The politics of organizing information on the web: computing centres and natural languages

Peter Jakobsson
Södertörn University
peter.jakobsson@sh.se

Fredrik Stiernstedt
Södertörn University
fredrik.stiernstedt@sh.se

Abstract

This paper is an exploration of the methodologies, economics and politics of organizing information on the web, through a historical-comparative analysis of Google. The paper centres on two cases that reveal interesting tensions in contemporary attempts at organizing knowledge and information. The first case deals with natural and artificial languages as tools for knowledge, working with the historical case of Gottfried Wilhelm Leibniz and his interest in a universal language as well as his pioneering contributions to etymology. The second case looks at the dialectics of centralization and decentralization as illustrated by the early 20th-century project of bibliographer Paul Otlet. Together they are used to evaluate Google's utilization of techniques from computer science to extract knowledge from search queries and unstructured web-data, both of which are stored and indexed in Google's computing centres. In the concluding parts of the paper this is considered from the perspective of "audience production". In the case of Google everything users do on the web is potentially of economic value for the company. E-mails, search queries and web-pages are raw materials that can be mined in order to reveal information of our most private longings: the intentions, desires and interiors of human users.

Introduction

The purpose of this paper is to form an understanding of the politics of organizing knowledge and information on the web. Folksonomies and self-organizing systems for classification utilised online are surrounded by narratives of liberation, de-centralization and de-hierarchization.[1] As such these are said to manifest a distinct shift from previous attempts in universal organization and represent a new era in the practice and politics of classification. Outside of engineering circles to little is however known about how these systems actually work. This paper offers a historical-comparative analysis of Google, the main contemporary attempt in exploiting self-organization for a universal ordering of human knowledge, and a powerful force as well as cultural dominant in the information society of the 21st century. Through relating the methodologies and epistemologies of Google with two of the major projects aiming at the systematization of human knowledge in the history of modernity we aim at forming a better understanding of the contemporary contestations over knowledge, language and power on the web. Gottfried Wilhelm Leibniz (1646-1716) and Paul Otlet (1868-1944) – authors/philosophers/inventors/ engineers – are frequently evoked as 'fore-fathers' of the information age. The epistemologies and methodologies they represent are however perhaps too often seen as surpassed by contemporary attempts at organizing information. In this paper we use Otlet and Leibniz as a starting point to analyze Google's methods of organizing human communication on the web. What we attempt to show in this paper is first

how, paradoxically, the decentralized modes of organization employed by Google, relies heavily on centralization. Secondly, we analyze how the techniques used by Google radically extend previous methods of commodification of media audiences by practically including everything on the web in the creation of economic value for the company.

The politics of organizing information

Classification is often met with suspicion. The inescapable arbitrariness of taxonomies always open up for suspicions of hidden power-structures and unsolicited agendas.[2] The ties between bureaucracy, central administration and classification are often evoked as implicit as well as explicit critique of this way of approaching the world. In the light of this there is no wonder that technologies that promises more decentralized and self-organizing ways of structuring knowledge are seen as promising alternatives to taxonomization. Michael Zimmer for example argues that folksonomies "possess the ability to subvert the structured, hierarchical categorization of more traditional physical collections of information".[3] He also claims that such systems makes us less of passive observers for whom information is served and instead, quoting Stephen Werner, makes us "an integral part in the machine's production of narratives of knowledge".[4] Folksonomies, as self-organizing systems, promises to free users from imposed structures of knowledge. As a way of organizing information this has become a visible trend for web-users over the years. Earlier web technologies relied more on formalization and categorisation. Communities, centred round common interests and identities, has been replaced by social networks stressing the multiple and contingent relationships between persons.[5] Centrally administered categories as a way of finding information on the web has been replaced by search tools and ways of tagging material that relies on the users to sort out the relevant links between different kinds of material.[6] Sites serving to the audiences of music, literature, and film relies on the viewing and purchase patterns of their users to give advice to other users on what they might like. Map services expect the users themselves to supply the information on what is valuable and relevant in a certain city or country.

Search engines, such as Google, are also part of this ideology of self-organization and decentralization. There is a potential in search technology that promises more equal opportunities of being heard and consequently also of finding information that is produced outside the system of mass media.[7] Google, for example, dress up their technology in the discourse of democracy by evoking the process of "voting" as analogous to their way of deciding relevance in search results, stating that "democracy on the web works".[8] Google's PageRank algorithm, as well as other search engines, has however been criticized for promoting some kinds of material over other and being inherently biased.[9] Thus the democratic appeal of search has been somewhat tempered. In this paper we want to highlight another aspect of search engines that puts forward the dialectics of self-organizing systems, as they on the one hand promises greater freedom but also greater control.

The paper is organized in a chronological fashion. Although chronology is not part of the main argument of the paper this way of structuring the material works best for readability. The first part of the paper then is about Leibniz and highlights his interest in both artificial and natural languages. These two interests will in the last part of the

paper be put against each other in an analysis of the research fields of Natural Language Processing and Machine Learning as two ways that Google organizes information. This, we will argue, show the appreciation and valuation of the disorganized, everyday communications that are taking place in the globalized information networks. Natural language processing, as used by Google, is a way of both organizing information and commdodifying language. The second part of the paper concerns Otlet and his construction of a world centre of information. This theme also resurfaces in the last part of the paper were we analyze parallels between Otlet's attempts and Google's. In the case of Google, the company's methods of organization rely heavily on the access to storage capacity and most of all computational power. It is the company's huge computing centres that open up for new relationships between knowledge and natural languages, and consequently for a politics of everyday language. Since access to and availability of computing power are increasingly monopolized and centralized by companies such as Google. Together these two themes, natural languages and centralization, form the basis for our conclusion that deals with the commodification of web communication in self-organizing systems of information such as Google's search engine.

It is worth pointing out that the turn to natural languages in organizing knowledge is not new in itself, especially not to researchers within these fields of computer science. The methodologies have been around since the 1950's, and have significantly increased in status since the 1980's.[10] The point we want to make is about the politics, not the technologies, of this kind of organization. Neither is the intertwining of self-organization, decentralization and new methods of commodifying language and audiences new to media- and communications research, but it is rarely substantiated in the way that we try to do in this paper. This paper can be seen as part of the turn towards the hardware and software of informational and cultural production in order to map the workings of power in the networked society.[11] The analysis thus achieved lets us question the oversimplified equation between social organization of knowledge and emancipatory potentials of new technologies.

The universal language

Leibniz has often been pointed out as being one of the forefathers of the information society, paving the way for a "mechanization of processes of thought"[12] and laying the foundations for the modern computer through his binary calculus.[13] The expansion of the field of mathematics into all human fields of knowledge and experience that Leibniz promotes has been interpreted, by for example Armand Mattelart, as the very intellectual basis for the information society.[14] One of the primary reasons why Leibniz should represent such a *mathesis* of modernity is arguably his interest in constructing a universal language: a life-long project that at least at some times involved a substitution of alphabetic expression for numerical, with the expressed hope of escaping the arbitrariness of words and minimizing the space for interpretation and uncertainty.

