Bernhard Rieder, Department of Media Studies, University of Amsterdam

81.498 Words: the Book as Data Object

-- preprint –

published as: Rieder, Bernhard (2013). 81.498 words. The book as data object. In J. Kircz & A. van der Weel (Eds.), The Unbound Book. Amsterdam: Amsterdam University Press, 57-70.

« Is a digital library a machine or an institution? » (Agre 2003)

1. Introduction

After the entrance – or 'incursion' according to certain commentators – of Amazon.com into the book market in 1995, a second U.S. company, Google Inc., has provoked strong reactions in the world of print in recent years. Already Amazon.com's business model and logistics relied heavily on certain features of networked computing and the company became a poster child for a Web 2.0 imaginary of how to do business online. As Wired's Chris Anderson (2004) famously argued, 'shelf space' had become virtually infinite for online sellers, enabling them to offer a larger catalogue than even the largest brick-and-mortar store. Tim O'Reilly (2005) lauded Amazon.com's strategy to make 'a science of user engagement' and to exploit 'user activity to produce better search results' by encouraging its customers to rate, comment, and browse, registering every purchase and every click, analyzing the data, and reordering rankings and navigational pathways accordingly. According to O'Reilly, it was this fixation on data that gave Amazon.com the critical edge over its competitors. Very much in line with the Web 2.0 catechism, Google Inc. set out in 2004 to carry the book's encounter with algorithms and databases a step further by announcing the plan to digitise *all of the books ever published*. Together with book-oriented social networking platforms like LibraryThing or Goodreads, the two companies have become the main actors in a series of transformations that are set to affect how books are found, read, and understood. Google's ambitious goals and the company's track record when it comes to its self-assigned mission 'to organize information on a global scale in order to make it accessible and useful to all'[1] beg for particular analytical attention and this paper will therefore concentrate on Google Books[2] as perhaps the most pervasive effort to *perform* the book as a data object.

Focusing on a set of specific questions related to books as digital documents, there are many important pieces of the larger puzzle that I will *not* take into account. I will neither retrace the story of the digitisation project led by the company from Mountain View since 2004, nor reference the debates that seem to arise after every announcement of yet another library agreeing to have its collections scanned. Issues concerning copyright and remuneration of authors and editors are equally beyond the scope of this paper, and I will therefore leave the various legal initiatives that

[1] http://www.google.com/about/corporate/company/, accessed October 21, 2011.

[2] http://books.google.com/.

are currently seeking to clarify the relations between the different actors involved uncommented. Finally, I do not want to add to the discussion whether a private American company should have the right to appropriate significant portions of foreign national heritage or not. While these matters are certainly very important, I would instead like to argue that services like Google Books raise certain fundamental questions about the organisation of knowledge in contemporary Western societies that deserve to be discussed in detail. For civilisations organised around writing, changes affecting the various techniques and practices surrounding the written word are bound to have important implications. And although it is very difficult to talk about these changes without slipping into overarching arguments or petty moralising, they deserve to be discussed and thought through.

A starting point for the argument I want to make could be a somewhat naïve question: what kind of *object* is a digital collection containing the full-text of more than 10% of all books ever published,[3] a collection accessed by millions of people each day? Why would a company build such a collection and what is its 'value'? Is such a very large database simply the logical continuation of the library, a version that is perhaps more practical and easier to manage but ultimately comparable to the Library of Congress or other national libraries? Or, on the contrary, are there enough breaks and shifts to justify talking about the emergence of a new sort of object and, potentially, a new configuration of knowledge? To begin to account for these questions, it is useful to look more closely at this latest incarnation of the eternal dream to *capture* all the world's knowledge that is Google Books.

