CENTER FOR ADVANCED INTERNET STUDIES

Renée Ridgway

The Autonomous Surfer

CAIS Report

Fellowship
Mai bis Oktober 2018

GEFÖRDERT DURCH

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

The Autonomous Surfer

Research Questions

The Autonomous Surfer endeavoured to discover the unknown unknowns of alternative search through the following research questions: What are the alternatives to Google search? What are their hidden revenue models, even if they do not collect user data? How do they deliver divergent (and qualitative) results or knowledge? What are the criteria that determine ranking and relevance? How do p2p search engines such as YaCy work? Does it deliver alternative results compared to other search engines? Is there still a movement for a larger, public index? Can there be serendipitous search, which is the ability to come across books, articles, images, information, objects, and so forth, by chance?

Aims and Projected Results

My PhD research investigates Google search – its early development, its technological innovation, its business model of the past 20 years and how it works now. Furthermore, I have experimented with Tor (The Onion Router) in order to find out if I could be anonymous online, and if so, would I receive divergent results from Google with the same keywords. For my fellowship at CAIS I decided to first research search engines that were incorporated into the Tor browser as default (Startpage, Disconnect) or are the default browser now (DuckDuckGo). I then researched search engines in my original CAIS proposal that I had come across in my PhD but hadn't had the time to research; some are from the *Society of the Query Reader* (2014) and others I found en route or on colleagues' suggestions. The other focus was on privacy search engines and I wanted to discover what their business model is, if it is not user data. I also wanted to research a model of the commons, which is YaCy, a p2p distributed and decentralised search engine.

In the following I will provide a few examples of alternative (and privacy) search engines. Part of my fellowship focused on discourse analysis, reading texts and articles, but I also did hands-on querying. I wanted to learn more about gathering data and the different methods for making data visualisations. I was fortunate to be accepted to attend a conference in Barcelona where I learned some tools for data visualisation (e.g. Gephi) and presented my YaCy research as a CAIS fellow. Furthermore, I also worked on two journal articles, one of which is still under review. During my fellowship, CAIS financed my event application for an *Indexathon* on October 18–19, 2018.

Some Alternative Search Engines

One of my research questions was whether there was a way to search with serendipity and I began with The Serendipity Engine, which was shown to me by Leuphana Fellow Katrina Jungnickel. Together with Aleks Krotoski, Jungnickel made a test for users called 'Your Serendipitousness', drawn from answers to questions that fit into seven different scales. The scales are assigned 'weighting', a form of ranking that reflects importance and is based on psychological tests: Social Support, Creativity (x^2 weighting), Physical Well-Being, HeadRAM (x^2 weighting), Attention (x^2 weighting), Access to Knowledge and Grit. They propose that search engines are recommendation engines and that serendipity engines would be a much messier entanglement of humans, objects, tangible and intangible things. Although serendipity is a happy accident on the one hand, it also requires having the sagacity to recognize the wisdom or insight, along with knowing what to do with it.

Hyphe is developed by a group of researchers at SciencesPo médialab in Paris, France. Hyphe is a web corpus curation tool featuring a research-driven web crawler that builds a web corpus based on and organised by the web user. It works with 'web entities' that determine how web pages are grouped by URL (Uniform Resource Locator). With its flexible and dynamic memory structure it allows the user to change the definition of the web entity's boundaries. After playing around with the Hyphe Demo Tool, I decided I would like to experiment more and potentially create a research project around using it.

Wolfram Alpha is a 'computational knowledge engine' or 'answer engine' that does not return search results as a list of documents that might answer the query but attempts to answer the query itself from its own archive of 'curated data'. Bing, DuckDuckGo and other engines use Wolfram Alpha along with Apple's Siri, SVoice by Samsung and the speech recognition software for Android's Iris and Blackberry.