De Arte Combinatiora is Leibniz's first serious development of this project, but it is elaborated in many of his other writings.[15] The direction that his thinking on this matter took, as reflected in these writings, shifted somewhat during his life. A basic premise of his thoughts on a universal language was however that all concepts are in reality only combinations of other more simple and fundamental ones, except for the

most fundamental which are the basic building blocks of all other concepts. Furthermore, he believed that if it was possible through philosophical analysis to arrive at these most fundamental concepts, and if they were fitted with appropriate symbols, they would spell out the alphabet of human thought, a *universal characteristic*. With a logical system of symbols and rules for calculation, the human thought could then be subjected to the same kind of reasoning used in arithmetic. The power of analysis and rational thought evident in this project is perhaps why Foucault puts Leibniz as "the gravitational centre" of classical thought.[16]

Leibniz was not in any way unique in his time in trying to construct a universal language.[17] The decline of Latin was definitely a contributing factor for this, but otherwise the motives diverged. Some of the projects were motivated by the promise of increased international trade, whereas others saw the construction of universal languages as a way to achieve peace and order; the universal language would serve humanity as a whole.[18] Leibniz represented the latter stance as he thought that a completely rational language would make conflicts unnecessary. Famously, the goal of his project was that two men would solve any conflict by sitting down, and say to each other: "'Let us calculate, Sir', and thus by taking to pen and ink, we should soon settle the question."[19]

The quest for a universal language should also be seen against the background of a common conception of the 17th century, namely that natural languages constituted a barrier for the mind, which needed to be overcome or razed in order to achieve true knowledge. On this point there has been many rivaling interpretations of Leibniz's thought. Some have argued that he sought to replace natural languages and others that his universal language meant as complimentary to vernacular languages.[20] More important for us however is the question of *why* human languages did present such a problem to rational thought? This problematic of natural languages is with us to this very day and the consequences of how this question is tackled are central in what follows.

The first answer to this question has to do with the ambiguity of words. An illustration of this problem is found in *New essays on human understanding*, a reply to Locke written by Leibniz.[21] In this fictitious dialogue Philalethes (Locke) emphasizes the richness and ambiguity of natural languages and asserts that any dictionary of human thought has to be given by way of examples. Theophilus (Leibniz) on the other hand is firm in the belief that the signification of words can be reduced to a "determinate number of significations" and defined by "substitutable paraphrases".[22] For Leibniz, a language that would be useful in discovering new truths through calculation required such clear definitions: hence the need to construct a universal characteristic. Another solution to the problem of ambiguity would however be to construct a system that collects and indexes examples of uses of words in an automatic manner, which is the solution that we will look at in the case of Google.

The second answer as to why human languages are problematic for the 17th century philosopher has to do with the development of these languages:

> The situation is such that [specifically human] needs have forced us to abandon the natural order of ideas, for that order would be common to angels and men and to intelligences in general. It would be the one for us to follow if we had no concern for our own interests. However, we have had to hold fast

> to the order which was provided by the incidents and accidents to which our species is subject; this order represents the history of our discoveries, as it were, rather than the origin of notions. [23]

The historicity and seeming contingency of language is thus a second problematic of using human language as a tool for knowledge. This Leibniz's learned from his interest in etymology.[24] His view on natural languages, as evident in his *New Essays on the human understanding*, was that that they are not totally conventional in their signification. Human languages are the result of their history, but equally important, they are the result of the meeting between the world and the human mind. Natural languages are the result of our sense of being in the world: "[L]anguages have a certain natural source, namely the harmony between sounds and affections which the sight of things excites in the mind."[25] Onomatopoeia is for example a case in point for Leibniz:

> The latin *coaxare*, applied to frogs, corresponds to the German *couaquen* or *quaken*. It would seem that the noise these animals make is the primordial root of other words in the Germanic language. Since these animals make a great deal of noise, we connect it with chatters and babblers, who we call by the diminuitive *quakler* […] And since those sounds or noises of animals testify to the presence of life, and tell us that something living is there before we can see it, in old German *quek* signified life or living.[26]

In other writings he went beyond onomatopoeia to search for such relationship. For example, 'nose' fittingly enough begins with an *n* since this letter is pronounced through the nose.[27] Realizing however that such relations were not to be found in all, or even the majority, of the words used in the languages of his own time, he posited that languages change for numerous reasons, often by pure chance: "words have passed by means of metaphors, synecdoche, and metonymies from one signification to another, without our always being able to follow the trail".[28] Establishing the origin and development of languages required extensive historical research, but could in theory be accomplished.

Besides the construction of an artificial language then, another possible way towards knowledge were empirical studies of the spread of words and grammar; in short, etymology. The knowledge that could be gained from such studies were however somewhat different from the universal knowledge that could be obtained through the universal characteristic. In his *Ermahnung an die Teutsche ihren Verstand und Sprache besser zu üben* he lays down two principles about the German language that he according to Aarsleff soon would consider to be true of human languages in general: First, "The bond of language, of social customs, and even of the common name unites individuals in a powerful though invisible manner and produces as it were a sort of affinity." and secondly, "Language is to be regarded as a bright mirror of the understanding."[29] The study of human language could thus teach us about the origins of nations and communities of people through the way that language reflects the experiences and histories of certain people. And secondly, language in the way that it mirrors the mind, could learn us more about the workings of the intellect and its relationship to the world.

Leibniz interest in natural languages, as distinct but also intertwined with his interest in artificial languages, has been noted upon by scholars such as Rutherford and Aarslef but does not seem to have evoked the same amount of attention as other

aspects of his philosophy.[30] Leibniz's ideal was probably the philosophical language, but he nevertheless held that since we are beings of both sense and reason we have no option but to hold on to our natural languages as they have developed, by chance or by reasons given from our motivated involvement with the world.[31] This line of thought in Leibniz, quite different from the path towards "the mechanization of thought" hailed by Wiener, could perhaps be used to put forward a more sympathetic image of Leibniz – in a time less convinced by overzealous rationalism. That is however not the point to we want to make. Instead we want to point to the fact that the problematic of natural languages as discerned by philosophers in the time of Leibniz is very much still with us. The empirical approach to meaning in language, overlooked by those who points to Leibniz as a fore-father of the information age, is perhaps the most salient approach to organizing information on the web today. This does not however imply an escape from arbitrary classification and democratic ideals of knowledge, but has totally other consequences, as will be explored in the case of Google.