2. The book, the algorithm, and the database

A description of any website is likely to be outdated the moment it is published. My goal is therefore to outline, through a review of its current form, the general ideas that guide Google's book platform. The core functionality consists, of course, in giving users access to the books that have been scanned by the company or provided, already in digital form, by a publisher. Depending on the different legal regulations, access can extend to all the pages, to some pages, to extracts, or to metadata only. Books that are in the public domain can generally be downloaded in PDF format or as a text file created from the images of scanned pages through automatic optical character recognition (OCR). Other features allow users to interact with books in different ways: an identified user can create his or her own 'library' made of 'shelves', which boils down to a relatively basic system of labelled lists; unsurprisingly, there is a powerful search engine that scouts both the metadata and the full text of books. This last feature leads up to the argument I want to put forward here: the processes of scanning and text recognition embed the book into a technological configuration that certainly includes the characteristics of both a catalogue and a reading device. But beyond these familiar elements lurks an industry-leading information system, a database consisting of the full-text content of the processed books complemented by a steadily growing set of supplementary information, which is 'mined' by a variety of algorithms.

This extension points to a similar shift in search techniques that took place on the Web in the late 1990s, namely the rise of the algorithmic full-text search engine, epitomised first by AltaVista and

[3] According to its own affirmations, Google had scanned 15 million books out of an estimated 130 million in October 2010 (http://booksearch.blogspot.com/2010/10/on-future-of-books.html, accessed October 21, 2011).

then by Google Search to name only the most emblematic, and the fading of Web 'catalogues', such as the Yahoo Directory, which were compiled manually by human editors. The different ways of processing contents – manual cataloguing vs. automated crawling and indexing – imply, as a corollary, very different architectures of knowledge, i.e. different ways of ordering and relaying information. Indexing the full textual content of a document opens the door to automated processes that, in a sense, can be understood as ways of *reading* or *interpreting* – 'explaining the meaning of' (OAD) – a book. These processes imply a *particular* understanding of the book as a structured set of words and they 'project' it in certain ways, both individually and as part of a larger corpus. It is therefore not surprising that Google Books offers, beyond the more traditional search functions, representations or 'views' of a title that are produced by algorithms harvesting the textual material. The cloud of words and word combinations that can be found under the 'common terms and phrases' heading on a book's overview page mobilises statistical techniques to find units that are particularly 'significant' for the work in question.[4] Each term is also a hyperlink that leads to a list of passages, taken from the scanned pages, where it can be found.

Fig. 1: The word cloud from Google Books for *The Order of Things* by Michel Foucault

In keeping with the requirement of simplicity that characterises Google's philosophy when it comes to interface design, this feature is rather basic and more understated than the various views and measures that Amazon.com provides in this area. The company is also not hesitant to remove a feature that does not empirically demonstrate its usefulness to users; the 'places mentioned in this book' feature, which generated a geographical map populated by the town names detected in the text, has thus disappeared in the latest redesign.

Google shows its technical prowess when it comes to linking a book with others and to connecting it with additional data sources. The 'popular passages' feature is particularly intriguing: it provides a list of phrases, taken from a selected title, that frequently appear in other books and, copyright permitting, directly link to the relevant pages. In addition, Google lists the references to a title that originate from the Web, from other books, and, via Google Scholar, from the world of scholarly

[4] Note that these measures of relevance make use of the statistical distribution of a word or word combination throughout the entire corpus. Put simply, a word is considered significant when frequent in a certain title and rare in the entire corpus. The quality and size of the corpus therefore influences how well an individual title can be analyzed.

publications. The strategy to invest broadly in harvesting a large variety of data clearly bears its fruits here. Finally, there are both editorial and user reviews, again taken from different data sources or directly submitted through an interface on the site.

All these features give rise to dense networks that connect (almost) any title to a multiplicity of documents and pieces of information. With reference to literary analysis, one could say that Google Books produces different types of *transtextuality*, often taking the material shape of a hyperlink, placing a text 'in an open or hidden relation to other texts' (Genette, 1982, p. 7). The techniques *building* these relations have been developed in at least three distinct directions. First, the system generates a series of what Genette (1987) would call *paratextual* representations (alternative 'tables of contents', generated from the full text: clouds, lists, maps, text statistics, etc.), which aim, on the one side, at 'explaining' a book by providing an overview of the most 'significant' words, the most quoted passages, and so on, and, on the other, at establishing direct navigational entry points into the text. The goal is to *show* us and *direct* us to 'what matters' in the book, as seen through the lens of statistical calculation. Second, different forms of *intertextual* relations between works (direct citations, shared passages, etc.) are scouted for by another set of algorithms and, once again, transformed into both representations and opportunities for navigation. These techniques can go as far as comparing the statistical distributions of words between books and calculating similarity coefficients to indicate 'related' works. Third, Google Books and similar services add a *classificatory* layer that includes basic metadata (author, publication date, etc.) as well as subject classification, which is either based on traditional organisational systems such as the Dewey Decimal Classification or on collaboratively filtered folksonomies established from tags or lists. Algorithmic procedures based on concept extraction or document clustering are not (yet) common but both Google and Amazon.com propose 'readability' scores that situate a book inside of the full corpus by expressing its difficulty in relation to other works. Whether these measures are actually *pertinent* or *useful* is a question that is beyond the scope of this article but they perfectly reveal the eagerness to use calculations to express aspects that we would readily classify as 'cultural'.

