Towards an Open Web Index: Lessons From the Past

Michael Völske,[†] Janek Bevendorff,[†] Johannes Kiesel,[†] Benno Stein,[†]
Bauhaus-Universität Weimar, Germany
Maik Fröbe,[*] Matthias Hagen,[*] Martin-Luther-Universität Halle-Wittenberg, Germany
Martin Potthast,[‡] Leipzig University, Germany

Abstract

Recent efforts towards establishing an open and independent European web search infrastructure have the potential to contribute notably to more resilient, more equitable, and ultimately more effective information access in Europe and beyond. In this article, we recapitulate what we believe to be important goals and challenges that this effort should aspire to, and review how past web search endeavors have fared with respect to these criteria.

In a nutshell: For the past twenty years, no independent search engine has been able to establish itself as a fully viable alternative to the major players. The attempts so far underline the importance of both collaboration to close the scale gap and innovative new ideas for funding and bootstrapping a new contender. Future entrants to the web search market are well-advised to take note of existing approaches to collecting relevance feedback data in a privacy-protecting manner.

INTRODUCTION

The Open Search Foundation strives for a more open, diverse, and transparent web search landscape that offers a variety of independent search engines. Today's web search market, where a single market leader has held onto a global market share of over 90% for more than a decade, is in many ways the opposite of this ideal. Specifically from a European perspective, various pundits advocate for a more independent domestic digital infrastructure [28, 29].

On occasion, such arguments can appear borderline chauvinistic in their favoring, or even regarding as somehow inherently superior, domestic technology for its own sake. Nevertheless, we believe that there are good reasons to aspire to establish a new, viable, alternative. Since a good portion of contemporary economic and social activity depends on information access facilitated by web search to some extent, search engine monoculture presents a risk to the societies and economies that depend on it. The recognition of this risk is nothing new, as evidenced by a long history of efforts directed toward establishing alternatives; however, none so far can be considered fully successful.

This article is first and foremost an appeal to heed the lessons that can be learned from the challenges that alternative web search engines have encountered in the past. In order to provide a framework and context for these efforts, we summarize in the following what we consider to be the primary goals and challenges that an open web search engine must address.

(1) **Independence.** In today's interconnected world, access to information can be considered almost as critical to day-to-day life as fundamental infrastructure such as water supply and hospitals. The information flood being unmanageable without effective retrieval, web search indexes should be regarded as serving an essential public need, and like other essential infrastructure should be resilient to breakdown, and thus redundant [1]. From a global point of view, being as independent as possible from existing providers achieves this redundancy.

(2) **Scale.** Creating and maintaining a useful web index requires significant computational resources, as documents need to be crawled, stored, indexed, and re-crawled to maintain freshness [3]. For instance, Google reported an index size exceeding 100 PiB in 2017 spread across fifteen major data centers worldwide [12, 15].

(3) **User data.** Bootstrapping a successful search engine generally hinges on the chicken-and-egg problem of user data acquisition. While a web search frontend can certainly function without tracking users [2, 27], it proves difficult if not impossible to provide the best possible service without relevance feedback in a competitive environment where user tracking is the norm [7]. Any potential gains in search effectiveness through user data collection must be reconciled with the related, equally important, and yet somewhat contradictory goal of protecting users' privacy, which poses technical challenges of its own [4, 31].

(4) **Market penetration.** Closely related to the previous issue, even the best piece of technology is useless unless it is used. Key to the development of an open search infrastructure is a marketing strategy that overcomes the barriers to market entry due to the virtual omnipresence of one contender. It is futile to think that an alternative search engine will reach a wide adoption based only on it being against the incumbent. A clear-cut unique selling proposition must be obtained and defended, or else end users will stick to what they are used to and what is most convenient.

(5) **Funding.** Commercial search providers tend to finance their operations through selling ads, or selling search engine result page (SERP) placement directly [21]. From the perspective of a new contender in open search infrastructure, this approach appears forlorn. Not only might it contradict aspirations toward independence, but the market situation is vastly different today compared to more than 20 years ago, when digital advertising first established itself. This makes finding a funding model compatible with open search a key challenge that must be overcome.

(6) **Transparency.** In their role as "gatekeepers to information," the major commercial online services are coming under increasing scrutiny, due in large part to the opaqueness of their operations [14]. Clearly, in order to be trusted, an open search engine cannot be a black box; at the same time, a highly-transparent search engine brings its own challenges, such as being more easily exploitable by parties interested in manipulating its rankings [20].