The universal book

Themes of universal categorization and universal knowledge from the early stages of European modernity are revived again during the "technological modernism" of the 20th century.[32] Industrialism cultivated an ideological emphasis on efficiency through improved design and engineering and the large-scale corporate capitalism needed systems for control and standardization. The scientific management of labor – Taylorism – required a "constant flow of data from the production and marketing processes upon which management could base its decisions".[33] Hence a new regime of information.[34] One of the front figures of this time, often pointed out as a forefather of the information society is the Belgian documentalist Paul Otlet.[35]

Otlet was from early on drawn to positivism and its scientific method, its rejection for metaphysics and its utilitarian ethic of good for Humanity. All of these themes are recurrent in the life and work of Otlet. But another feature of positivism was certainly of equal influence on the career path he chose: the firm belief, within positivism, in the possibility as well as the necessity of synthesis. For Auguste Comte, for example, the "positive generalities" would be able to organize all of human reality and would gradually lead to a kind of unity in science.[36] As suggested by Bernt Frohmann Otlet's project rested upon the insight of *language* as being subjective and ambivalent: and in consequence threatening meaning and truth.[37] Therefore Otlet, as Leibniz before him, sought to create and implement an artificial language that would capture only the objective *facts* within a document. Classification becomes for him not primarily a question of grouping together texts written within the same discipline, or treating the same subject: bibliographical classification with Otlet leaves the surfaces of documents in order to use their inner workings, the actual contents (the facts) as units for classification. In this way he sought to increase simplicity and decrease instability and variability in texts: "to reduce all that is complex to its elements" and to secure "unvarying meaning".[38]

Otlet found a system that could complete such a task in the American Dewey decimal classification, which he modified to become the Universal Decimal Classification (UDC). It was constructed in such a way that it had a greater flexibility than previous systems for bibliographical description. It could easily encompass new knowledge by

its infinite but orderly extensibility. It worked as a code where each number (between 0 and 9) was assigned a quality that was "identical in all the combinations of which it is made part".[39] Through divisions, in classes, groups, divisions and sub-divisions, it would hence be possible to express utterly complex themes and contents in series of numbers.

The aspiration of Otlet's system goes beyond descriptive bibliography and classification. Otlet wants to do away with the 'languageness' of language and solve the 'problem' of ambiguity and invent a system of signification that would be as unambiguous and lucid as the facts within a document. This was possible only, he reckoned, if reading, and eventually the practice of writing itself, was 'freed' from the human author. The ambition was that the introduction of the bibliographical system would mean an abandonment of reading as an act to "slavishly follow the author through the maze of a personal plan, he attempts to impose on those who read him".[40] This gathers the paradox in Otlet's work, as in modern thought in general, on the one hand subjectivity has to give in for the objective, the fact: qualities has to be transformed to quantities, numbers, the human has to disappear. But all this in the name of the very same human, the freedom for the individual (reader) to not any more be forced to follow the plan so meticulously outlined by the author. Or to put it in another way: what Otlet wants to establish is the identity of the document instead of that of the author, an autonomy of information.

Besides Otlet's emancipatory ideals there is one other part of his project that interests us in this paper, the creation of the Mundaneum - the construction of a world centre in order to fulfil an organisation of all the world's information.[41] As shown by Charles van den Heuvel Otlet himself often used the architectural metaphor of the factory to visualize and explain this world centre.[42] In an actual drawing, titled 'Laboratorium Mundaneum', Otlet illustrates his project by mountains of raw material – information – tapped into a factory building with smoking chimneys and transformed from the disorderly mass of 'journaux, revues, lois, livres, brevets, statistiques, correspondence'[43] to the end product of properly packed and categorised rolling out from the factory on a train of box cars (driven by a locomotive named UDC).

It was, however, not only in drawings and on a metaphorical level that the Mundaneum was conceptualised as a factory. It was also highly dependent on infrastructures and technologies developed through industrialist mass-production and modes of organising labour. The database, swelling over time and eventually containing about 16 million entries needed a storage facility of 150 rooms. The problem of space and organisation of space was hence a very real and tangible issue for Otlet and his collaborators. In 1910, Otlet and La Fontaine first envisioned a "city of knowledge", which Otlet originally named the Palais Mondial (World Palace), that would serve as a central repository for the world's information. In 1919, soon after the end of World War I, they convinced the government of Belgium to give them the space and funding for this project, arguing that it would help Belgium bolster its bid to house the League of Nations headquarters. They were given space in the left wing of the Palais du Cinquantenaire, a government building in Brussels. The Palais Mondial was briefly shuttered in 1922, due to lack of support from the government of Prime Minister Georges Theunis, but was reopened after lobbying from Otlet and La Fontaine. Otlet renamed the Palais Mondial to the Mundaneum in 1924.[44] The Mundaneum, during its time of operation between 1919 and 1934, out of necessity

adapted to methods of the contemporary factory, with "division of labour, centralization and standardization", as Otlet himself put it in a speech at the *Second International Conference of Bibliography* held in Brussels in 1897.[45] Furthermore, the project itself had been unthinkable were it not for the contemporary inventions in information technologies, as for example the abandonment of previous ledger-based systems for card-based systems within bibliography. Through these the databases could increase their flexibility as cards could be moved around, edited and re-arranged. As described by Rayward this development took its inspiration from the contemporary Taylorist and Fordist ideas of "flexibility, correctability, currency, cumulativeness, and cooperative formation, maintenance and use".[46] The technology of card-in-cabinet and shelf systems for storage was a technology that was invented and got a wide diffusion with the rise of the industrialism and large-scale state bureaucracies of Otlet's time and it also implicated another technology: the cards themselves. The standard in the Otletian archive was the 3x5 American catalogue card and the project of his not only rest upon a standardisation of bibliographical cards, but also a possibility to mass-produce such cards.[47]

This mode of organizing the Mundaneum expresses a dialectics between centralization and de-centralization in Otlet's projects that we will argue resurfaces in the case of Google. Because at the same time as the Mundaneum was sought to be a "world centre" the UDC was a highly decentralised mode of organising knowledge. The numerical system was flexible and only worked if the input to the system was made by a multitude of different actors. And as previously mentioned, at the very heart of card-based systems is their allowance for change and the possibility to alter the relations between elements in the system without it affecting the system as such. The distribution of the UDC and its universality was a vehicle to create such a decentralised mode of information management. As we will see, these themes, combined with the Leibnizian interest in formal- as well as natural language processing are the basis for Google and its contemporary operations.

The universal medium

If we today feel that Leibniz's and Otlet's quests for universal knowledge are a bit pretentious or even embarrassing, Internet giants such as Google are met with no such feelings. On the one hand, libraries and archives of all sorts have opened the door to their collections for the Google scanning teams, and on the other, an increasing amount of users regularly turn to them for information on all things under the sun. The dream of ever-accessible and universally useful knowledge has been transformed into a form that we seem to accept as viable and worth striving for. Cybertheorists of the 90's seem to have paved the way for an acceptance of the business models of today.[48]

Since its inception the web has grown exponentially, calling for solutions to problems of organization and navigation. Google is famous for its founders Larry Page and Sergey Brin's patented PageRank Algorithm, a solution to search that uses 'information within information' to achieve the task of organization.[49] This 'information within information' is so called backlinks. Google's search engine calculates the relevance of search results by acknowledging the fact that sites with more incoming links are likely to be more relevant to most web-users. This method of organization consequently lets the users of the web play a part in deciding what kind

of search results will show up in Google's hit lists; if many web-users link to a particular site, this site will be displayed at the top of the hit list. This method of organization is present in other Google products as well. Google Earth, for example, relies on the user's to tag maps with relevant information. Even Google's advertising system relies on this method: Ads that are clicked on often will also show up more often on site's that are powered by Google's ad-system.