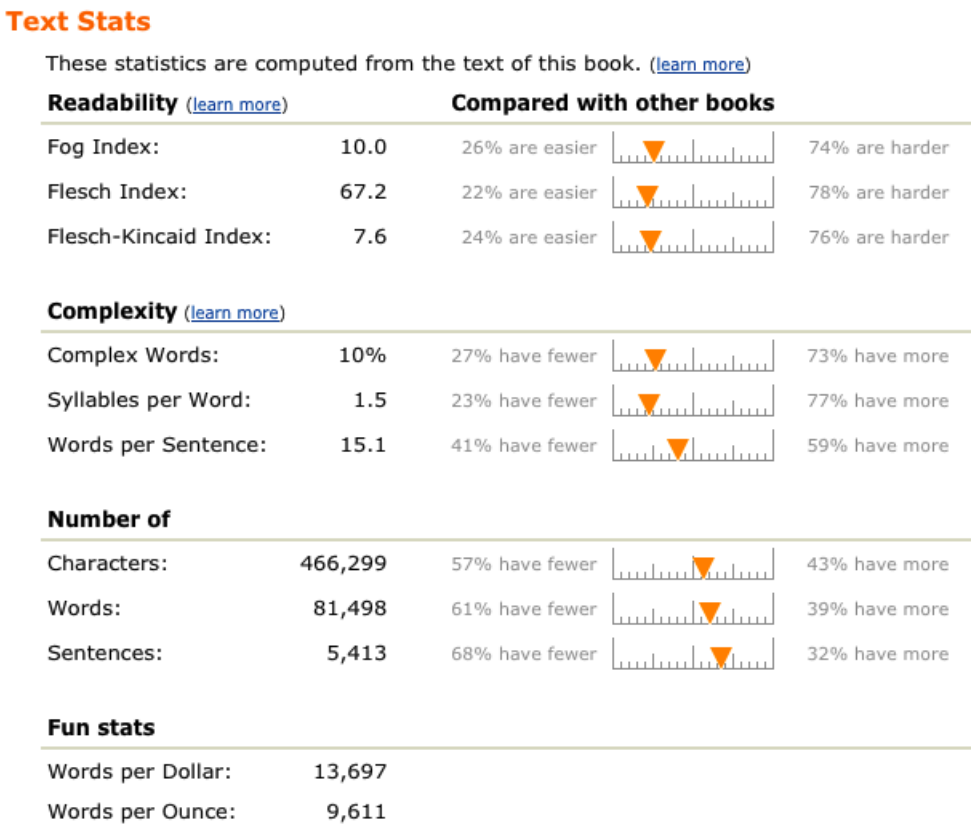

Fig. 2: Amazon.com's text statistics for Victor Hugo's *The Hunchback of Notre Dame*. Readability indices are calculated from the size of a title's vocabulary, the length and complexity of words and sentences, and other factors. The 'fun stats' show that such indices can be created from almost any information, in this case from the book's price and weight, in relation to the number of words it contains.