[*] <first-name>.<last-name>@informatik.uni-halle.de
[†] <first-name>.<last-name>@uni-weimar.de
[‡] martin.potthast@uni-leipzig.de

In what follows, we examine a variety of past attempts at building alternative search engines with respect to how they address—or fail to address—the above goals and challenges. Following that, we outline a few ideas that we believe can help future endeavors fare better in these respects.

LESSONS FROM THE PAST

There have been many previous attempts at addressing the challenges facing alternative search providers. Table 1 surveys previous attempts at establishing general-purpose search engines, with a focus on recently-active endeavors, and evaluates them with respect to the goals and challenges identified in the previous section. While at first glance it seems that the number of available options is quite large, closer inspection makes clear that many of them do not satisfy our criteria. The following sections examine how each of the challenges has been tackled in the past.

Attempts at Fully Independent Alternatives

For a search engine to be fully independent, it needs its own crawling infrastructure to feed its own index and serve it to users with their own algorithm. The price tag of crawling and indexing the whole web can be put at around one or two years of time and well over one billion Dollars in cash [8], so to build a search engine fully independent of existing competitors is a hard if not impossible challenge for any newcomer. Despite the sheer volume of niche search engines, hardly any of them can actually be considered independent. Besides Microsoft's Bing and the long-established regional search engines Baidu and Yandex, only few operate their own index infrastructure and their own ranking algorithms. The recent shutdown of the Cliqz search engine has shown once again how high the entry barriers to the search market are, especially for companies who want to develop and maintain their own technology stack. Cliqz was a search engine that was "private by default" with a custom index and a ranking based upon the "Human Web" [6], a sink for anonymized and unlinkable user click and web traffic statistics totally devoid of directly or indirectly user-identifiable information. The goals of the Human Web were (1) to make after-the-fact record linking impossible, so data can only be used for its original intended purpose, and (2) to minimize the amount of data sent to the server by aggregating and cleansing data on the client. The approach was innovative and unlike many others in the industry, yet Cliqz failed to attract a critical mass of users to fund the endeavour and the search engine had to shut down, particularly in light of a looming and potentially long-lasting financial dry spell caused by the global COVID-19 pandemic [9]. Cliqz' legacy is a fading impression of a thorough attempt at developing an independent search infrastructure, but through their blog also an array of valuable insights into the depths of the search engine business, which are otherwise inaccessible to spectators outside the industry.

Qwant as another European competitor has set the goal to become a fully-independent search engine, but as of now it is still using Bing to improve the rankings; whether due to technical difficulties maintaining a complete index, showing relevant results without direct user feedback, or other reasons, is unknown. The technical difficulties of maintaining a central infrastructure is avoided by YaCy, an entirely distributed search engine [19], which has so far remained more of a technical curiosity than a practical and widely used search engine, however. Smaller contenders with independent search indexes include GigaBlast, Mojeek, and Exalead, which have all been active for more than fifteen years, but don't seem to match the search result quality of the major search engines [23].

Scaling a New Search Engine

Indexing the web requires a massive investment in infrastructure and it is easy to get lost in naive assumptions about the amount of resources needed, while on the other hand, we can assume that one does not need to outscale Google in order to provide a useful and competitive search engine. Overall, the size of the indexed web is estimated at around 60 billion* web pages, which easily exceeds the largest Common Crawl to date by an order of magnitude. Today, the Internet Archive stores around 60–70PB of archival data. The Wayback machine alone had indexed well over 20PB of data as of 2018 [18]. This number serves as a good base estimate for a representative sample of the indexable web. To provide a live index with complete full-text search and timely updates, however, the storage requirements can easily multiply. By seeding the crawler with only the top-ranked domains, the size of the index can be reduced significantly at the cost of completeness. A simple full-text index of a 1.6–2.1-billion document Common Crawl snapshot can be built at around 20–30TB with an additional 30TB for holding the original cached HTML pages [2]—about the size of Google's index back in 2004 [11]. Such an index contains no multimedia content, no user data, no knowledge graphs, and no recent updates. Adding overhead for redundant data storage at a factor of at least 1.5 or 2.0, additional space for storing fresh crawls, user logs, archival of old data, as well as space for processing and indexing new data, it is safe to assume that a minimum capacity of 50PB has to be planned for at the low end with an additional 25% on top to provide sufficient room for rebalancing data in the case of outages and as a general buffer before new hardware has to be acquired. Considering Google's 100PB index, these estimates are extremely conservative and the actual storage requirements for a real competitor may easily scale up to the Exabyte mark. As a more practical example, the Qwant index has a size of "several hundred terabytes" with 2PB of archival data [26] and yet the search engine still sees the need for complementing their ranking with results from Bing.