PageRank is important for Google, but is not the only source of the company's success. In this paper we would like to dig a little deeper into Google's various strategies of organizing knowledge. In order to achieve this purpose we will need to go beyond sweeping comparisons – between centralization and decentralization; top-down and bottom-up; inductive and deductive methodologies – and engage the epistemologies and methodologies of the contemporary approaches to organizing knowledge in the same detailed way as we can appreciate Leibniz's and Otlet's attempts at organization. In the case of Google this can however be problematic since, like other web-companies, its methods and technologies are in most instances well-guarded secrets. We have tried one way of getting past this obstruction in this paper. On Google's website there are links to papers and conference publications that has been written by the company's employees. Such papers give some hints of what kind of expertise Google is hiring and also says something of what kind of research is rewarded within the company. Indirectly then they can be used to analyze the firm's operations and even the different strategies and epistemologies behind knowledge organization on the web as a whole.

On the site "Papers written by Googlers" a huge number of different publications is listed and it has not been possible to go through all of them. But as a starting point their classification can give a clue to what kind of research is undertaken by the company's employees. There are 101 articles, books and conference presentation under the non-descript title of *Algorithms and theory*. One also finds 88 publications under *Distributed Systems and Parallel Computing,* 84 under *Machine Learning,* 65 under *Natural Language Processing,* 51 under *Audio, Video, and Image Processing,* 49 under *Security, Cryptography, and Privacy,* 48 under *Human-Computer Interaction and Visualization* and 47 publications under *Artificial Intelligence and Data Mining*. Besides these there are also several categories with considerably fewer publications.

Two of the fields and directions covered in the published material will be covered in this paper: Google's computing centres (*Distributed Systems and Parallel Computing)* and the second and third largest categories: Machine Learning (ML) and Natural Language Processing (NLP). It is here we can find the relatively new direction for the quest of universal classification and organization, and interestingly also, the basis for the ideology of decentralization and self-organization that runs through so many web-related discourses. The epistemology and methodologies of Natural Language Processing, and its interest in human languages in opposition to artificial languages, will be the main object of comparison in relation Leibniz and Otlet.

ML and NLP are fields within computer science and both off-shoots or subfields to research on Artificial Intelligence. They are, if you like, what is left of the quest for artificial intelligence since the more grand claims of that field has stranded. ML is a

more general term and its methods can be used for NLP. Since it is these uses that interest us here we will in the following only use the latter term. NLP concerns the interactions and communications between computers and humans using human languages, both written and spoken. A basic research question for this field would, for example, be to get a computer to understand the intentions of a web-user from his or her use of human languages. We will argue that it is this mixing of human intentions, desires and interiors with science and technology, at the most basic level of NLP, which makes the field interesting for corporations such as Google.

The field has a history as long as the history of the modern computer. Throughout this history there has been two paradigms competing for the hegemony of the field, the "rationalist" and the "empiricist" paradigm.[50] Pragmatic concerns have however made the two paradigms come to terms with each other and share the scientific burden of the field.[51] However, during the last twenty years, technological developments, and as we will see, changes in the modes and directions of social communication has brought with them changes in the field. This change is in the following quote described by one of Google's employees:

> The application of statistical methods to natural language processing has been remarkably successful over the past two decades. The wide availability of text and speech corpora has played a critical role in their success since […] these methods heavily rely on data. Many of the components of complex natural language processing systems […] are statistical models derived from large data sets using modern learning techniques.[52]

The availability of corpora is thus crucial to the techniques of NLP, and the availability of these has increased since more and more collections of texts are digitized. The basic problem that availability of corpora in conjunction with statistical methods can help solving is the ambiguity of meaning of human languages; the constant deferral of meaning that haunts formal systems. A problem that in this paper was previously illustrated by the dialogue between Leibniz and Locke. In that era the book, as a static and bounded artifact, was clearly limited in providing definitions through examples, since any list of examples can never be exhaustive. In a dynamic system, such as constantly upgraded web-indexes, that can automatically compare any given statement with a large corpus of naturally occurring statements this disadvantage is overturned. The possibility of using, for example, the whole World Wide Web as a corpus, has contributed to the appreciation of the strength of a statistical approach to language.

Most researchers within NLP rely on more limited corpuses, for example newspaper archives or medical databases; a limitation that is guided by research interests but also by practical (i.e. financial) reasons. A company like Google can hypothetically use not only the web as a whole, but also e-mail correspondences, and all the other archives that they have scanned and indexed into their databases. The supply of enormous sets of data is obvious in the papers written by Googlers, facilitating 'experimental settings' containing, for example, "*50 million unique, fully anonymized search queries* in English submitted by web users to the Google search engine in 2006".[53] Or in another case the "experiments relied on the unstructured text available within a collection of approximately *100 million web documents* in English".[54]

These quantities of data bears witness of an aspect of Web 2.0 that is rarely noticed or taken into accounts of search engines: the fact that it requires massive computational power to operate. Despite the rhetoric of abundance and the post-fordist economy (implying its independence from industrial materialities), Web 2.0 is not at all a weightless, digital non-space: it is not only sets of algorithms allowing decentralized users to self-organise in networks independent of material concerns. It is, for example estimated that Google alone needs about 1 000 000 servers to keep the search engine going. And as the amount of data archived, stored and managed by Google is increasing (and it constantly is), the problems of how to keep the data pushes Google to, in an increasing pace, build new data centres all over the world, investing more and more of its resources in new servers, more computational power.[55] But not only does Google have to create new data-centres, the constant lack of space also pushes Google engineers to develop more advanced system-architecture and design more energy-efficient servers farms.[56] This is obvious in the research articles that Google publish online. Even though this part of Google is the one least public and the one the company try to keep most secret, a fair amount of the research papers deals with questions of energy-efficiency, power-provisioning, system architecture and 'failure trends in large disk drive populations'.