The transition from printed paper to electronic reader, at least for the moment, leaves the compositional logic of the book as a linear sequence of words, sentences, and chapters largely intact. The full-text book database however is set to treat it as a raw material that can be *deterritorialised* – reduced to textual atoms and their frequencies – and *reterritorialised* – reassembled in groupings of algorithms and interface elements that (re)frame it and make it navigable in different ways. Tables of contents and word indices have provided alternative ways of navigating a book for centuries, but the digital machinery clearly pushes much further in this direction. While electronic readers strongly affect the book as a commodity, online book platforms extract it further from its traditional *configuration* (bookstore, library, paper, reading chair, etc.) and place it into a new set of relations (database, algorithm, screen, interface, etc.).[5] There is an analytical distinction to be made between a cognitive component – the platform as 'vision machine' (Virilio 1988) that introduces layers of automated perception – and 'grammars of action' (Agre 1994) that, embedded in the interface, define modes of interaction and structure navigational pathways. As we have seen above, these two dimensions are deeply intertwined and although we still lack empirical confirmation, it seems clear that the practices of searching, browsing, reading, and understanding will all be affected by this pervasive change of 'device' (*dispositif*).

[5] This increasing 'detachment' of information, first from its physical and then from classificatory constraints, has prompted David Weinberger (2007) to declare that 'everything is miscellaneous', in the sense that a piece of information can 'hang on many branches [of a classificatory system], it can hang on different branches for different people, and it can change branches for the same person if she decides to look at the subject differently' (p. 83).

The question of whether these transformations represent a 'loss' or an 'enrichment' is clearly too full of cultural, political, and aesthetic *a priori*, to treat it seriously here. It seems sensible, however, to remember that the famous expression 'ceci tuera cela', with which the archdeacon in Hugo's *The Hunchback of Notre Dame* (81.498 words!) laments the revolution in both knowledge and power associated with the printing press, is not a law of nature. The history of media has shown time and again that the addition of new devices and formats is rather a story of accumulation, negotiation, and diversification than one of replacement. The shift from an 'Order of the Book' to a digital 'Libroverse' is marked by continuities, discontinuities, and a complicated set of unfolding *consequences* (Van der Weel 2011). But even if the book is not on the verge of being killed off, it is important to note that the logic of database and algorithm lodges it in new configurations, which not only allow for new practices of searching and reading but also imply shifts in collective and cultural forms of circulation and appropriation. It is therefore crucial to consider the economic motivations of a commercial company set on scanning tens of millions of books and on making them accessible in the specific ways I have discussed over the last pages. I will, again, focus on the computational exploitation of data and present it as a complex source of value.

3. Why digitise millions of books?

Despite its 'mission' cited above, it is imperative to treat Google first and foremost as a publicly traded company, with a responsibility to its shareholders and 30.000 employees to make profits. This does not mean that (short-term) commercial goals are the only considerations taken into account by decision-makers. Google's stock structure[6] and considerable cash reserves provide its top executives with significant leeway to plan with the medium and long term in mind. Business considerations are certainly crucial but the company has also shown that it is willing to explore and develop opportunities that do not have a clear or immediate potential to generate revenues. In the case of Google Books, I would argue that there are at least three very obvious sources of (potential) income:

a) On the Web, every page served can be outfitted with ads. In short, the more pages served, the better. A very large collection of books represents, in a sense, valuable advertisement real estate. By providing a comprehensive and useful service, Google can further increase the number of pages – and thus ads – it delivers to users, thereby growing its 'circulation', to use a newspaper term. The company already serves ads with search results on Google Books and also provides links to online shops where users can purchase a title, earning a commission if a title is purchased.

b) Since the opening of the Google eBookstore[7] in the United States in late 2010 it is clear that Google intends to compete with online booksellers, including Amazon.com's Kindle Store and

[6] Like many other companies, Google distinguishes between 'class A' (one vote per share) and 'class B' (ten votes per share) shares. This structure allows the founders and top-level executives to keep full control over the strategic direction of the company while extensively tapping the markets for capital.

[7] http://books.google.com/ebooks.

Apple's iBookstore. A very large number of books available for free can certainly help attract customers who can then be 'guided' to the commercial offer and perhaps be bound to the platform.[8]

c) Google seeks to present itself as – and indeed seeks to become – a one-stop counter for all informational needs and desires. A book service fits perfectly into this logic and promises to further reduce the frequency with which a user has to leave the Google universe to go to other sites. This not only helps with committing users to an ever expanding platform but also to realise significant economies of scale by sharing technological knowhow (modules, algorithms, best practices, etc.) and infrastructure (data centers, data delivery facilities, etc.). We regularly underestimate the immense technological knowledge and engineering skill involved in building large-scale, high-performance Web systems and forget the competitive advantage Google has when it comes to both improving existing services and developing new ones.