Storage alone obviously does not make a search engine, and serving an index of only a few terabytes to millions of users with billions of daily requests already requires a high-availability deployment of several hundred if not thousands of servers (or an equivalent number of cloud instances) in addition to the raw storage space. Hence, it does not come as a surprise that many players avoid these costs of maintaining their own indexing infrastructure entirely by using the indexes of their competitors. Examples for these types of meta search engines are DuckDuckGo and Ecosia (using Bing) and Startpage (using Google). This approach eliminates the most crucial hurdles of indexing the web, acquiring user click data, and building a useful ranking from it. It does not, however, solve any problems of dependence on competitors, and despite potentially being able to aggregate results from multiple

*https://www.worldwidewebsize.com/

Table 1: The search engines that this paper discusses, active years, most recent Alexa rank, country of headquarters, and their approaches to the identified problems.

Search engine	Years active	Alexa rank	Country	Independence	Scale	User data	Funding	Transparency
MetaGer	1996–today	64,210	DE	No	Uses Bing + Scopia	None	Ads, donations	Open source
Google	1997–today	1	US	Yes	Own datacenters	Own traffic	Ads	Closed
Yandex	1997–today	62	RU	Yes	Own datacenters	Own traffic	Ads	Closed
Startpage.com	1998–today	1,895	NL	No	Uses Google	None	Ads	Closed
Baidu	2000–today	5	CN	Yes	Own datacenters	Own traffic	Ads	Closed
Gigablast	2002–today	19,819	US	Yes	Own datacenters	None	B2B, donations	Open source
YaCy	2003–today	–	–	Yes	Decentralized	None	Donations	Open source
Exalead	2004–today	47,873	FR	Yes	Own datacenters	Own traffic	B2B	Closed
Mojeek	2004–today	414,308	UK	Yes	Own datacenters	None	B2B	Closed
Wikia Search	2007–2009	–	US	Yes	Community-moderated	User contribution	Ads	Open source
DuckDuckGo	2008–today	182	US	Hybrid	Uses Yahoo, Bing	None	Ads	Open source
Bing	2009–today	38	US	Yes	Own datacenters	Own traffic	Ads	Closed
Ecosia	2009–today	471	DE	No	Uses Bing	None	Ads	Closed
Qwant	2013–today	7,408	FR	Hybrid	Uses Bing + own index	None	Ads	Closed
Cliqz	2015–2020	52,948	DE	Yes	Own index	Human web	Ads	Mostly closed

search engines, the results will hardly outmatch those of any individual backend search engine. A unique selling point of most meta search engines therefore continues to be privacy, where the service pledges not to track users, while playing middle man to the search backend that does. In the end, such a search engine still indirectly relies on user tracking, where instead of tracking their own users, they are exploiting the fact that other users are willingly trading their data for a superior ranking. So even though the search engine itself protects its own users, it is more of an aftermarket product which does not solve any of the fundamental issues, and which would not be able to operate if other parties were not willing to collect user data in their stead.

Marketing Search and Addressing the User Data Issue

Setting aside arguments for or against user tracking, a new player on the market has to obtain this kind of data in order to attract users to their platform—a hard problem for a fresh and barely-frequented service. Buying the data from third parties lends itself as the most obvious solution, but not without establishing new dependencies on the companies one has set out to defeat in the first place; needless to say that Google and Facebook make for a combined share of more than 80% of the tracking market [31].

We have seen a number of approaches to get around the user data dilemma. In 2018, the French government decreed that all government agencies use Qwant as their default search engine instead of Google [16], thus generating and locking in a portion of users the search engine would not have had on a free market. The effectiveness can probably be measured in the upper hundreds of thousands or lower millions of users, but it remains a drop in the bucket even on just a national scale with a total market share of still less than 1%.[†] Moreover, if rolled out to the general public, the compatibility of such a forced approach with our ideals of freedom of choice on the one hand and innovation by competing on a level playing field on the other hand, is at least highly debatable.

