.[20] In this respect, digitising all world's books bears no substantial difference from what has been always done in the past to protect works from the risk of getting lost on damaged, namely: reproducing them in many copies and store these copies in safe places.

However, if the main reason for digitising books was preservation, mass digitisation does not work well to this end. Librarians have pointed

---

[17] Peter Branley, director of technology for the California Digital Library, quoted in Kelly: "Scan This Book!", p. 8.

[18] "Digitisation is more than a technical option, it is a moral obligation" (*The new Renaissance*, Report of the "Comité des Sages" on bringing Europe's Cultural heritage online, Luxembourg: Publications Office of the European Union 2011, p. 14).

[19] See e.g. Directive 2001/29/EC, art. 5(2)(c): "Member States may provide for exceptions or limitations to the reproduction right [...] in respect of specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage". Art. 5(2)(d) allows Member States to permit the "preservation" of recordings of broadcasts "in official archives [...] on the ground of their exceptional documentary character".

[20] Sergey Brin, co-founder of Google, reminds that "The famous Library of Alexandria burned three times, in 48 BC, AD 273 and AD 640, as did the Library of Congress, where a fire in 1851 destroyed two-thirds of the collection." He comments: "I hope such destruction never happens again, but history would suggest otherwise" ("Google hits back at book critics", *BBC News* (9 October 2009)).

out the incongruity between the exigencies of preservation and those of projects like Google Books, whose digital copies are beyond the standards of preservation-quality copies.[21] Most importantly, digitisation is not merely a matter of "reproducing" content. Making digital copies is only preliminary for other operations to take place, namely activities that have nothing to do with preservation. Some of these activities are directly related to a further purpose for digitisation, namely that of enhancing the dissemination of works.

As it is often stated, mass digitisation primarily purports to enhance the *access to* our past and present knowledge that is deposited in books. This "access argument" is sometimes referred to as the goal of building the "universal library of all human knowledge". Brewster Khale, the founder of the Internet Archive, puts this goal in evocative words: "We can provide all the works of humankind to all people of the world. It will be an achievement remembered for all time, like putting the man on the moon".[22] Whereas such achievement would certainly be attractive and worth-pursuing, question remains as to whether mass digitisation is just this, namely a technical enterprise aimed at building the "universal library". The point is not whether this goal is worth pursuing. It is whether mass digitisation will create a "library" at all – that is "a place set apart to contain books for reading, study, or reference."[23] Preservation and dissemination represent the very mission of libraries, the reason of their existence. However, in mass digitisation, books are not digitised to be preserved from calamities. Nor – perhaps less obviously – are they digitised (only) to be read by people. Books are digitised to be further *used* in a different way.

Digitisation: From Reproduction to Computation

Any kind of reproducible instance, including images, films and sound recordings, can be digitised. But what does this mean? Contrary to what is

[21] Coyle: "Mass Digitisation of Books" ("There is an undeniable conflict between 'mass' and 'preservation' for the digitization of hard copy materials"). Moreover, to fit the purpose of preservation, digital copies could simply be stored in data silos and wait for the passing of the dark ages. Yet, this is clearly not the case. Digital copies are not just stored, but at the same time they are *used* – and, as will be shown in a while, they are used in an unprecedented way.

[22] Quoted in Kelly: "Scan This Book!", p. 1.

[23] *Oxford English Dictionary*, entry "library".

commonly thought, the conversion in digital format is not just a reproduction of the work in a different "medium", or, otherwise put, a different technical representation of the same object. This could have still been the case with the so-called "analogue" reproduction techniques, such as photography or reprography.[24] To be sure, a digital copy is, technically, a representation through discrete values ("digits") of an object which is represented through continuous values (sounds, colours, meanings). The essential trait of digitisation lies precisely in this shift from "continuous" to "discrete" values. This shift marks a step where it is not the reproduction as such that is relevant, but the fact that the reproduced object becomes fully *computable*. A digitised work is primarily a work that can be processed by computers.