In addition to these relatively straightforward elements – advertising, sales, network effects / economies of scale – it seems that the case of Google Books is particularly suited to discuss certain computational aspects whose potential value is a lot less obvious. The two central directions, here, are first the processing of the contents themselves and second the capture and analysis of interactions between users and books.

d) Although attempts to create software that would be 'intelligent' in the same way as human beings have so far proved unsuccessful, even very simple programs can perform functions that could be described as 'cognitive', functions that can provide significant cues in knowledge production processes. As Wendy Chun (2011) phrases it, software can create 'order from order', in the sense that algorithmic processing makes it easy to ascertain frequencies, patterns, and relationships in large amounts of data and to represent these as coefficients, lists, graphs, and other forms. The automatic creation, for example, of a concordance or index for a text – the main uses of computers in early 'digital humanities' from the 1950s onwards (Schreibman et al. 2004) – may seem banal, but the manual work required for such a task can be enormous and indices are very helpful things after all. The functionalities of the Google Books service presented above are mostly based on relatively well-understood statistical techniques (counting words and collocations, calculating probabilities, comparing frequency distributions, etc.) that nonetheless produce results that can be interesting and useful.[9] I would therefore like to propose to see the book database as a latent reservoir of knowledge that is at least partially exploitable by algorithmic approaches. At the moment, we are not yet able to assess fully what Jean-Claude Guédon (2008) called the 'computational potential' of these reservoirs, but it is certainly revealing that the contracts between Google and its partner libraries are generally quite favourable for both the libraries, in particular the European ones, and the end users, while strictly blocking other search engines from access to the digital files. Books have played a central role, over the last centuries, in organising the fundamental conversations a culture has with itself and with other cultures. They document the richness and diversity of human imagination and ingenuity. More than any other medium, books are at the very

---

[8] According to Darnton (2011), this attempt to create a bookstore rather that a pure search tool is one of the reasons why the collective settlement between Google, the Authors Guild, and the Association of American Publishers was not approved by the courts in March 2011.

[9] There are important limits due to polysemy, the difficulty to establish context, and the general complexity and situatedness of language and meaning. My point is, however, that even relatively unsophisticated techniques can deliver very interesting results if applied to large volumes of high quality data.

core of how we understand and invent ourselves. Even if only a fraction of this wealth can be harnessed by algorithmic means, there is enormous potential for all kinds of (commercial) applications.

From an economic standpoint, we can already see a set of areas where large book databases are starting to become valuable assets, beyond the immediate interest of end user services like Google Books. In the field of machine translation, for example, statistical techniques have proven their superiority over approaches based on language modelling, but large amounts of high-quality text data are needed to 'train' the systems. Having access to a large number of book titles in different translations is evidently very useful for keeping an edge in this competitive area – and Google progressively builds translation capabilities into many of its services, most recently into their online office suite Google Docs. Concept extraction and related fields are another area where high volumes of first-rate data are crucial for producing usable results. Already in the late 1950s, the information pioneer Hans-Peter Luhn showed (Luhn 1959) how a basic statistical analysis of word adjacency could reveal conceptual relations inside text documents. The ability to build representations of thematic clusters or term relationships – in the form of semantic networks, thesauri, ontologies, and so on – is certainly not enough to fully attain the level of *meaning* in a non-superficial sense of the word, but it can go far in improving search results and providing a better overview for users. A research project associated with Google Books provides a third, and rather striking, example for illustrating the computational potential of a very large book database. By making their main research tool available to the public on the Google Labs[10] site, the project, going by the telling name 'culturomics',[11] has gained considerable attention. Using a (large) subset of Google Books as their dataset – in all 5.2 million titles in seven languages – the researchers have shown that it is possible to produce perpectives on cultural, political, social, and economic developments in society by analyzing and visualising the frequency variations of 'word grams'[12] over time (Michel et al., 2011). The stated aim is to *measure culture* and certain historical events such as Nazi censorship of Jewish authors do indeed show up very clearly in the data. My goal, here, is not to weigh in on Adorno's famous dictum that, talking about his uneasy collaboration with Paul Lazarsfeld in the late 1930s, 'culture might be precisely that condition that excludes a mentality capable of measuring it', but to point to the fact that serious efforts are underway to *make all this data talk* in one way or another. The particular performativity of the resulting *speech*, whether it will have intellectual merit or create significant economic value, remains largely to be seen. It is clear however that the data pool underlying Google Books is already one of the richest document collections on human history and culture ever compiled.