Google has become famous for its so-called "distributed computing".[57] The basic idea is to aggregate computational power through networking a multitude of standard, off-the-shelf servers and hard drives from consumer brands. Instead of trying to build one "super-computer", every single server does its assigned tasks, networked to the other computers of Google, to form something of a digital assembly line where everyone fulfils their small part of a grander whole. The information is distributed over many computers, and the answer to every search query involves a mass of servers, delivering their parts of stored information pathways. The copies of material published online that Google keeps in their server-farms is stored in three copies distributed again over a range of hardware. Even the indexes of archived information are atomized in this way, stored on many different machines.[58] This organisation, in which every machine is assigned specific tasks also makes every server replaceable and it is possible, as done at Google to count on hardware failures: failure is not an exception, as stated in one of the research-papers published on Google's website, 'failure is the norm'.[59] Decentralization of hardware is however not as straightforward as it seems. The system architecture of Google, as described in the Google-papers 'Web Search for a Planet' and 'The Google File System' instead paints another picture. In order to simplify the system architecture the file system of Google, the Google "archive", has adopted a hierarchical structure with "one single master" to control the labouring servers in every network: an informational supervisor to control all movement within the network.[60]

In this way the factories of Google resembles the description Otlet gave of the principles for the Mundaneum, that should simultaneously rest upon division of labour and centralization. The Taylorist idea of scientific management of labour is hence perfected as the labourers – to a higher degree than ever before – is made up of machines and the raw-materials are immaterial to a higher degree than before.[61] Furthermore, the server-farms rest upon the same industrial limitations and conditions as any manufacturing in industrial society: supply of energy, water and space. One of the central questions for Google, as elaborated in the research-paper 'The Price of Performance' is how to "afford the computational capacity you need".[62]

As previously stated NLP is aimed at facilitating communication between humans and machines by making machines 'smarter', which is also the reason for Google's interest in NLP. Larry Page's dream is to develop "the ultimate search engine" that "understands" what the user means and what he or she wants.[63] In this task Google seem to face the same difficulties as once did Paul Otlet. The ever-increasing mass of "mountains of information" has to be categorized in order to become searchable. Not only do Google manage this problem through "help[ing] people create structure that aids search"[64] through the Google Base-system, nor does it only rely on exploring the already structured data of the deep web or annotation schemes such as *Flickr*, that manage information through exploiting users own labeling of content. Google also needs to process unstructured text, the mess and chaos of the online worlds, in order to make it searchable and manageable. The main promise of NLP is in this respect that it can help construct large knowledge bases from seemingly unstructured corpuses of texts – since it can help the machines to understand in what category to place a given piece of information, with only minimal 'external constraint', that is to say manually specified taxonomies.[65] One such interesting way of minimizing the regulation of constraints in the generation of labels, or classes, and still achieve the wanted-for disambiguation of search results, is to combine Google with Wikipedia. As "instances of the same class (e.g. different people) or different classes (e.g. a type of snake, a programming language or a movie) may share the same name in the query"[66] the effectiveness of web search could be greatly improved if search results could be grouped together according to "the corresponding sense"[67] rather than being presented in a "flat sense-mixed list of items". For example the name John Williams has a multitude of meanings and a great many people share this name. In order for a search engine to deliver relevant results it has to understand which of the John Williams that are asked for, and in order to deliver the correct results it has to be able to group the different Williams: the musician, the wrestler et cetera. And in order to make such groupings – without documents that have been pre-processed, classified or tagged – it has to be able to 'understand' what the document is about. The search engine has to learn what kind of different Williams there are and their respective qualities and as argued in the Google research-paper, this could be done through using the Wikipedia-dataset (of 1,783,868 queries) to train a "named entity disambiguator".[68] In this example it is also all too obvious how the mining of categories and classifications in large text corpora ultimately exploit the work of users who create a seemingly unordered mass of information in their handling of digital media. Other similar examples of how Google creates classes and classification systems through combining NLP with statistical models are their interest in implicit relations; the billions of relations between named entities on the web.[69] As stated by Culotta et al

> In order for relation extraction systems to obtain human-level performance, they must be able to incorporate relational patterns inherent in the data (for example that one's sister is likely one's mother's daughter or that children are likely to attend the same college as their parents).[70]

Hence, in order for classification to be successful the machine(s) has to learn how human relations work and how humans classify their everyday life. The search engine has to "capture human knowledge"[71] in order to fulfill its task of an organization of the unstructured data of the web.

Echoing Paul Otlet and his notion of the facts within documents as the object of bibliographical classification, Google-engineer Marius Pasca in his research-papers predicts a coming World Wide Web of Facts

> Although the information in large textual collections such as the Web is available in the form of individual textual documents, the human knowledge encoded within the documents can be seen as a hidden, implicit Web of classes of objects (e.g., named entities), interconnected by relations applying to those objects (e.g., facts). The acquisition of an extensive World Wide Web of facts from textual documents is an effort to improve Web search that also fits into the far-reaching goal of automatically constructing knowledge bases from unstructured text.[72]

For him this is possible through introducing data-mining not only of documents (web-pages, e-mails, books et cetera) but also of search queries as such: the search queries which are the most prevalent and direct input from the human users in to the system of organizing knowledge. This is what Pasca calls the "wisdom of the (search) crowds" to which millions of web users contribute daily.[73] The movement from pre-specified semantic relations between documents and facts, towards the relations created and revealed by the "real-world interest[s]"[74] expressed in search queries, for Pasca means to get a grip of not only human knowledge, but the very process of knowledge creation: to get direct access to the cognitive as well as emotional dimensions of 'real-world' web users. The technology of Google in such ways translates our vernacular into computer code, and back again – making it possible to systematize and hence anticipate not only the answers we are looking for but also the questions we ask: Google aims to construct a search engine that not only is good at taking commands but that with the help of the sheer mass of information lets Google understand its users from within and make their desires visible and transparent. Informational culture ultimately turns not only everything but *everyone* into information

Conclusion

Over the course of this paper we have assembled a history of the information society using Leibniz and Otlet to highlight the dynamics of the contemporary attempts to organize all the world's information. As our narrative suggests the organization of knowledge in informational capitalism/culture can be understood as an amalgamation of themes from the pre-industrial Leibnizian visions as well as the industrial and modernistic bibliography of Paul Otlet. Google, as an organizer of information combines utilization of synthetic languages with development of natural language processing. Google also utilizes an industrial and large-scale handling of the information-commodity and (re)produces ideological narratives of public good, democratization and increased rationality. But rather than being a more democratic way of organizing information we argue that the turn towards the web-users as the source of building taxonomies and classification is more adequately assessed if seen from a traditional political-economic perspective; as novel ways of commodifying language and the web-audience.

Others have noted that search means new ways of 'producing' an audience and in this helped to create a "crisis in the ratings industry" as a whole.[75] Hence, in radio as well as in television new modes of audience measurement have been introduced, methods that mimic the 'clickability' of online services.[76] The newness of the search

economy's way of producing an audience is however still under debate. Fernando Bermejo argues that on the web, what is actually sold are words; it is search terms that are the new commodity that search businesses sell to advertisers. Companies buy specific search strings that will trigger the appearance of their ads. The ads are only paid for once they are clicked on; a system that has the consequence that different search strings will be differently valued because of a differing market interest. For example "globalization textbooks" costs 0.05 EUR, while "Cultural studies graduate programs" costs 3.54 EUR.[77] The political economy of Google is then different because the audience no longer "works" for media companies through watching, reading or listening to commercial messages.[78] What is collected, and hence what is sold, is not watching, listening or reading: but typing and clicking. Instead of receiving, what is sold is activity, the doings of the web.[79]

What is missing in this picture however is what we have tried to show in this paper: the intimate relation between methods of organization of information and the production of audiences, or clicks. In an article published on Google website, explaining the uniqueness of the Google file system, the authors hint to the importance of large data sets:

> The file system [. .] is widely deployed within Google as the storage platform for the generation and processing of data used by our service as well as *research and development efforts that require large data sets.*[80]

This research and development arguably include the 'research' done on the 'audience' or the users of Google. A research that is highly automated, and which generates revenues for the company. Hence, the same methods that are used to serve users with good search results are used to serve advertisers with clicks. Google are, in both of these cases, relying on not only our search strings, but on all information possibly gathered from the web and our Internet communications. For example, only by understanding the content of an e-mail can the most relevant ad be placed there, and getting a computer to understand the semantics of an e-mail requires NLP or similar techniques. The explicit goal of Google, to produce a search engine that understands the user's wishes, is in for example G-mail complemented with the dream of understanding the user through his or her texts, the e-mails she sends and receive. And through the mapping of implicit relations between "named entities" of online documents the company seeks to understand the inner workings of human users. To informatise knowledge, longings, passions, wishes, dreams, ideas, emotions and practices that are never explicitly put into words in the form of search queries.