The perspective according to which works are represented as systems of computable values, and as units of values, is not created by digitisation. Rather, digitisation presupposes it. Still, the conversion of (continuous) values into (discrete) values, that is into "digits", is an essential step to capture a unit of value into a purely computational perspective. A continuous value is already a value, *i.e.* an instance or a unit of measurement "to count with".[25] However, its computational potential remains still largely unexpressed. Counting with continuous values requires humans who dedicate themselves to deploy their own capacity of calculus, both individually and collectively. The human capacity of calculus organises itself into accounting divisions that take the form of what we nowadays call "sciences".[26] Yet, these divisions rely on human capacity of calculus, that is on a resource which is limited. The potential of values requires that this limit is overcome. Machines, and namely computers, are so far the most efficient tools to replace man as operator of the calculus with values.[27] Since the only limit to computability is the computation power

[24] A reprographic copy of book's pages is still a book's copy, although maybe in a more portable fashion.

[25] The understanding of a painting as "visual art" or of a book as "text" is the effect of this counting with works in terms of (continuous) values.

[26] For instance, semiotics is the accounting division that deals with continuous values in the form of "texts".

[27] "So far" means: insofar as computers are not replaced by the next "thing". Computers are just the (transitory) consequence of the deployment of capacity of calculus. The plugging of all world's knowledge "directly into human brain with thin white cords" (Kelly: "Scan This Book!", p. 2) might well be the next stage, in the wait of the subsequent stage where digital bites will be transformed into edible molecules ("molecularisation of knowledge").

that can be supplied at a given stage, humanly-driven calculus is fatally replaced by the more powerful machine-lead computation.

Digitisation is the transition from (humanly-driven) calculation to (machine-lead) computation. But what is a "digit"? It is a symbol to represent numbers, and it is referred to the ten *digita* (fingers) of the hand. In the decimal system, these correspond to the basic symbols of numbering. In the context of computers, a binary digit or "bit" is the basic unit of a code capable of representing *any* value, through numbers, as a binary series of values (noughts and ones), and these in turn as a sequence of electric impulses.

The act of digitising an instance such as a literary work presupposes, implicitly, four subsequent conversions. First the instance is translated into continous values (*e.g.* a series of words "expressing" certain "ideas", which in turn an expressions of "altural trends" etc.); then values are represented through numbers, which then are converted into binary digits and eventually are turned into electric impulses. The driving principle of this fourfold conversion may be understood by reference to the etymology of the Latin word *digitus*. While *digitus* is commonly referred to the root **deik* (to indicate, to point out, to identify), similarly to *dicere* (to tell), Pianigiani relates it to the root **dak* or **dek* instead. This latter root bears the meaning of "catching" and "grabbing", as in the Greek words δέχομαι (to take, to gather), δόκος (trap, pitfall), δοκάνη (space capable of receiving something) and δοχός (capacious, able to hold). This meaning resembles that of the German and English word "Finger", which, according to Pianigiani, comes from *fangen* (to catch, to capture, to entrap).[28] The two etymologies – "to indicate" on the one side, "to capture" on the other – are not in conflict with each other. The *digitus*, the finger, is simultaneously both an indicator and a gatherer: it gathers the sense of something *while* pointing to it, and vice versa.[29] It is because the *digitus* has this twofold trait that the "digit" can be what it is, namely: the universal identifying-and-capturing operator of the computation of everything. Yet a peculiar shift occurs here. Where the *digitus* points out, the digit *identifies*; where the *digitus* gathers, the digit *captures* and *seizes*. Hence the digit is, simultaneously, the identifying warder and the sizing agent of the

[28] Ottorino Pianigiani: *Vocabolario Etimologico della Lingua Italiana* (1907), entry "dito". This etymology is not recognised by Duden, which relates *Finger* to *fünf* (five), see *Herkunftswörterbuch* (*Duden* Bd. 7) (Mannheim: Dudenverlag 1989), entry "Finger".

[29] In the well-known Zen saying "when the finger points to the moon, the silly men looks at the finger", the "moon" is precisely the whenever gathered-by-pointing instance, which the *digitus* aims at.

cyberspace. Digitisation is the process of identifying each instance as source of value and securing it in its full computability.

In this respect, mass digitisation is not primarily a project of reproducing works for preservation or access, plainly because it is not in essence a mere reproduction technology. Mass digitisation is the project of securing all world's works in their full computability. It is no surprise to read the following reported words of an anonymous Google engineer: "we're not scanning all those books to be read by people. We're scanning them to be read by Artificial Intelligence".[30] The use of the verb "read" in this last sentence is not just geek slang; books are actually "read" by artificial intelligence. This does not only take place in the sense that the viewing of books in digital format is necessarily mediated by computer programs and applications. Computers "read" books in the straightforward sense that, by supplying a surplus of computational power, they replace humans in the very function of calculating with each and every book as an instance capable of generating value.