A more market-oriented approach was pursued by Cliqz with their popular anti tracking browser extension Ghostery. Combining the seemingly incompatible goals of collecting data and preserving users' anonymity, Ghostery was used to fuel the rankings of the Cliqz search engine by tapping the accessible and already-penetrated anti-tracking market for data collection via the aforementioned Human Web approach. In the end, it was not enough to keep the search engine alive, but Ghostery and the Human Web live on as a clever idea and a data source that feeds off the ubiquity of Google's tracking codes on the web without actually relying on any single company.

Finally, the installation of sponsored browser toolbars (primarily as side load of third-party installers), has long served as an entry point for advertisers and data collectors into users' computers (recall the infamous Yahoo, Bing, or Alexa toolbars to name only a few). However, considering the rising popularity of central app stores, the growing share of mobile devices, stricter privacy regulations, the practical security hazards of installing unknown software, and the generally unilateral value these toolbars provide compared to extensions like Ghostery, the future of "unwanted" browser toolbars appears questionable at best.

We consider overcoming the user data issue with innovative, sustainable, and privacy-respecting ideas a key component of a successful search business; yet besides the technical infrastructure aspects, it remains one of the hardest problems to solve. Fortunately, promising approaches like the Human Web already exist, which may very well get more traction in the future. Given the desensitized nature of the data, a model such as this may even be viable as a public resource not under the control of any one entity—unless of course unforeseen progress is made in understanding users' queries and matching them with documents.

Table 1 also presents the search engines' most recent Alexa Top Sites ranking retrieved from the Wayback Machine in May 2020,[‡] as well as their country of headquarters. This highlights the fact that while the Chinese and Russian contenders can be considered realistic domestic alternatives, the same cannot be said about any of the European efforts so far, especially when considering those who operate their own indexes and infrastructure.

[†] Market share according to Statcounter https://gs.statcounter.com/search-engine-market-share/all/france

[‡] web.archive.org/web/20200501000011/http://s3.amazonaws.com/alexa-static/top-1m.csv.zip

Funding Web Indices and Search Engines

Like all services, both web indices and search engines require an investment both up front (e.g., development costs) and during operation (e.g., server costs). As Table 1 shows, most search providers today sell advertisements (ads) to cover their costs. Some search engines place ads similar to a genuine result into the search engine result page. However, a huge number of users is necessary to earn enough ad revenue for sustaining a fully independent search engine: "several hundred thousand daily users" are still not enough [9]. Several search engines save costs by not having their own infrastructure (see the lessons about scale). The decentralized peer-to-peer approach of YaCy reduces costs for the company even more, making it possible for the company to rely on donations—both in the form of money and code. As a special case, the Exalead search engine serves as a showcase for the search technology of the company—which they sell to businesses—and therefore is not supposed to generate income at all. Gigablast and Mojeek also provide services for business customers for income, but their main focus is the public search engine.

The means of funding described above all assume a search engine, which poses the question whether an open web index can be funded on its own. One source of income could be to sell the right for a commercial use of the index, similar to how some search engines today already buy access to the search of Bing, Google, or Yahoo (cf. Table 1). Another option is to request and rely on public funding. In the proposal for the Open Web Index, the author suggests that several states—like the European Union—are needed to fund the cost [22]. However, as the author notes, it would be important that the index is still operated without government influence. It is unclear when or if at all such funding will be available, and how reliable it would be. States have funded research in search engine technology in the past, most prominently the five-year Quaero research programme, but assuming this kind of funding is speculative. Though one might also be tempted to see an open web index as having similar value as public radio stations (both supposed to give unbiased access to information and other content), it might still be hard to argue for similar funding.

Attempts to Make Search More Transparent

All major search engines show result pages to users that do not explain why the listed documents appear in their order. Conversely, a transparent search engine would clarify the results so that they become trustworthy for users [10]. The existence of ads is often the only transparent part of commercial search engine result pages since ads are set apart from organic results. A fully transparent search engine can establish and justify users' trust by opening and explaining its behavior and internals. For example, the listing of ads can be enriched by the advertisers' price and the associated click-through-rate [17]. However, that disclosure of internal features becomes unusable for everyday users, since web-search pipelines combine at minimum hundreds of features [5] or use deep-learned models [25].