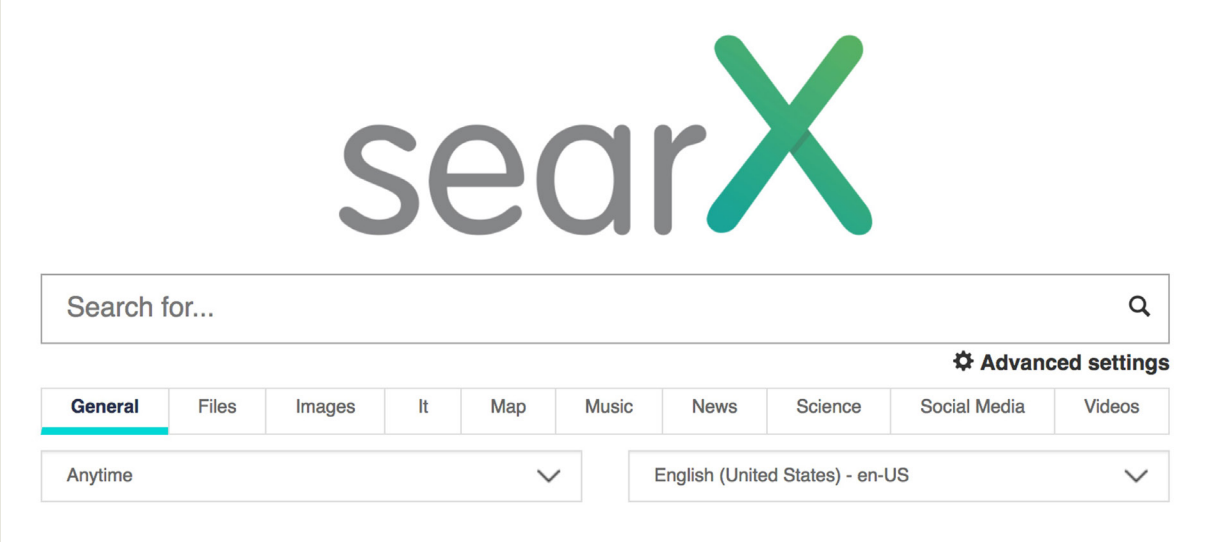
Privacy Search Engines: Revenue Models

Dirk Lewandowski's recent overview of the history of search engines gives a timeline of search engines, along with events, updates and technological innovations that have affected their development. However, it does not discuss business models and therefore I wanted to find out how search engines earn revenue, if there are different models that provide free services for data (as with Google). Ultimately, I would like to publish an article that describes these business models, their intricacies and differences.

One search engine advertising privacy the last years is DuckDuckGo that has gained ground on Google, because many users can now (since 2015) choose it as their default search engine for commercial services (smartphones and computer browsers such as Safari and Firefox). It is still the default search engine for the Tor (The Onion Router) Browser that enables anonymity online. According to Gennaro, DuckDuckGo fuses proprietary crawlers with APIs from websites, making it a 'hybrid' engine (2018a, 2018b). DuckDuckGo also offers suggestions when users type in queries, has a range of filter settings and offers 'Bang' where users can add a '!' after the query and the website name in order to search just that site. Their business model is quite simple, there is 'untracked advertising' based on users' keywords, where advertising is shown relative to the keyword. If users come to Amazon through DuckDuckGo and buy something, DuckDuckGo takes a small commission, called 'affiliate marketing' that DuckDuckGo declares does not determine the ranking (Gennaro 2018 a, 2018 b). It seems however that it sources its ads from Yahoo, which has a search alliance with Microsoft, which owns Bing, and that they operate on Amazon's servers.