Computing With Books: From Information Retrieval to Value Extraction

Once books are digitised, they become available for automated processing by computers. Such processing can extend from basic operations, that are functional to make books and the information therein contained timely and efficiently retrievable by users, to more sophisticated procedures aimed at directly extracting and elaborating information. These last activities are commonly referred to as data mining and text mining. "To mine" means: to extract value from books *qua* texts and data, as one can dig out minerals from the earth *qua* natural resource. The "value" that is extracted has not only, and not primarily, an economic meaning.[31] One can provisionally call it "informational value". The worth of such value is the fact of being employed in a computation process. When a value is compatible with computation and is usable in computation, it is then *useful*. Usefulness, in the sense of compatibility with a computation process, orients the extraction of value from digitised copies of books. To this

[30] Reported in George 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.

[31] As a matter of fact, the economic value of mass digital repositories is still largely uncertain.

end, processes of data mining and text mining are put in place to discover new knowledge from old resources, to locate patterns that remain invisible to human eyes and generally to find interesting, previously unknown associations.[32] The driving idea is that "the collective intelligence of a library allows us to see things we can't see in a single, isolated book".[33] Text mining consists in a range of natural language processing techniques applied to masses of texts. It is used, for instance, in the context of biology and medicine, where such techniques applied to wide corpora of scientific literature (articles, books, reports, etc.) may "automatically generate and rank hypothesis that a scientist can test in a laboratory".[34] Text mining "goes far beyond its cousin, information retrieval", by providing the "means to rapidly drill down to individual facts, rather than, like information retrieval, providing many documents to wade through".[35] For instance, statistical analysis of co-occurrence of biomedical concepts describing genes, drugs and diseases, in large corpuses of articles and papers is capable of discovering novel relationships between concepts "that have high probability of being biologically valid".[36]

This move from information retrieval to value extraction is a main driver in mass digitisation of books as well. The fact that books in Google Books and in other digital repositories become "searchable inside" is one of the most spectacular effects of digitisation.[37] However, this is only the condition for other kinds of automated processing to possibly take place,

---

[32] Text mining is applied, for instance, to wide corpora of scientific literature (articles, books, reports, etc.) "to automatically generate and rank hypothesis that a scientist can test in a laboratory" ("Text Mining and IP", Submission to the Independent Review of Intellectual Property and Growth (*Hargreaves Review*) from the National Centre for Text Mining, University of Manchester, May 2001, p. 2). Text mining "goes far beyond its cousin, information retrieval, [...] by providing means to rapidly drill down to individual facts, rather than, like information retrieval, providing many documents to wade through" (*ibid.*).

[33] Kelly, "Scan This Book!", p. 4.

[34] "Text Mining and IP", Submission to the Independent Review of Intellectual Property and Growth (*Hargreaves Review*) from the National Centre for Text Mining, University of Manchester, May 2001, p. 2.

[35] Ibid., pp. 2-3.

[36] Frijters et al.: "Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases", in *PLoS Computational Biology 6* (9) (2010) (reported in "Text Mining and IP", p. 6).

[37] "Search inside!" is a registered trade mark of Amazon Inc. Unlike other large-scale repositories, such as the Internet Archive and the partners of Europeana, Google Books allows users to perform search queries *inside* the books. This is the most visible point of departure from the library standard. See Pamela Samuelson: "Google Books is Not a Library", in *The Huffington Post* (13 October 2009).

including text mining. For example, textual analysis on a large corpus containing books in many languages may be used to refine and improve automatic translation.[38] But computational power driven by usefulness and implemented by machines takes over gradually most of the activities that are carried out by humans. Similarly to what occurs in the context of scientific literature, the analysis of sequences of words that repeat identically or just similarly across multiple books can provide information on connections between the "key ideas" that are "contained" in a book, more or less independently from their exact literal expression.[39] These technologies make clear that the main value of a corpus of millions digitised 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 [*scil.* the Google Books corpus] would not be made available for the purpose of reading the works. Instead, this group of works is intended to made available for computational analysis *on* works".[40] It is a fact that the whole corpus "cannot be read by a human".[41] Seemingly naturally, humans are then replaced by machines in the effort of "reading" "all these books". Human reading becomes nothing but an inefficient and perhaps obsolete technique of extracting information. Computers discover information that no human intelligence can ever extract, such as quantitative and qualitative data on "cultural trends" over centuries and across languages, migration of "ideas" from one place to another, evolutions of linguistic and cultural phenomena, influences of an author over other authors and vice versa.[42]