e) There is a second major element that has to be taken into account when investigating the computational potential of this collection, namely the 'capture' (Agre 1994) of how people actually interact with it, how they search, navigate, browse, and *read*. Computer interfaces can register every query, log every click, measure every timespan, and store every cursory mouse gesture; and, generally, they do. For users, the most visible application of statistical processing of interactions is

---

[10] http://ngrams.googlelabs.com.

[11] http://www.culturomics.org.

[12] A 'gram' can be one ('1-gram') or several words ('n-gram'). The culturomics project fully counts *all* grams from one to five over the 5.2 million books.

in features like spell checkers or the different implementations of suggestion systems[13] in Google's Web search engine and other services, but potential uses can go much further. The data produced by capturing user interactions represent a significant addition to the book contents themselves and they can be used to improve ranking mechanisms and classificatory structures, to evaluate functionality, and to personalise different site elements. More generally, the correlation of user profiles with book contents can produce data that represent 'paths of meaning' – word relations, semantic networks, etc. – derived from actual practice instead of mere text mining. These data 'speak' not only about books, but also about the intimate relationship that individuals and *cultures* – 'webs of significance' according to Geertz (1973) – entertain with these artifacts. The ability to better understand what people *do with books*, what they are looking for and how they read them, may end up being more valuable commercially than the contents themselves. The Google Flu Trends[14] project is a practical example showing how a rather straightforward analysis of user data, in this case search queries, can be used to accurately predict outbreaks of disease. This particular model can surely not be transferred 'as is' to just any other domain but it gives us an idea how real-world data can produce impressive results without any kind of elaborate processing. It is certainly not excessive to expect that the culturomics project's goal to 'measure culture' can easily be extended from books to interactions with books and from a purely historical perspective to the examination of current cultural trends.

The obvious commercial application of captured interactions, beyond improvements to the sites themselves, would be the further enrichment of user profiles for better ad targeting in the various Google services and partner sites. The current pricing system for advertisement, which determines the price an advertiser pays for a user click by means of keyword auctions, has the effect that the company has strong incentives to not only strive to serve *more* advertising but *better* advertising, i.e. advertising that leads to higher conversion rates and thus to a higher sales price for a click. It is hard to overstate the economic importance of user profiling for personalised ad placements, by far the most important source of revenue for Google. Targeting based on incredibly detailed accounts of user habits promises not only to increase the likelihood that a visitor purchases a product on the advertiser's site or to extend the duration of a visit, but also to reduce the level of 'ad resistance' by actually showing fewer but more 'relevant' ads to each user. In short, the high quality of data contained in books can help Google to both learn more about user interests and about the structures of meaning that exist in any culture.

4. A new logistics of reading

Issues of privacy are certainly crucial in this context; but they are also highly visible and I would therefore like to put forward an equally important but less visible question: which type of 'epistemic culture' (Knorr Cetina 1999), which ordering of knowledge, is favoured by the mechanisms that drive a service like Google Books? According to José van Dijk (2010), '[k]nowledge is not simply conveyed to users, but is co-produced by search engines' ranking systems and profiling systems' (p. 575) and we should therefore treat a large-scale book database as a 'knowledge system' (Van der Weel 2011) that implies its own logic. Beyond the concrete

[13] http://www.google.com/support/websearch/bin/answer.py?answer=106230.

[14] http://www.google.org/flutrends/.

techniques, this means that we have to examine the standards and principles that govern the selection and parameterisation mechanisms that decide how to 'show' a book, where to link, and in what order.