Searx is an open-source meta-search engine with the principle of providing privacy to its users, as it blocks tracking cookies and user-profiling-based results modification. Run by volunteers, it was inspired by the Seeks project (ended 2016) that desired to give the user more control over their results, relying on a distributed 'collaborative filter' that lets the user personalise and share the results as a type of p2p user-sourced ranking. Searx provides a link to the site, unlike a 'tracked redirect link as used by Google', and by default Searx queries are submitted via HTTP POST to prevent users' query keywords

from appearing in webserver logs. Searx aggregates results from around 70 different engines including major ones and sites like Wikipedia and Reddit, where the engines used for each search category can be set with a 'preferences' interface. These settings will be saved in a cookie in the user's browser, rather than on the server side, since for privacy reasons, there is no user login. Searx does not offer suggestions but instead offers a self-hosting feature and makes it the only metasearch engine with no logging. Searx can proxy websites, yet there is a chance that this might 'break' the websites visited. It has a non-proprietary license with MIT. At the present moment I am considering researching Searx for a forthcoming project on alternative search because it offers the opportunity for users to download search results in RSS file, .json and .csv and I could set up my own version of a SearX instance.



Ranking

Users' Understanding of Search Engine Advertisements by Dirk Lewandowski (2017) gives an in-depth and detailed analysis regarding his mixed methods study (a survey, a task-based user study, an online experiment) to find out how users (1000 Germans) are generally unable to distinguish between advertisements and organic results. His conclusion is that many users (42%) cannot differentiate between ad-based SERPs (Search Engine Result Pages) and organic (without ads). He also recommends that search-based advertising be amended.

Indexing

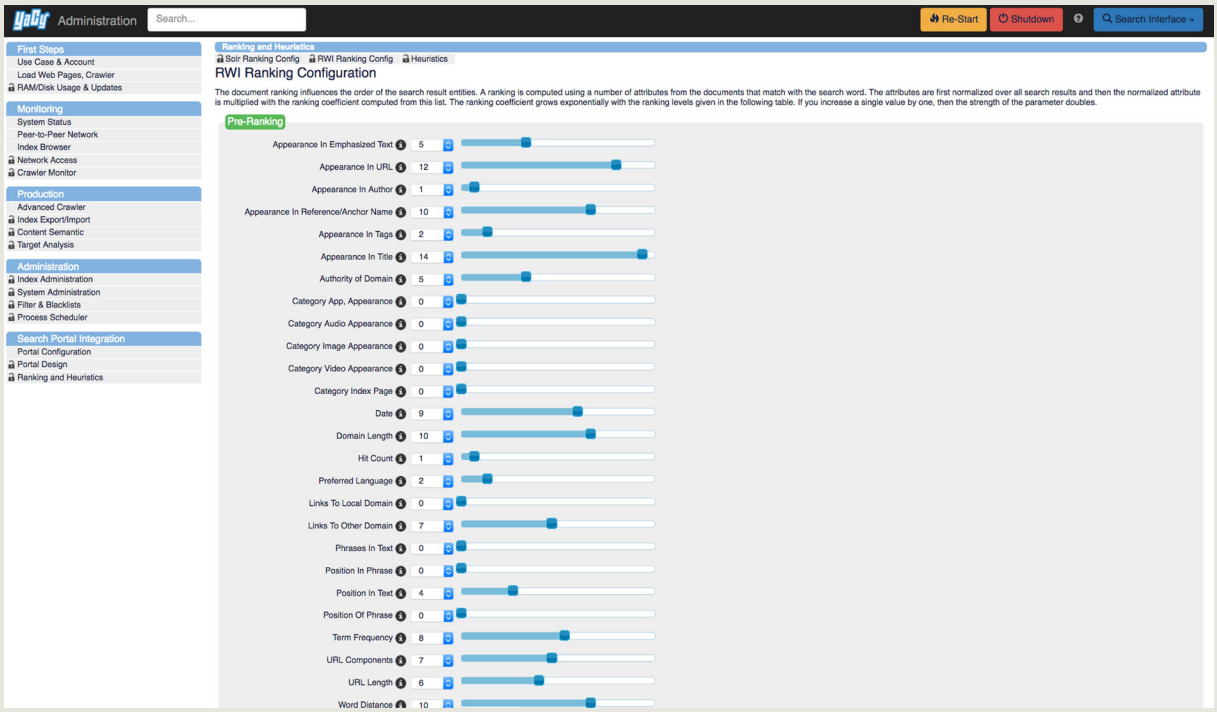
The concept of the index is crucial to the structure of a search engine, being the 'middle part' after the crawler has gathered the data and before the results are returned as rankings. As stated above, Google's index is the largest in the world and proprietary. Historically there were two major attempts in Europe to create a 'public index' (Quaero and Theseus), which both stranded for numerous reasons. One of the larger looming questions in the field of alternative search is whether a public index of the web is needed that would be accessible to a variety of search engines. Such an index 'would facilitate competition on the search engine market and allow for lots of smaller search projects to be realized' (Lewandowski 2014). This non-proprietary index would enable counter strategies of exploring new methods in finding information through search engines. I wanted to find out more about it and how there was a continuation of interest in Europe for a non-proprietary index. Through our email correspondence, I learned that Lewandowski has been working on the Open Web Index. Lewandowski is a professor of information research and information retrieval at the Hamburg University of Applied Sciences, and the head of Search Studies.