Extraction of value from digitised books does not only take place via text mining and quantitative analysis on texts (so-called "culturomics").[43]

---

[38] Franz Och: "Statistical machine translation live", in *Official Google Research Blog* (28 April 2006) (available at: http://googleresearch.blogspot.com/2006/04/statistical-machine-translation-live.html).

[39] Bill N. Schilit and Okan Kolak: "Exploring a Digital Library through Key Ideas", in *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.

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

[41] Jean-Baptiste Michel et al.: "Quantitative Analysis of Culture Using Millions of Digitized Books", in *Science Express* (16 December 2010), p. 1.

[42] Ibid.

[43] Ibid. p. 1.

Data gathered from search queries are used to improve search engine algorithms and targeted advertisement. Since, "the very worst [search] algorithm at 10 million words is better than the very best algorithm at 1 million words",[44] the more books are digitised and made searchable to users, the better the search engine works; in this respect, books are just collections of "words", *i.e.* data, to feed search results.[45] These data are useful in extracting information on users' behaviours, which may help directing advertisements to people running searches for a book or an author's work. To this end, data on the uses of books – how many times a book is being searched, by whom it is searched, for how long it is browsed, etc. – may be processed to create databases of user profiles.[46] Analysis of books' content *via* text mining, combined with processing of data on search queries on books, may prove to be a powerful means of extracting value from books without the books even being made available to the public for reading.[47]

Carried out on digital copies, all these activities share the quality of being *automated*. This means that they are performed through serial application of a prearranged set of instructions (an "algorithm") in order to obtain a result. After having determined the algorithm, the human intervention is limited to correct and refine the procedure where necessary.[48] In this respect, automatically implemented algorithms can replace manually executed methods. We have observed that computation has decidedly moved from information retrieval to value extraction. However, retrieval and extraction are not substitutive; they are fully complementary activities. The search engine technology is the organising principle that connects retrieval and extraction.

---

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

[45] Incidentally, it can be observed that words in the books that form the corpus of Google Books are only accessible from Google's search engine; they cannot be located by other engines like Yahoo! or Bing.

[46] By running users' queries against the database of book content, Google "could compile dossiers on individual users", which "would allow Google to personalize advertisements to or aim products at specific users" (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, p. 7).

[47] For a discussion of these activities in the light of copyright law, see Maurizio Borghi and Stavroula Karapapa, "Non-display uses of copyright works: Google Books and beyond", in *Queen Mary Journal of Intellectual Property*, vol. 1. n. 1 (2011), pp. 21-52.

[48] Or, to move the process to a different format of computation: for instance, in-laboratory validation of hypothesis generated through text mining ("Text Mining and IP", p. 6).

The case of search engine algorithms as applied to digitised books is particularly telling. While the activity of indexing and cataloguing books in a library presupposes the application of a method, or simply a scheme of classification, whose outcome depends ultimately on human judgment, in a mass-digital repository the organising principle is an algorithm which takes into account an increasing and virtually unlimited number of factors. These include, as we have briefly discussed, metadata on the book, as well as statistical associations with other books and other digital resources on the basis of search queries run by users on the book's content. This is made possible by an elemental, and almost unnoticed quality that books acquire when they are transplanted from ink and paper to digital format, namely the fact that they become searchable – and hence computable – on a word-by-word basis. This means that search is not just performed "amongst" books on the basis of external indicators, such as the title, the table of contents or the librarian's classification. By transforming printed pages into a text document, books become "transparent" to search engines. Consequently, the ordering principle of books is the "rank" that the algorithm assigns each time to the book (as to any other "digital resource") in the list of results corresponding to the particular search query of the moment. In lieu of stable classification ordering, books become "fluid" instances at the disposal of each computation need, however transient and ephemeral.[49]

The ordering principle is nothing but the instantaneous request to supply "useful information", that is a value compatible with the computational process of the moment. In a way, computation anticipates the requested value and *validates* it. In other words, it certifies that the value has been always already extracted *as* worth retrieving, *i.e.* useful. The search engine algorithm is the validator of universal computation.