Consequently, transparent and explainable search engines are still an open and essential research topic [10]. Hence, it comes at no surprise that none of the surveyed search engines tries to clarify the ranking of results to its users. A few search engines (cf. Table 1) follow the simple approach to increase their transparency and trustworthiness by publishing their system as open-source software. Still, most surveyed search engines operate in a fully closed manner. Unfortunately, the search engines that publish their algorithms as open-source have only a small market share (MetaGer, Gigablast, YaCy) or do not act as a fully independent search engine (DuckDuckGo). The resulting gap of transparent search engines that index representative parts of the web and maintain a non-negligible market share is still to be closed.

The Open Web Index proposal [22] is perhaps the most comprehensive call for a fully transparent search infrastructure so far. The idea, which in its basic form exists since the web search monopoly first began to emerge [30], centers around the web index as a service, which lays the groundwork for derivative services on top. The web index itself is open to everyone and deployed on top of public, distributed, and shared infrastructure. Without the massive burden of maintaining a web-index, a multitude of services and search engines may arise. The resulting ecosystem makes it simpler to solve transparency issues that current search engines face, since the infrastructure part, which encompasses substantial business value and investments, is shared anyway. The huge funding challenge that comes with the development and maintenance of a public web index is a major weakness of the Open Web Index proposal. The suggestion mentioned in the proposal, that the EU should fund the web index through a foundation seems unlikely for the next few years when we consider the recent shutdown of Cliqz [9].

CONCLUSIONS, IDEAS, AND OPEN QUESTIONS

Data as a public good and search as critical public infrastructure have repeatedly sparked the desire for establishing a wider and more diverse competitive landscape in a heavily monopolized market. However, the extreme entry costs and the many failed attempts require more critical thinking as to what we as a society want to achieve, the way we tackle the problem, and why. We strive for a more grounded discussion beyond national interests and fear-mongering with regards to foreign mega corporations. While the problem does undeniably have a severe political dimension in the way monopolists are allowed to use their platform for promoting and locking in users to their other products, there are also other aspects that merit or deter the creation of an independent European web indexing infrastructure and the immense costs and narrow chances of success make critical reflection on our motivation even more crucial.

Collaborate to Achieve Scale

We strongly endorse previous calls to tackle the immense scale challenges that must be overcome on the way to a practically useful search infrastructure through collaboration [24]. An important question is how a joint effort of European public computing infrastructure can reach the scale required, considering that this infrastructure already has other primary scientific missions that take up most of its capacity.

Recently, university data centers are converging more and more toward the role of Infrastructure-as-a-service (IaaS) providers—the Heidelberg University cloud infrastructure heiCLOUD[§] being an example where this transformation has advanced particularly far. While chiefly a consequence of the ever more complex information

[§]https://heicloud.uni-heidelberg.de/en/

technology needs of academia, this development is also grounds for optimism regarding these institutions' technical capacity to support new endeavors such as an open web search platform.

In terms of tangible, recommended courses of action, it appears prudent to take a detailed inventory of public computing infrastructure that might be able to donate some capacity; on that foundation, one may reach out and work towards strategic collaborations. National governing bodies such as the German *Zentren für Kommunikation und Informationsverarbeitung in Lehre und Forschung* (ZKI e.V.) or EU-level research institutions such as CERN may be of assistance in this regard.

Create an Open Web Data Exchange

The recent collapse of the Cliqz search engine was precipitated at least in part by a lack of confidence in the availability of significant public funding toward establishing an open web search infrastructure in the near to mid term. That situation being what it is, alternative avenues towards a sustainable open search infrastructure must be explored. Managing access to the datasets and computing infrastructure that make web search possible will be a key to success in this regard.

A question that has so far gotten rather little attention is how to bring together the different parties interested in an open web index, and to create incentives for contributing resources such as datasets, storage, compute, person hours, or funds. We suggest that an "open web data exchange" might allow to trade such contributions to the web index for the ability to benefit from its use to a greater degree; the more an organization contributes, the more they are allowed to benefit, to the point where companies are able to monetize a service derived from the open web index, while financing its further development in turn.