According to YaCy, a vital link is missing between the user and the information: 'free search'. In other words, in an era of hyperlinks and monopolistic infrastructures, which only make information visible for the good of the company, YaCy instead is the missing link that reflects search engines as a 'primary (and public) good' in the Rawlsian sense of the word. The general philosophical premise of YaCy is that information should be 'transparent, accountable and accessible for everyone'. As it is open-source and free software 'everyone can see how information is obtained for the search engine and displayed to the user' (BitcoinWiki 2020). Connecting the independent search-engine user with free content, YaCy embodies a decentralized technology that enables the user to not only consume content but to produce it. YaCy also is a private search engine, as the data is not collected or stored on servers.

Indexathon on October 18–19, 2018

At the end of my fellowship I organised an Indexathon, where 10 researchers were invited to come and test out alternative search methods, using the p2p search engine YaCy. Specifically, I was interested in indexing (p2p in the case of YaCy) on the one hand, and the results of searching on the other. The past months I had been testing out YaCy with various keywords in regard to my research, taking screenshots and saving webpages of what I found when searching, first with 'autocomplete', then comparing 'ranking' and understanding the different criteria of 'relevance'. I also wanted to discover if and how serendipity could play a role.

At the Indexathon I did a small test with the participants to search simultaneously with the keyword 'privacy' and compared the different results. Some preliminary conclusions show that each individual user received different results with YaCy, based on what they had searched and accumulated on their computer, dependent on their amount and depth of crawling, along with their local index. Michael Christen also explained his well-designed and informative back-end interface and I would like to carry out more tests/studies, create visualisations and produce an article that elucidates ranking (please see image). At the end of the two days, each participant gave a short description about their topic of research and what they attempted with YaCy the past days.



Mace Ojala, IT University Copenhagen, decided to feed YaCy a seedlist of URLs about 'legacy software' (his research topic), which he had collected earlier. Regarding the corpus, he also wanted to see if he could tap into YaCy, bypassing the graphical interface so that he could connect other workflows of Python projects with YaCy. He showed his search results using Jupyter notebook (an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text).

Winnie Soon and Geoff Cox (Aarhus University) are working on a book chapter project called 'queries' for the Aesthetic_Programming_Book to be published by Open Humanities Press and used the Indexathon to think through the concept of 'aesthetic programming' for their research. 'Hello World' is a programme used to introduce novice programmers to a programming language and to think through relevance and ranking. They began by creating a seedlist for YaCy, then experimented with the depth of crawl, and the patterning and distance for visualisations of the YaCy crawler.

Marcell Mars (Coventry University) and Dusan Barok (University of Amsterdam) worked on 'shadow' libraries, how to make something relevant through 'curation' by filtering collections. They indexed a few distributed and decentralised websites to YaCy and would like to federate the websites, only using the YaCy interface to see how quickly they could find books. They installed YaCy on a server to act as one 'peer' and started by adding three websites (Monoskop, Ubu and Memory of the World) but didn't have enough time to complete.