* * *

To summarise the points that have been touched so far, mass digitisation carries on a threefold purpose, namely *preservation* of, *access* to and *computation* with books. The third purpose seems to be not only the most prominent one, but also that which orients the other two as well. Computation is carried out by machines, and it includes information retrieval and value extraction. Retrieval and extraction are the two sides

[49] "Google's method of relying on the collective and active judgment of millions of Web users seems in the abstract to realize one of the most influential theories of epistemology: American pragmatism" (Vaidhyanathan: *The Googlization of Everything*, p. 60).

of the same coin, whose organising principle is the search engine as supreme validating instance.

One might argue that machines are here simply new intermediaries to reach more efficiently the task that humankind has always faced, namely that of reading and interpreting present and past knowledge and building new knowledge on it. Although machines may displace man as the sole reading being in this world, they do not replace him as the last beneficiary of reading. Mass digitisation does not get rid of man. Or does it?

One can speculate on whether, and to what extent, automated processing of books will change the "reading habits" and the general attitude of humankind towards books, and on whether these potential changings will bring about threats or opportunities. These are however issues that are deliberately left outside the scope of this chapter. The point here is not to appraise a sociological or cultural development in the man-book relationship; it is rather to question the emergence of an unprecedented understanding of what a book is. More precisely, the point is not the individual or collective relationship with books, but the meaning of the relations to books *as a whole*. Books may be still individually experienced as they have always been, namely by reading, studying and referencing. However, the *whole* of which each book forms part, is about to change essentially. I will address this change in the following.

What are Books?

Although from a historical point of view the printed book was born around the end of the 15th Century, it was not until three centuries later that its sense was determined. To an extent, the true date of birth of the modern book is 1785. It was then that Kant addressed explicitly the question *what is a book?* in the *Metaphysik der Sitten*. Kant's answer to this question is as simple as rich in implied meanings: the book is a speech that someone, by means of a mediator, holds before the public. This characterisation of the book as a representation of a public address, namely as the "mute instrument" of a public act of speech in one's own name, is the premise to the Kantian argument on the unlawfulness of unauthorised reprinting. The mediator, that is the publisher, speaks to the public in the name of the author. Reproducing a book without the author's consent is comparable to speaking in someone else's name without having a mandate to this end, or alternatively to compelling someone to speak against

his will. This is a wrong against the autonomy of the human being as speaker, and this is why counterfeiting books is considered unjust.[50]

The book is the carrier of an autonomous, freely determined, public act of speech. Autonomy of speech is an end in itself, and its pursuit does not need to be justified on other grounds – for instance, on the ground of the "social utility" that such act may bring about. However, autonomy is far from being a mere individual claim. What is at stake with autonomy of speech is not individual self-expression. This becomes clear once we observe the fact that people struggle to defend the right to speak publicly. Why, for instance, do people fight for the freedom of the press? It is because, as Kant observes, "if this freedom is denied, we would thereby lose a very potent means for proving the correctness of our own judgments and we would be left to error."[51] The ultimate reason to defend the freedom of the press, that is the freedom to speak publicly in one's own name, is because this is a means for proving the truth of one's own judgment. Denying or undermining such freedom means to be left at the mercy of untruth. To be sure, individual self-expression may be a worth-pursuing value, but it is certainly not an ideal which is worth fighting for *as a people*. Furtherance of truth, not self-expression, is the very reason books exist.[52]

A book conveys a speech in one's own name, and this is the addressing of one's own judgment to the public in order to have the truth of this judgment proved by other intelligences. This understanding of what a book is rests upon a determination of truth as *correctness*. For Kant, as for the whole modern thought, truth is essentially the correctness of judgment, namely the accordance of the judgment with its object.[53] In this respect, the address of one's own judgment to others is a means, although indirect, of proving the correctness of the judgment itself. This is because

---

[50] Immanuel Kant: *Die Metaphysik der Sitten*, AB 127-130. See also Immanuel Kant: "Von der Unrechtmäßigkeit des Büchernachdrucks", in *Berlinische Monatsschrift*, vol. 5 (1785), pp. 403-441 (reprinted and translated in Lionel Bently and Martin Kretschmer (eds.): *Primary Sources on Copyright (1450-1900)* (2008), www.copyrighthistory.org). For an understanding of the subject matter of copyright as an "act of speech" see Abraham Drassinower: "Authorship as Public Address: On the Specificity of Copyright vis-à-vis Patent and Trade-Mark", in *Michigan State Law Review*, vol. 1 (2008), p. 1999.