This means that far from only relying on clicks, Google takes advantage of, at least potentially, all our communications over the Internet. In order to get a better understanding of us as consumers, everything we do on the Internet becomes potential audience labor. This is the general epistemological promise of audience production in new and digital media – and what has produced a crisis of measurement – to informatise also the human interiors: to provide direct address to the user's souls. The abandonment of arbitrary, top-down systems of classification does consequently have wider consequences than liberating us from these very systems. They do indeed make us integral parts of the machines of knowledge production, but not only as subjects but rather as *objects* of classification. The successful selling of the idea of collaborative and collective intelligence as the force of organisation and knowledge production enhances a very effective informationalisation of the world. One in which

we begin to move beyond an index of knowledge to an index of everything.[81] And most importantly we move towards an informationalisation that includes not only all types of ”documents” in an Otletian sense, but everything from our most private longings to our most public concerns; the mapping of our bodies, interests and personalities in this way creates a global database of dividuals[82], subjects of informational culture.

[1] Introna and Nissenbaum, “Shaping the web: why the politics of search engines matters.”
[2] Bowker and Star, *Sorting things out* .
[3] Zimmer, “Renvois of the past, present and future,” 110.
[4] Werner, “Using the Encyclopédie ,” 270.
[5] Bassett, *The arc and the machine* .
[6] The first approach was successfully employed by Yahoo, but later complemented with search capabilities.
[7] Lev-on, “The democratizing effects of search engine use: on chance exposures and organizational hubs.”
[8] Google, “Corporate Information - Technology Overview”; Google, “Corporate Information - Our Philosophy.”
[9] Introna and Nissenbaum, “Shaping the web: why the politics of search engines matters”; Diaz, “Through the Google Goggles: Sociopolitical bias in search engine design”; Goldman, “Search engine bias and the demise of search engine utopianism.”
[10] Dale, Moisl, and Somers, *Handbook of natural language processing*.
[11] Fuller, *Media ecologies* ; Fuller, *Software studies* ; Galloway, *Protocol* ; Kittler and Johnston, *Literature, media, information systems* .
[12] Wiener, *Cybernetics or control and communication in the animal and the machine*, 20
[13] e.g. Davis, *The universal computer* .
[14] Mattelart, *The information society*
[15] Leibniz, “Dissertation on the art of combinations, 1666 (Selections)”; Leibniz, “Preface to a Universal Characteristic (1678-79)”; Leibniz, “Towards a heuristics for discovery”; Leibniz, “Introduction to a secret encyclopedia”; Leibniz, “Two prefaces to the general science.”
[16] Foucault, *The order of things* , 63
[17] . Bacon, Dalgarno and Wilkins are some of the most well known representatives of this movement.
[18] Pombo, *Leibniz and the problem of a universal language*.
[19] Leibniz, *Leibniz: Selections*, 15.
[20] Cassirer, *The philosophy of symbolic forms*, is most famously associated with the substitution hypothesis. Pombo, *Leibniz and the problem of a universal language,* holds that the search for the philosophical language required ridding natural languages of their inconsistencies and peculiarities. Dascal, *Leibniz, language, signs, and thought*, however takes the position that Leibniz was more ‘modern’ than his contemporaries and preempts later perspectives on language where languages are not a hindrance but the possibility of rational thought
[21] Leibniz, Remnant, and Bennett, *New Essays on Human Understanding*.
[22] Ibid., 332-333.
[23] Ibid., 277.
[24] An interest that was fueled by the opportunities presented to him while visiting archives and old texts when writing the history of the House of Braunschwieg-Lünenburg, Aarslef, *From Locke to Saussure*.
[25] Leibniz, “On the connection between words and things.”
[26] Leibniz, Remnant, and Bennett, *New Essays on Human Understanding*, Book III, 282.
[27] Aarsleff, *From Locke to Saussure*.
[28] Leibniz, Remnant, and Bennett, *New Essays on Human Understanding*, Book III, 283.
[29] Leibniz quoted in Aarslef, *From Locke to Saussure*, 85.
[30] Rutherford, “Philosophy and language in Leibniz.”
[31] Ibid.
[32] Buckland, “European modernism and the information society .”
[33] Black, Muddiman, and Plant, *The early information society*
[34] Robins and Webster, *Times of the technoculture*
[35] Rayward, *European modernism and the information society*
[36] Rayward., *The Universe of Information.* It is also important to note that these aspirations by no means only were held by Otlet, but was shared by many of his contemporaries independent of each other. The search for new ways to classify human knowledge had in part been spurred with new intensity in the mid-19th century following the industrialization of the printing business, coupled with the advent of cheap binding materials that created an explosion in publishing
[37] Frohmann, “European modernism and the information society .”
[38] Otlet, *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet,* p. 149, 153.
[39] Otlet, Quote from Rayward, *The Universe of Information,* 67.
[40] Otlet, *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet,* 79.
[41] And in a fashion quite similar to the Mundaneum of Paul Otlet, Google has created a symbolic, as well as material, centre in Silicon Valley: the famous Googlplex attracts pilgrims of the digital age and generates a sprawling weblog literature, see Jakobsson & Stiernstedt.
[42] van den Heuvel, *European Modernism and Information Society.*
[43] Newspapers, journals, legal documents, books, certificates, statistics and correspondence.
[44] Rayward, *The universe of Information.*
[45] Ibid.
[46] Rayward, *European Modernism and the Information Society,* 12.
[47] Ibid. For a history of the card and catalogue see Krajewksi, *ZettelWirtschaft*.

[48] E.g. Lévy, *Collective intelligence*
[49] Page and Brin, "Method for node ranking in a linked database."
[50] Dale, Moisl, and Somers, *Handbook of natural language processing*.
[51] Ibid.
[52] Mohri, "Statistical natural language processing."
[53] Paşca, "Organizing and searching the world wide web of facts -- step two," 103.
[54] van Durme och Pasca , "Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction," 1244 (added emphasis).
[55] These physical centers of Google are shrouded in mystery. No one, outside the company, really knows where they are located, and no one knows how many these data centers are. But according to Miller there are at least 12 major data centers within the US and five more in Europe. There have been rumors of new data centers in Asia, for example in Taiwan, as well as speculations of new data centers in Lithuania (ibid., 2008). In 2009 it was made public that a new major data center was to open in Finland in 2010. It is estimated that Google spent about 2.4 billion dollars on four new data centers in 2007 and that as much as 30% of all servers shipped by the entire computer industry go into the data centers of Google, Amazon, Yahoo, eBay and Microsoft and a few others.