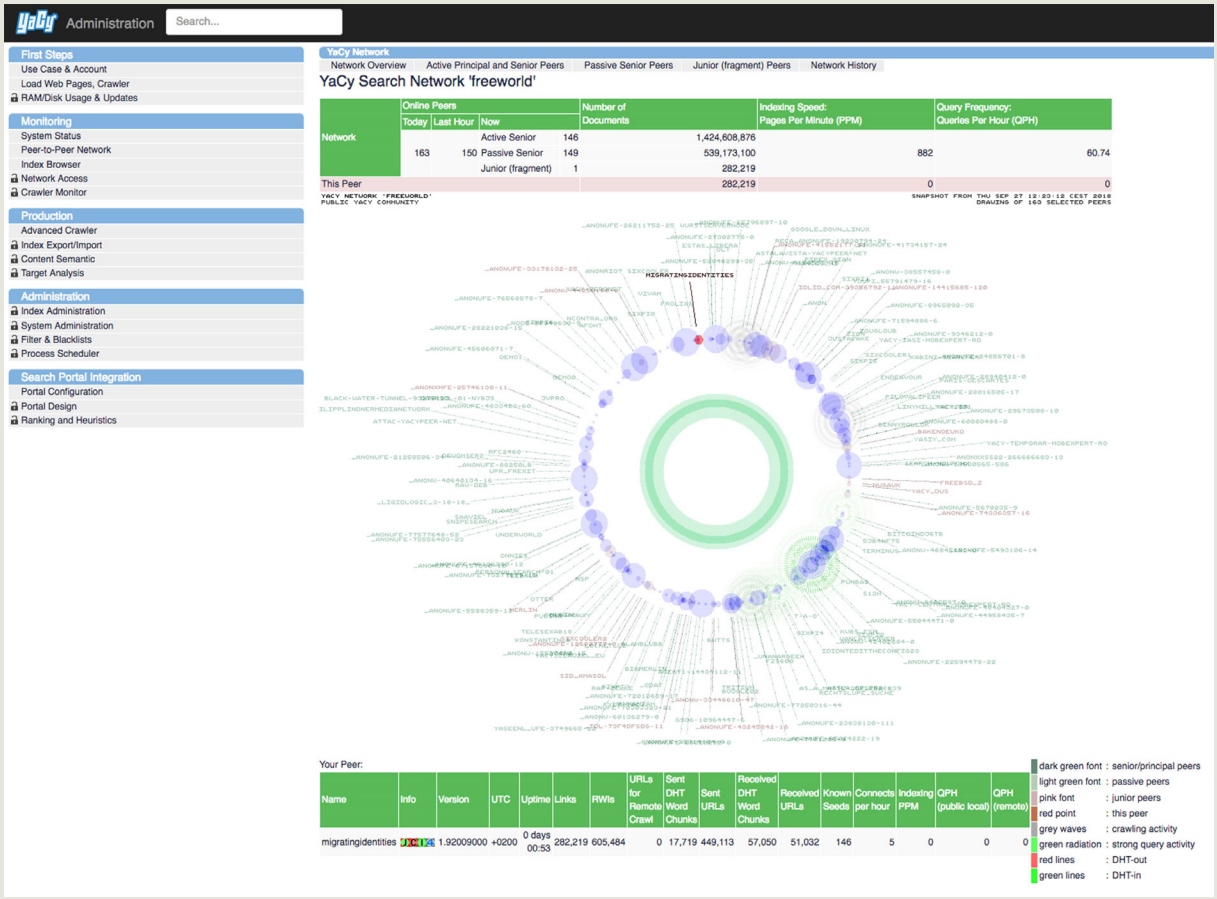
Bryan Newbold (Internet Archive) first showed Open Library, which contains several million books that are scanned by Internet Archive, yet it is centralised and expensive. FatCat (Fat Catalogue) is the name of his project, collecting PDFs of papers from the open web, and he compared that interface to YaCy (YaCy is much more graphical). The project could use YaCy so that people could find these books through metadata to make them more accessible and users can edit metadata about books. Using a 'development' machine at the Internet Archive he tried to crawl the directory of 'Open Access Books' and pull out content. Yet he decided that YaCy would be easier for people than Internet Archive's software.