[51] Immanuel Kant: *Anthropologie in pragmatischer Hinsicht*, A 128.

[52] This point is developed in Maurizio Borghi: "Copyright and Truth", in *Theoretical Inquires in Law*, vol. 12, no. 1 (2011), pp. 1-27.

[53] Heidegger has repeatedly clarified this point and its meaning in his writings. See e.g. Martin Heidegger: *Einleitung in die Philosophie* (*Gesamtausgabe* Bd. 27) (Frankfurt a.M.: Klostermann 1996), p. 267.

"the presumption, at least, arises that the agreement of all judgments with each other, in spite of the different characters of the subjects, rests upon the common ground of the agreement of each with the object, and thus the correctness of the judgment is established."[54] Incidentally, one can observe that this means of proving the truth, *i.e.* the correctness of one's own judgment, is equally important when it comes to "rational" truth, that is when it comes to judgments whose correctness does not depend on subjective facts. These are typically the judgments that belong to mathematical and philosophical knowledge. Although, as Kant observes in the *Anthropology*, "in philosophizing we do not need, and we should not need to appeal to the judgment of others to corroborate our own" (*i.e.* truth is not a pragmatic instance, "what men think is true"), the fact that a view which someone has expounded publicly finds no support may still "give rise to suspect of being in error"[55] (*i.e.* truth calls for accordance between men). And in this respect mathematics itself "is not privileged at all", since if, initially, there was no "perception of the fact that the judgment of the land-surveyor regularly agreed with the judgment of all the others working diligently and carefully in the same domain, mathematics itself would not be able to be free from the fear of falling into error".[56]

Both the book as a single instance and the whole of books receive a specific determination from this understanding of truth. First, the book as such is the representation of a speech, and namely a publicly addressed judgment, which means that the book is an instance provided with own internal coherence. A speech has a beginning, a core and an end. But, most importantly, it has its own peculiar and even unique way of drawing to the source, that is, of crafting the language and the concepts that articulate the address to the public.[57] Whatever is the "value" of a book, each book is at least presumptively a fully autonomous and self-standing instance, which represents a "unique" act of addressing a judgment to the public. In this respect, the "physicality" of books, *i.e.* the fact that they consist in bound volumes separated from each other, represents more than a causality of technological progress. Books are not just accidentally physical containers of words that would otherwise flow unrestricted in

---

[54] Immanuel Kant: *Kritik der reinen Vernunft*, A 821-1, B 848-9.

[55] Immanuel Kant: *Anthropologie in pragmatischer Insicht*, A 128-9.

[56] Ibid., A 129

[57] The concept of "originality" in copyright law is rooted in this trait of the work, see Drassinower: "Authorship as Public Address" (2001).

the air; they are coherent acts of meaning addressed to other human beings in order to be tested in their truthfulness.[58]

Second, books as representations of acts of speech are instances addressed to human beings. They convey the likelihood of being appropriated by others. Their very concept is dependent upon a human coalescence, which is nourished by a shared responsibility towards truth. In this respect, the whole of books is not a crowded and confused mass of more or less interrelated self-expressions, but is rather a *sphere*, namely a "public sphere".[59] In this public sphere, authors and readers are, so to speak, "brothers in truth". They share the load of the furtherance of truth, by means of testing reciprocally the correctness of publicly addressed judgments .

What Books Become, or the "Single Liquid Fabric"

Are mass-digitised books still instances of the "public sphere"?

As we have discussed earlier, mass digitisation is described as the technical enterprise that realises the "old dream" of having all past and present knowledge housed in one place. However, at the same time, the conversion of the totality of books into a fully computable whole cause knowledge to appear as a mass of information that overwhelms our capacity to deal with. This is not just an accidental side-effect of the "migration" of all knowledge in the universal form of digital bits. On the contrary, it is its actual purpose. Knowledge *must* appear as overwhelming human intelligence *in order* that the computational power of machines replaces the human capacity of calculus. This replacement brings about a peculiar alteration in the understanding of what our humankind has called "books", "libraries", "public sphere" and "reading".