One could operationalize such an idea by way of a distributed API credit system in which resources contributed to the search index are exchanged for requests. Appropriate consensus algorithms can be used to ensure proper attribution of contributions. For instance, Proof-of-Space-based blockchains [13] have already been employed to implement peer-to-peer cloud storage[¶] and could just as well track contributions of storage space towards the open web index. Other contributions may be more difficult to valuate, as the worth of a dataset may become apparent only once it is materialized in a concrete service. Hence, one may want to re-evaluate such contributions over time like in a real-world trading exchange.

Find Small Early Wins

A new open search engine should strive to be useful for some tangible purpose as early as possible. To this end, it appears prudent to initially focus the new web index's efforts towards niche use cases that are currently not well served by mainstream search engines, which may be the case, e.g., due to lack of specialized data, or lack of market incentives. Harvesting such low-hanging fruits first can both kindle initial interest by the public and thus secure further funding and also make use of existing resources that are already under the control of entities outside the global search business. Potential candidates would be search for citizen oversight (public records, laws parliamentary debates), digital assets of national and local libraries, public datasets (e.g., using CERN's Zenodo platform), and academic open-access publications. While there are frontends for the latter two provided by the big search companies, they can in many aspects be seen as almost lackluster compared to the quality standards of their main web search business, which leaves breathing room for other contenders (of which there are already quite a few, such as Semantic Scholar, ResearchGate, and a number of field-specific preprint repositories).

[¶]Such as https://storj.io

Converge on Common Goals

In the end, we have to ask ourselves: Are we content with building an alternative Google or do we want a different product that fits our needs better? Is better privacy alone a sufficient incentive for users to switch to an alternative or is it not rather a fundamental design decision and thus a byproduct of the user-facing service? Do we even build privacy into our products or do we merely rely on (supposedly) superior geo-political circumstances? Why can we not trust Google to handle their primary asset responsibly in the first place? Are we willing to invest in the massive infrastructure costs for building a competitive web index or do we focus on an underrepresented niche? What resources is building a new web index worth to us, and how do we best trade off goals of digital resilience against building data repositories for other efforts such as combating cancer, fighting infectious diseases, or understanding the origins of the universe?

We believe that with the right model and incentives, we can build a valuable and sustainable data infrastructure, but focusing on ideology or technical specifications for building a web index will not be enough. Many more issues need to be solved such as how and what kind of user data is being acquired, how a new web index serves needs that are not already met by existing solutions, how users can benefit so they start using it, how data owners can be incentivized to contribute their own data, and obviously how such a Herculean endeavor can be financed sustainably.

REFERENCES

[1] M. Baram. Resilience and Essential Public Infrastructure. In S. Wiig and B. Fahlbruch, editors, *Exploring Resilience: A Scientific Journey from Practice to Theory*, SpringerBriefs in Applied Sciences and Technology, pages 33–40. Springer International Publishing, Cham, 2019.

[2] J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In L. Azzopardi, A. Hanbury, G. Pasi, and B. Piwowarski, editors, *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, Mar. 2018. Springer.

[3] B. B. Cambazoglu and R. Baeza-Yates. *Scalability Challenges in Web Search Engines*. Morgan & Claypool Publishers, Dec. 2015.

[4] A. Catarineu, P. Claßen, K. Modi, and J. M. Pujol. Preventing Attacks on Anonymous Data Collection. *arXiv:1812.07927 [cs]*, Dec. 2018.

[5] O. Chapelle and Y. Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*, pages 1–24, 2011.

[6] Cliqz GmbH. Human Web—Collecting Data in a Socially Responsible Manner. https://web.archive.org/web/20191203182249/https:

//0x65.dev/blog/2019-12-03/human-web-collecting-data-in-a-socially-responsible-manner.html, Dec. 2019.

[7] Cliqz GmbH. Is Data Collection Evil? https://web.archive.org/web/20191202175142/https://0x65.dev/blog/2019-12-02/is-data-collection-evil.html, Dec. 2019.

[8] Cliqz GmbH. A New Search Engine: Cliqz Journey. https://web.archive.org/web/20191205214556/https://0x65.dev/blog/2019-12-05/a-new-search-engine.html, Dec. 2019.

[9] Cliqz GmbH. Farewell From Cliqz. https://web.archive.org/web/20200501153126/cliqz.com/en/magazine/farewell-from-cliqz, Apr. 2020.

[10] J. S. Culpepper, F. Diaz, and M. D. Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA, 2018.