Janneke Adema (Coventry) crawled Open Humanities Press and will crawl more open access publishing sites upon return.

Jurij Smrke (Coventry) wanted to use YaCy as a way to crawl webpages for his own research, which contain useful information. By using the built index, he could later extract the information from book metadata. He first used Wikipedia and DuckDuckGo to aggregate information about universities and their 'research citation rankings' (his research topic) and started crawling four sites. He used YaCy to search for funding, ranking and found a couple of interesting leads/links to continue on.

Jan Gerber (Open Media Library) is interested in large collections of media such as Pad.ma, which he tried to index into YaCy. He then installed YaCy on his server and also attempted to index Tor's hidden services to see the difference between what his results would be using YaCy, compared to the rest of the groups' searches with 'privacy' (Renée's experiments). He used the search engine Torch to first get results, then picked sites to index to YaCy.

After the short presentations Michael Christen also presented his YaCyGrid as a designed object, calling it a 'supercomputer simulator', which crawls, loads and parses. It is next generation YaCy but the YaCy Grid is more elastic, with a nice distribution function so that it can scale easily. Christen then gave a lecture on 'The Future of Search' about his open-source personal assistant, SUSI (susi.ai) that incorporates a better query language, in combination with the 'accumulated advantage' of machine learning and social media metadata.



The Search Engine of the Future

Michael Christen's newest development in search technology, SUSI (Social Universe Super Intelligence), started out as an investigation into formulating search requests (queries) in a different way – with natural language. If the answer to all questions is the search engine of the future, where is the free and open-source answer? Personal assistants (like Alexa, Siri) need a whole ecosystem that is connected to each other and to real world and skills – the digital assistant ecosphere. SUSI needed a unification data structure and by aggregating different search sources, YaCy p2p search engine, WolframAlpha, DBpedia, and Loklak Search (that collects messages from social media), these became 'SUSI Thought'. DBpedia is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects. This structured information resembles an open knowledge graph (OKG) which is available for everyone on the Web. A knowledge graph is a special kind of database, which stores knowledge in a machine-readable form and provides a means for information to be collected, organised, shared, searched and utilised. Google uses a similar approach to create those knowledge cards during search. We hope that this work will make it easier for the huge amount of information in Wikimedia projects to be used in some new interesting ways.

In order for SUSI to develop it needs conversation rules, drawing on first order logic deduction, which goes back into data memory and where it becomes 'reflection memory'. With each cycle, more information is generated and the learning increases, thereby producing more knowledge. The result became software – the AI assistant called 'SUSI Mind' –, a language driven search engine aggregation system. The idea behind it is that users create their own skill set for SUSI and the code is open-source and freely available, with everything that is learned backed up at github.

Ultimately the goal is to aggregate user knowledge, which leads to an acquisition of more knowledge for the user. People can send a pull request and then SUSI developers can look to see if they are valid. In the future SUSI will 'dream' functions as the plug-in technology where SUSI can write her own skills. What Wikipedia is for knowledge, SUSI should be for skills, with SUSI learning from the skills of users.