[56] Fan, Weber, and Barroso, "Power provisioning for a warehouse-sized computer."
[57] Vise and Malseed, *The Google story*
[58] Barroso, Dean , and Hölzle , "Web Search For a Planet: The Google Cluster Architecture."
[59] Pinheiro, Weber , and Barroso, "Failure Trends in Large Disk Drive Population."
[60] Ghemawat, Gobioff, and Leung, "The Google file system."
[61] This de-humanization of the information-labour is in a peculiar way underlined by the very human language used to describe how information moves – it migrates, or becomes orphaned – and how communication between the labouring servers and their master goes through "heartbeat-messages to give it instructions and collect its state", see Ghemavat et al.
[62] Barroso, "The Price of Performance," 49.
[63] Vise and Malseed, *The Google story*, 282
[64] Madhavan m.fl., "Structured Data Meets the Web: A few observations," 1.
[65] van Durme and Pasca , "Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction."
[66] Bunescu and Pasca, "Using encyclopedic knowledge for named entity disambiguation," 9.
[67] Ibid. 9.
[68] Ibid. 11.
[69] e.g. Paşca m.fl., "Names and similarities on the web."
[70] Culotta, McCallum, and Betz, "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," 296.
[71] Paşca, "Organizing and searching the world wide web of facts -- step two," 102.
[72] Ibid. 101.
[73] Ibid. 102.
[74] Ibid. 109.
[75] Bermejo, "Audience manufacture in historical perspective," 149
[76] e.g. Balnaves m.fl., "Introducing Ratings in Transition."
[77] You can get an estimation of the cost of different search strings at https://adwords.google.com/select/KeywordToolExternal.
[78] The understanding of audience commodity as an aggregate of this "worktime", packaged as ratings, pre-processed through broadcasting schedules and surveyed through the ratings industry's work, has been a commonplace of media analysis at least since the "blindspot debate" in the 1970s.
[79] Bermejo, "Audience manufacture in historical perspective."
[80] Ghemawat, Gobioff, och Leung, "The Google file system" (added emphasis).
[81] Halavais, *Search engine society*
[82] Deleuze, "Postscript on the societies of control."

References

Aarsleff, Hans. *From Locke to Saussure : essays on the study of language and intellectual history*. Minneapolis: Univ. of Minnesota Press, 1982.

Balnaves, Mark, Liz Ferrier, Gail Phillips, and Tom O'Regan. "Introducing Ratings in Transition." *Media International Australia Incorporating Culture and Policy*, no. 105 (2002): 6-9.

Barroso, Luiz Andre, Jeffery Dean , and Urs Hölzle . "Web Search For a Planet: The Google Cluster Architecture." *IEEE Micro* (April 2003): 22-28.

Barroso, Luiz André. “The Price of Performance.” *Queue* 3, no. 7 (2005): 48-53.
Bassett, Caroline. *The arc and the machine : narrative and new media*. Manchester: Manchester univ. press, 2007.
Bermejo, Fernando. “Audience manufacture in historical perspective: from broadcasting to Google.” *New Media Society* 11, no. 1-2 (Februari 1, 2009): 133-154.
Black, Alistair, Dave Muddiman, and Helen Plant. *The early information society : information management in Britain before the computer*. Aldershot: Ashgate, 2007. http://www.loc.gov/catdir/toc/ecip073/2006033557.html.
Bowker, Geoffrey C., and Susan Leigh Star. *Sorting things out : classification and its consequences*. Inside technology, 99-1592600-6. Cambridge, Mass.: MIT Press, 1999.
Buckland, Michael. “On the Cultural and Intellectual Context of European Documentation in the Early Twentieth Century.” In *European modernism and the information society : informing the present, understanding the past*, Ed. W. Boyd Rayward, 45-59. Aldershot, Hants, England ;: Ashgate, 2008. http://www.loc.gov/catdir/toc/ecip0720/2007023683.html.
Bunescu, R., and M. Pasca. “Using encyclopedic knowledge for named entity disambiguation.” In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 2006.
Cassirer, Ernst. *The philosophy of symbolic forms. Vol. 1, Language*. New Haven, 1977.
Culotta, Aron, Andrew McCallum, and Jonathan Betz. “Integrating probabilistic extraction models and data mining to discover relations and patterns in text.” In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 296-303. New York, New York: Association for Computational Linguistics, 2006. http://portal.acm.org/citation.cfm?id=1220835.1220873.
Dale, Robert, Hermann Moisl, and H. L. Somers. *Handbook of natural language processing*. New York: Marcel Dekker, 2000.
Dascal, Marcelo. *Leibniz, language, signs and thought : a collection of essays*. Foundations of semiotics, 0168-2555 ; 10. Amsterdam: J. Benjamins Pub. Co., 1987.
Davis, Martin. *The universal computer : the road from Leibniz to Turing*. New York: Norton, 2000.
Deleuze, Gilles. “Postscript on the societies of control.” *October* 59 (1992): 3-7.
Diaz, A. “Through the Google Goggles: Sociopolitical bias in search engine design.” In *Web search: multidisciplinary perspectives*, Eds. Amanda Spink and Michael Zimmer. Berlin Heidelberg: Springer, 2008.
van Durme, Benjamin , and Marius Pasca . “Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction.” Chicago, Illinois, 2008.
Fan, Xiaobo, Wolf-Dietrich Weber, and Luiz Andre Barroso. “Power provisioning for a warehouse-sized computer.” In *Proceedings of the 34th annual international symposium on Computer architecture*, 13-23. San Diego, California, USA: ACM, 2007. http://portal.acm.org/citation.cfm?id=1250662.1250665.
Foucault, Michel. *The order of things : an archaeology of the human sciences*. Routledge classics. London: Routledge, 1970.
Frohmann, Bernd. “The role of Facts in Paul Otlet’s Modernist Project of Documentation.” I *European modernism and the information society : informing the present, understanding the past*, Ed. W. Boyd Rayward, 75-89. Aldershot, Hants, England ;: Ashgate, 2008. http://www.loc.gov/catdir/toc/ecip0720/2007023683.html.
Fuller, Matthew. *Software studies : a lexicon*. Cambridge, Mass.: MIT Press, 2008.
Fuller, Matthew. *Media ecologies : materialist energies in art and technoculture*. Cambridge, Mass.: MIT Press, 2005.
Galloway, Alexander R. *Protocol : how control exists after decentralization*. Cambridge, Mass.: MIT Press, 2004.

Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. “The Google file system.” *SIGOPS Oper. Syst. Rev.* 37, no. 5 (2003): 29-43.
Goldman, E. “Search engine bias and the demise of search engine utopianism.” I *Web search: multidisciplinary perspectives*, Eds. Amanda Spink and Michael Zimmer. Berlin Heidelberg: Springer, 2008.
Google. “Corporate Information - Our Philosophy.” http://www.google.com/corporate/tenthings.html.
---. “Corporate Information - Technology Overview.” http://www.google.com/corporate/tech.html.
Halavais, Alexander. *Search engine society*. Digital media and society series. Cambridge: Polity Press, 2009.
van den Heuvel, Charles. “Building Society, Constructing Knowledge, Weaving the Web: Otlet’s Visualizations of a Global Information Society and His Concepts of a Universal Civilization.” In *European modernism and the information society : informing the present, understanding the past*, Ed. W. Boyd Rayward, 127-155. Aldershot, Hants, England ;: Ashgate, 2008. http://www.loc.gov/catdir/toc/ecip0720/2007023683.html.
Introna, L, and H Nissenbaum. “Shaping the web: why the politics of search engines matters.” *The information society*, no. 16 (2000): 169-185.
Jakobsson, Peter and Fredrik Stiernstedt. "Googleplex and Informational Culture". In Riegert, Kristina and Staffan Ericson (eds.). *Media Houses: Architecture, Media and the Production of Centrality.* New York: Peter Lang, forthcoming.
Kittler, Friedrich A., and John Johnston. *Literature, media, information systems : essays*. Critical voices in art, theory and culture, 1025-9325. Amsterdam: G+B Arts International, 1997.
Krajewski, Markus. *ZettelWirtschaft : die Geburt der Kartei aus dem Geiste der Bibliothek*. Copyrights (Berlin) ; 4. Berlin: Kulturverl. Kadmos, 2002.
Leibniz, Gottfried Wilhelm. “Introduction to a secret encyclopedia.” I *The art of controversies*, Ed. Marcelo Dascal, 219-224. Dordrecht: Springer, 2008.
---. *Leibniz: Selections*. Ed. Philip Paul Wiener. New York: Scribner, 1951.
---. “On the connection between words and things.” In *Leibniz. Language, Signs and Thought.*, Ed. Marcelo Dascal, 189-190. J. Benjamins Pub. Co., 1987.
---. “Preface to a Universal Characteristic (1678-79).” I *Philosophical essays*, Eds. Roger Ariew and Daniel Garber, 5-10. Indianapolis: Hackett Publishing Company, 1989.
---. “Towards a heuristics for discovery.” In *The art of controversies*, Ed. Marcelo Dascal, 93-104. Dordrecht: Springer, 2008.
---. “Two prefaces to the general science.” In *The art of controversies*, Ed. Marcelo Dascal, 213-218. Dordrecht: Springer, 2008.
Leibniz, Gottfried Wilhelm, Peter Remnant, and Jonathan Francis Bennett. *New Essays on Human Understanding*, 1996.
Leibniz, Gottfried Wilhelm von. “Dissertation on the art of combinations, 1666 (Selections).” In *Philosophical papers and letters*, Ed. Leroy E. Loemker, 73-84. Synthese historical library, 0082-111X ; 2. Dordrecht: Reidel, 1969.
Lev-on, A. “The democratizing effects of search engine use: on chance exposures and organizational hubs.” I *Web search: multidisciplinary perspectives*, Eds. Amanda Spink and Michael Zimmer. Berlin Heidelberg: Springer, 2008.
Lévy, Pierre. *Collective intelligence : mankind's emerging world in cyberspace*. New York: Plenum Trade, 1997.
Madhavan, Jayant, Alon Halevy, Shirley Cohen, Xin (Luna) Dong, Shawn R. Jeffery, David Ko, and Cong Yu . “Structured Data Meets the Web: A few observations.” *IEEE Data Engineering Bulletin* 29, no. 4 (December 2006).
Mattelart, Armand. *The information society : an introduction*. London: Sage, 2003.
Miller, Rich http://www.datacenterknowledge.com/ (Last visited, 2009-04-17).
Mohri, Mehryar. “Statistical natural language processing.” In *Applied Combinatorics on words*. Cambridge Univ. Press, 2005. http://www.cs.nyu.edu/~mohri/postscript/appcowC4.pdf.

Otlet, Paul, and W. Boyd Rayward. *International organisation and dissemination of knowledge : selected essays of Paul Otlet*. FID publ., 99-0103415-9 ; 684. Amsterdam: Elsevier, 1990.

Page, Larry, and Sergey Brin. “Method for node ranking in a linked database.”

Paşca, Marius. “Organizing and searching the world wide web of facts -- step two: harnessing the wisdom of the crowds.” In *Proceedings of the 16th international conference on World Wide Web*, 101-110. Banff, Alberta, Canada: ACM, 2007. http://portal.acm.org/citation.cfm?id=1242572.1242587.

Paşca, Marius, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. “Names and similarities on the web: fact extraction in the fast lane.” In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 809-816. Sydney, Australia: Association for Computational Linguistics, 2006. http://portal.acm.org/citation.cfm?id=1220175.1220277.

Pinheiro, Eduardo, Wolf-Dietrich Weber , and Luiz Andre Barroso. “Failure Trends in Large Disk Drive Population.” San José, California, 2007.

Pombo, Olga. *Leibniz and the problem of a universal language*. Materialien zur Geschichte der Sprachwissenschaft und der Semiotik, 0721-6920 ; 3. Münster: Nodus, 1987.

Rayward, W. Boyd. *European modernism and the information society : informing the present, understanding the past*. Aldershot, Hants, England ;: Ashgate, 2008. http://www.loc.gov/catdir/toc/ecip0720/2007023683.html.

Rayward, W. Boyd, and Paul Otlet. *The universe of information : the work of Paul Otlet for documentation and international organisation*. FID publ., 99-0103415-9 ; 520. Moscow: Kniga, 1975.

Robins, Kevin, and Frank Webster. *Times of the technoculture : from the information society to the virtual life*. Comedia, 99-0953478-9. London: Routledge, 1999.

Rutherford, Donald. “Philosophy and language in Leibniz.” In *The Cambridge companion to Leibniz*. Cambridge Univ. Press, 1995.

Werner, Stephen. “The Encyclopédie "Index".” In *Using the Encyclopédie : ways of knowing, ways of reading*, Eds. Julie Candler Hayes and Daniel Brewer, 265-270. Studies on Voltaire and the eighteenth century, 0435-2866 ; 2002:05. Oxford: Voltaire Foundation, 2002.

Wiener, Norbert. *Cybernetics or control and communication in the animal and the machine*. Cambridge, Mass.: The Technology Press, 1948.

Vise, David A., and Mark Malseed. *The Google story*. New York, N.Y.: Delacorte Press, 2005.

Zimmer, Michael. “Renvois of the past, present and future: hyperlinks and the structuring of knowledge from the Encyclopedie to Web 2.0.” *New Media Society* 11, no. 1-2 (Februari 1, 2009): 95-113.