I would like to argue that the *problem* that these systems appear to respond to is not only that of a *representation* of knowledge (what is the world?), but also that of the *allocation of limited resources* (how to act in the world?). What book best fits a query? What is 'significant' in a book? What are the 'most closely related' books for a given title? In the end: what to read? The institutions that are traditionally invested in providing answers to these very questions – families, schools, libraries, mass media, churches, etc. – each operate according to their own principles, sometimes in collaboration, sometimes in conflict with one another. What online systems do is to add another logic to the pile, a logic that is somewhat different from the existing ones, however. Automated processes produce strange artifacts that may not seem all too different from taxonomies, dictionaries, reading lists, conceptual networks or catalogues on the outside, but when it comes to looking at how informational units are selected and ranked, we find mechanisms that are much more akin to *pricing mechanisms* in markets than to the ways scholars, educators, librarians or expert committees work. This argument needs to be explained more closely.

In neoclassical economic theory, the market is considered the most efficient form of resource allocation because it provides a quasi-cognitive function in a distributed way: it sets prices via the interplay between supply and demand. The price for an item in a market thus is an emergent property that takes into account the value perceptions of every participant in the system. Search engines and systems like Google Books mostly operate on the basis of similar mechanisms: the choices of Internet users, their navigational paths, their assessments and queries come together in setting the 'value' of each unit of information and just like a price, this value can fluctuate with the ebbs and flows of shifting interests. The result is a level of visibility for each unit of information and a very practical 'cost' expressed by how easily it can be found. 'Cheaper' books, the most popular titles, float on the top while others require a much higher investment of time and skill: one can only reach them at the price of coming up with an advanced query. As the 'right' price, the choice of 'good' books stems from a complex interplay between the book itself and its 'demand' as expressed by the number of citations, references, requests, visits, and so on. By transferring techniques from Web search to the world of books, the services offered by Google are able to provide not only *access* to knowledge, but also an *evaluation* of the knowledge it mediates, an evaluation that is *epistemic* in its own right. The rules by which this market-based ecology of knowledge operates are made operational through algorithms that are, in keep with tradition, unavailable to the public. Its *normative thrust* is therefore only available to critique on a very general level.

The choice of ranking techniques that are similar to market mechanisms mobilises many of the objections put forward by political economists: that the basic assumptions of rational choice and equal access to information defended by neoclassicists are fundamentally unrealistic; that markets therefore are prone to bubbles, disequilibria, inequality, exclusion, and herd behaviour; that powerful actors can largely skew things in their favour. The choice to organise visibility and access through market mechanisms is therefore very much disputable; it is a value choice that implies winners and losers. But are systems of knowledge not always (also) value systems? Certainly. What makes the Google model remarkable is the sheer amount of data it takes into account, its

algorithmic production method, and its radical, market-based *empiricism*. What makes it significant is the fact that it is based on the everyday practices of information search and knowledge production of a very large number of people.

5. Conclusions

The transformation of the book into a digital document is set to have important repercussions. For the publishing world, the transition to digital distribution promises to have implications as large as those already experienced by other sectors of the cultural industries. For the rest of us, the convenience of access, the advanced search features, and the different forms of transtextual navigation provided by services like Google Books has the potential to affect significantly how we find and make use of books. Thanks to its dominant market position and its immense financial and technical capabilities, the company from Mountain View plays a central role in these transformations. On these pages, I have concentrated on the computational aspect of the Google Books project, in particular the question of how new orderings of knowledge result from the coming together – so characteristic of our time – of content, database, algorithm, interface, and the capturing of user interactions and practices.

How do we have to evaluate all these changes around the book? As an epochal disruption? As a change of civilisation? I would propose, with Foucault, that what we are currently witnessing, 'is a day like any other, or rather, it is a day that is never quite like the others' (Foucault 2001, p. 1267). Technological changes, both *causes* and *consequences*, accompany changes in the economic, social, cultural, and political domains and interact in complex and contradictory ways with them. It would be a shame to hide these complex dynamics behind the oversimplified assessment that we are moving from one 'age' to another. But among the forces that make our days not 'quite like the others' is the transformation of our knowledge practices and the emergence, in gestation since at least the 1940s, of knowledge techniques that intimately related, on one side, to a dense network of concepts organised around a very particular understanding of information as a resource and, on the other, to a machine that can give these concepts a mechanical performativity. Google Books has provoked so many strong reactions because the project is placed precisely at the intersection between conflicting conceptions of how knowledge can and *should* be organised and mediated. The digitisation of the book transforms its status as both a commodity and as an object of knowledge and companies like Google and Amazon.com are central actors on both levels.