Speculating on Search

In the article *The Pataphysics of Creativity: developing a tool for creative search* by Hugill, Yang, Raczinski & Sawle (2013), the field of semantic search and its possibilities is brought to life through the musings on a pataphysical search engine *Syzygy*. The paper deals with 'patadata' structure (one step beyond metadata) and is an exploration into creativity that focuses on exceptions instead of norms. The search engine Syzygy 'aims to generate surprising, novel and provocative search results rather than relevant ones, in order to inspire a more creative interaction that has applications in both creative work and learning contexts' (Hugill et al. 2013, p. 237). The authors state that by using component-based architectures, which are pre-built, and interchanging them or making different arrangements, this would produce the intentional non-relevant returns through a semiotic system. The ways in which patadata could organise information pataphysically is a radical departure from 'inverted indexes' with its component infrastructures, how algorithms sort documents based on word groupings in traditional IR semantic systems, along with recognizing the fact that these systems of classification are outdated (Hugill et al. 2013, p. 249). The current architectures cannot support this proposed pataphysical algorithm as it requires precise data structures. The authors propose that there need to be new types of data structures, that 'component-based software architecture, graphical analogies and pataphysical algorithms would enable the possibility of non-ranked and unexpected search results' and that the Syzygy Surfer could provide 'inspirational learning through an exploratory search journey' to the user (Hugill et al. 2013, p. 249).

Preliminary Conclusions

There are alternatives to mainstream companies, with some of them echoing the model of free services for users' data, yet they have different advertisement models and do not keep user data. I also learned that 'searching is about retrieval of relevant answers and user relevance is the gold standard for comparison', which according to Tefko Saracevic (2013) is measurable (with information retrieval). There are various ways to adjust relevancy, as shown with the 'backend' of YaCy. However, alternative search is not only the result side of the algorithmic equation but of how the query is formulated. Asking specific questions has pushed the natural language query towards what Artificial Intelligence can comprehend. Instead of keyword search, the trend is moving to semantic or sentient analysis, such as with SUSI.ai.

Reflecting back on my fellowship and the Indexathon, I wonder if a 'federated index' could be created, similar to 'federated search', that aggregates the alternative indexes of other search engines and is open for anyone to use? This 'alternative search engine' would then be a non-proprietary index, which besides delivering results to the public is not based on an advertising model and would enable the development of specialised, academic or experimental projects.

Bibliography

Gennaro, Cuofano (2018a). *How Does DuckDuckGo Make Money? DuckDuckGo Business Model Explained*. Four Week MBA. Retrieved from https://fourweekmba.com/duckduckgo-business-model [12.06.2019].

Gennaro, Cuofano (2018b). *DuckDuckGo: The [Former] Solopreneur That Is Beating Google at Its Game*. Four Week MBA. Retrieved from https://fourweekmba.com/duckduckgo-vs-google [12.06.2019].

Hugill, Andrew, Yang, Hongji, Raczinski, Fania, & Sawle, James (2013). The pataphysics of creativity: developing a tool for creative search. *Digital Creativity*, 24 (3), 237–251.

König, René & Rasch, Miriam (Edit.). (2014). *Society of the Query Reader. Reflections on Web Search*. Amsterdam: Institute of Network Cultures. Retrieved from https://networkcultures.org/blog/publication/society-of-the-query-reader-reflections-on-web-search/ [27.01.2020].

Lewandowski, Dirk (2017). Users' Understanding of Search Engine Advertisements. *Journal of Information Science Theory and Practice*, 5 (4), 6–25.

Lewandowski, Dirk (2014). Why We Need an Independent Index of the Web. König, René & Rasch, Miriam (Edit.). (2014). *Society of the Query Reader #9: Reflections on Web Search*. Amsterdam: Institute of Network Cultures. Retrieved from https://networkcultures.org/query/2013/11/11/dirk-lewandowski-why-we-need-an-independent-index-of-the-web [27.01.2020].

Saracevic, Tefko (2013). *Relevance in Information Science: A Historical Perspective*. Lecture. Retrieved from https://www.youtube.com/watch?v=xBw2n-nf_TA [27.01.2020].

Table of Figures

Contact

Renée Ridgway
Da Costakade 158
1053XC Amsterdam
Netherlands