[11] A. Das and A. Jain. Indexing the world wide web: The journey so far. In *Next Generation Search Engines: Advanced Models for Information Retrieval*, pages 1–28. IGI Global, 2012.

[12] Data Center Knowledge. Google Data Center FAQ & Locations. https://web.archive.org/web/20170317032243/http://www.datacenterknowledge.com/archives/2017/03/16/google-data-center-faq/, Mar. 2017.

[13] S. Dziembowski, S. Faust, V. Kolmogorov, and K. Pietrzak. Proofs of Space. In R. Gennaro and M. Robshaw, editors, *Advances in Cryptology – CRYPTO 2015*, Lecture Notes in Computer Science, pages 585–605, Berlin, Heidelberg, 2015. Springer.

[14] R. Epstein. To Break Google's Monopoly on Search, Make Its Index Public. *Bloomberg.com*, July 2019. https://www.bloomberg.com/news/articles/2019-07-15/to-break-google-s-monopoly-on-search-make-its-index-public.

[15] Google LLC. How Google's Site Crawlers Index Your Site. https://web.archive.org/web/20170515125549/https://www.google.com/search/howsearchworks/crawling-indexing/, 2017 (accessed June 2020).

[16] C. Goujard. France is ditching google to reclaim its online independence. https://web.archive.org/web/20200426235334/https://www.wired.co.uk/article/google-france-silicon-valley, Nov. 2018 (accessed April 2020).

[17] W. Gross and S. Olson. Transparent search engine, June 2005. US Patent App. 11/006,078.

[18] V. Heffernan. Things break and decay on the internet—that's a good thing. https://web.archive.org/web/20180925130510/https://www.wired.com/story/wired25-virginia-heffernan-internet-archive-wayback-machine/, Sept. 2018 (accessed September 2018).

[19] M. Herrmann, R. Zhang, K.-C. Ning, C. Diaz, and B. Preneel. Censorship-Resistant and Privacy-Preserving Distributed Web Search. In *14-Th IEEE International Conference on Peer-to-Peer Computing*, pages 1–10, Sept. 2014.

[20] J. Kun. Who would transparency in search algorithms benefit? https://medium.com/@jeremyjkun/who-would-transparency-in-search-algorithms-benefit-6af6b0b388c3, Nov. 2016.

[21] L. S.-L. Lai. In Search of Excellence – Google vs Baidu. Dec. 2011.

[22] D. Lewandowski. The Web Is Missing an Essential Part of Infrastructure: An Open Web Index. *Communications of the ACM*, 62(4):24, Mar. 2019.

[23] LibreTechTips. Detailed tests of search engines: Google, Startpage, Bing, DuckDuckGo, metaGer, Ecosia, Swisscows, Searx, Qwant, Yandex, and Mojeek - LibreTechTips. https://libretechtips.gitlab.io/detailed-tests-of-search-engines-google-startpage-bing-duckduckgo-metager-ecosia-swisscows-searx-qwant-yandex-and-mojeek/, Feb. 2020.

[24] Open Search Foundation. A free and open Internet Search Infrastructure. https://opensearchfoundation.org/wp-content/uploads/2019/12/A-free-and-open-Internet-Search-Infrastructure-web.pdf, June 2019.

[25] Pandu Nayak. Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/, 2019 (accessed June 2020).

[26] Qwant SAS. How Microsoft tools strengthen Qwant. https://betterweb.qwant.com/en/how-microsoft-tools-strengthen-qwant/, June 2019 (accessed July 2020).

[27] Qwant SAS. Qwant Privacy Policy. https://web.archive.org/web/20200614104928/https://about.qwant.com/legal/privacy/, 2020 (accessed June 2020).

[28] M. Stothard. How Europe can dominate the next decade of tech. *Sifted*, June 2020. https://sifted.eu/articles/european-tech-startups/.

[29] A. Voss. Digital autonomy. *The Parliament Magazine*, Mar. 2020. https://www.theparliamentmagazine.eu/news/article/digital-autonomy.

[30] P. Wilford. OpenIndex: Creating a Public Internet Index. https://web.archive.org/web/20040423010643/http://openindex.org/, Apr. 2004.

[31] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol. Tracking the Trackers. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 121–132, Montréal, Québec, Canada, Apr. 2016. International World Wide Web Conferences Steering Committee.