The database and the algorithm frame *govern* the search for books and the way we read in a way that is fundamentally different from those that drive libraries or school curricula. I would suggest however that we do not yield to the temptation of attributing intrinsic value to either set of mechanisms. Both imply situated configurations of power relations and both produce their own vectors of emancipation and dominance. The question is rather how we can develop the intellectual tools we need to analyze and criticise the computational perspectives offered by the omnipresent interfaces and to develop well-informed practices. How should we debate the values that go into the design of information and how to develop regulatory frameworks that ensure a balance between democratic government and 'algorithmic governance' (Berns and Rouvroy, 2009)?

In this text, I have tried to show how the Google Books service can derive significant, and potentially valuable, informational 'intelligence' from the algorithmic exploration of both a very

large collection of books and a very large database of captured user interactions. The project should therefore not only be seen as a book distribution platform but, more importantly, as a manifestation of a much larger, and somewhat fantastical, attempt to redefine *knowledge* as *information* and thereby make it amenable to the software procedures. This redefinition implies significant changes in modes of access and use as well as a substantial rearrangement of the power structures than run through the anatomy of knowledge. The book as a data object thus sits directly at a cultural fault line that goes much deeper than the question whether we read on paper or on a screen. The intellectual and political challenges are considerable.

6. Bibliography

Agre, Philip: *Information and Institutional Change: The Case of Digital Libraries.* In: Ann P. Bishop, Nancy A. Van House, Barbara P. Buttenfield (eds.): Digital Library Use Social Practice in Design and Evaluation, MIT Press, 2003

Agre, Philip: *Surveillance and Capture. Two Models of Privacy*. Information Society 10(2):101-127 1994

Anderson, Chris: *The Long Tail.* Wired, October 2004 <http://www.wired.com/wired/archive/12.10/tail.html>

Berns, Thomas / Rouvroy, Antoinette: *Le corps statistique.* La Pensée et les Hommes, vol. 53, no. 74, 2009

Darnton, Robert: *Google and the Future of Books.* The New York Review of Books, vol. 56, no. 2, Février 2009

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden: *Quantitative Analysis of Culture Using Millions of Digitized Books.* Science 331 (2011)

Darnton, Robert: *Six Reasons Google Books Failed*. The New York Review of Books Blog, March 2011 <http://www.nybooks.com/blogs/nyrblog/2011/mar/28/six-reasons-google-books-failed/>

Foucault, Michel: *Dits et écrits. Tome 2. 1976-1988*. Paris: Gallimard, 2001

Geertz, Clifford: The Interpretation of Cultures. New York: Basic Books, 1973

Genette, Gérard: *Palimpsestes. La littérature au second degré.* Paris: Seuil 1982

Genette, Gérard: *Seuils*. Paris: Seuil, 1987

Guédon, Jean-Claude: *Who Will Digitize the World's Books?* New York Review of Books <http://www.nybooks.com/articles/archives/2008/aug/14/who-will-digitize-the-worlds-books/>

Knorr Cetina, Karin: *Epistemic Cultures. How the Sciences Make Knowledge* Harvard University Press, 1999

Luhn, Hans-Peter: *Auto-Encoding of Documents for Information Retrieval Systems*. In: Boaz, M.: Modern Trends in Documentation. London: Pergamon Press, 1959, pp. 45-58

O'Reilly, Tim: *What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software.* 2005

Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A Companion to the Digital Humanities*. Blackwell Publishing, 2004

van der Weel, Adriaan: *Explorations in the Libroverse.* In: Grandin, K.: Going digital: Evolutionary and Revolutionary Aspects of Digitization. Nobel symposium 147, Stockholm: Centre for History of Science, 2011, pp. 32-46

van Dijk, José: *Search Engines and the Production of Academic Knowledge*. International Journal of Cultural Studies, vol. 13, no. 6, 2010

Virilio, Paul: *La Machine de vision: essai sur les nouvelles techniques de représentation*. Paris: Galilée, 1988

Weinberger, David: Everything is Miscellaneous. The Power of the New Digital Disorder. New York: Times Books, 2007