In this respect, it is interesting to consider how technology itself sees the migration to the universal computability of books. One can appreciate technology's view from the words of one of its spokesmen, Kevin Kelly.[60] From technology's viewpoint – that is: from the perspective of

[58] The right of work's integrity flows from this understanding of the work as internally coherent act of speech.

[59] In Kant's terms the public sphere is the space of the "public use of reason". It is a logical concept, not, as for instance in Habermas, a sociological one.

[60] This is not said in a hyperbolical or, worst, ironical sense. By assuming Kelly as a spokesman of technology, I simply take seriously what he himself claims to be. In his last book, one can read: "In order to respond to technology, we have to figure out what technology wants. [...] Seeing our world through technology's eyes has, for me, illuminated its

the *plus*-trait which informs all instances of our age and compels each of them to turn into fully computable objects – books are just "isolated items, independent from one another", which lie on library's shelves "pretty much unaware of the one next to it."[61] In the eyes of technology, the self-sufficiency of books as autonomous and self-standing acts of speaking appears as *isolation*, namely a state of seclusion to which the slavery of paper tome has so far segregated them. Digitisation is there to free books from loneliness. As a matter of fact, "in the universal library, no book will be an island."[62] This is because "each word in each book is cross-linked, clustered, cited, extracted, indexed, analyzed, annotated, remixed, reassembled and woven deeper into the culture than ever before. In the new world of books, *every bit informs another*; every page reads all the other pages."[63] In the universal computability of everything, no "isolated" bit has a plausible value on its own. The value of a bit lies in the fact of informing other bits, and the resulting cross-informed whole is less a library than it is "a single liquid fabric of interconnected words and ideas".[64] The "public *sphere*" turns into a fully computable flat surface, where algorithms can sweep without restraint from one bit to another.

In this cross-informed universe, the new world of books becomes in fact a single "world's book", namely "one very, very, very large single text: the world's *only* book."[65] Hereafter, no book can subsist outside the "only" book. No book should *dare to* subsist aside from the liquid fabric where every bit informs another. Books as isolated instances, even those that "make sense in their own world", are of "little value" if they are kept outside the only world's book. Books that refuse to liquefy into the single fabric, books that are left outside connections and are not put in condition to "radiate [their] potential connections", will soon be "gasping for air", like a Web page outside the Web.[66] These nearly-floundering books may continue to exist, so long as they manage to do it. However, they are fated to be banned from the public sphere. They will be short of *credibility* – same as living organisms that are deprived from oxygen. They will be just like "pseudo- and parasciences", which are "nothing less, in fact, than

---

larger purpose. And recognizing what it wants has reduced much of my own conflict in deciding where to place myself in its embrace. This book is my report on what technology wants." Kevin Kelly: *What Technology Wants* (New York: Viking 2010), p. 17.

61 Kelly, "Scan This Book!", p. 3.

62 Ibid.

63 Ibid. (emphasis added).

64 Ibid., p. 5.

65 Ibid. (emphasis added).

66 Ibid., pp. 11-12.

small pools of knowledge that are not connected to the large network of science".[67]

This is not to say, however, that technology *wants* to undermine the role of books in our world. On the contrary, by offering them "to wire their texts into the universal library", technology provides books with the very last chance "to retain their waning authority in our culture".[68] Mass digitisation is the last opportunity for books to be authoritative voices in the public sphere. Yet, what is "authoritative" in the context of the "single liquid fabric of interconnected words and ideas"? The units of measure of authority are nothing but *potential connections* that radiate from the single bit. Connections generate accreditation. So, for instance, an "idea" is authoritative not by virtue of its being true, but on the ground that it radiates a number of useful connections. In this respect, automated processing on the single liquid fabric promises, among many other things, to provide "ready supply of authoritative voices".[69]

Yet the most critical shift that comes about with mass digitisation has to do with the peculiar stance of books vis-à-vis truth. And this shift cannot be seen through the eyes of technology. In the interconnected liquid fabric, books are still instances of the common furtherance of truth. However, truth has subtly shifted from correctness to *usefulness* – from correctness of human judgement to usefulness of machine-implemented computation. Together with millions of digitised books, is the task of questioning this shift that our age is about to leave to posterity.

---

67 Ibid., p. 13.
68 Ibid.
69 Schilit and Kolak: "Exploring a Digital Library through Key Ideas", p. 1.