


# Impact and development of an Open Web Index for open web search

Michael Granitzer<sup>1</sup>  | Stefan Voigt<sup>2,3</sup>  | Noor Afshan Fathima<sup>4</sup>  |  
 Martin Golasowski<sup>5</sup> | Christian Guetl<sup>6</sup>  | Tobias Hecking<sup>3</sup>  |  
 Gijs Hendriksen<sup>7</sup>  | Djoerd Hiemstra<sup>7</sup>  | Jan Martinovič<sup>5</sup> |  
 Jelena Mitrović<sup>1</sup>  | Izidor Mlakar<sup>8</sup> | Stavros Moiras<sup>4</sup> |  
 Alexander Nussbaumer<sup>6</sup>  | Per Öster<sup>9</sup>  | Martin Potthast<sup>10</sup>  |  
 Marjana Senčar Srdič<sup>8</sup> | Sharikadze Megi<sup>11</sup> | Kateřina Slaninová<sup>5</sup> |  
 Benno Stein<sup>12</sup>  | Arjen P. de Vries<sup>7</sup>  | Vít Vondrák<sup>5</sup> | Andreas Wagner<sup>4</sup> |  
 Saber Zerhoubi<sup>1</sup> 

<sup>1</sup>University of Passau, Passau, Germany

<sup>2</sup>Open Search Foundation, Starnberg, Germany

<sup>3</sup>German Aerospace Center (DLR), Cologne, Germany

<sup>4</sup>CERN, Geneva, Switzerland

<sup>5</sup>IT4I, VSB – Technical University of Ostrava, Ostrava, Czech Republic

<sup>6</sup>Graz University of Technology, Graz, Austria

<sup>7</sup>Radboud University, Nijmegen, The Netherlands

<sup>8</sup>A1, Ljubljana, Slovenia

<sup>9</sup>CSC – IT Center for Science, Espoo, Finland

<sup>10</sup>Leipzig University and ScaDS.AI, Leipzig, Germany

<sup>11</sup>Leibniz Supercomputing Centre, Munich, Germany

<sup>12</sup>Bauhaus-Universität Weimar, Weimar, Germany

## Correspondence

Michael Granitzer, University of Passau, 94032 Passau, Germany.

Email: [michael.granitzer@uni-passau.de](mailto:michael.granitzer@uni-passau.de)

## Funding information

European Commission, Grant/Award Number: 101070014

## Abstract

Web search is a crucial technology for the digital economy. Dominated by a few gatekeepers focused on commercial success, however, web publishers have to optimize their content for these gatekeepers, resulting in a closed ecosystem of search engines as well as the risk of publishers sacrificing quality. To encourage an open search ecosystem and offer users genuine choice among alternative search engines, we propose the development of an Open Web Index (OWI). We outline six core principles for developing and maintaining an open index, based on open data principles, legal compliance, and collaborative technology development. The combination of an open index with what we call declarative search engines will facilitate the development of vertical search engines and innovative web data products (including, e.g., large language models), enabling a fair and open information space. This framework underpins the EU-funded project OpenWebSearch.EU, marking the first step towards realizing an Open Web Index.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

## 1 | INTRODUCTION

Leveraging the Web as a data source presents technical challenges, necessitating substantial infrastructure investments and expertise in collecting and processing its heterogeneous data. Legal and ethical privacy concerns aside, large commercial web search engines like Google, Bing, Baidu, or Yandex maintain their market dominance due to economies of scale as well as having both the data and advanced proprietary processing tools ready (Baeza-Yates & Maarek, 2012). This barrier disadvantages researchers and innovators outside major internet companies (Lewandowski, 2015), resulting in a rich-get-richer effect (Hagiu & Wright, 2020). Their market dominance also hinders policymakers from opposing unwelcome corporate objectives, as no alternatives exist to replace non-compliant commercial search engines, which makes them a gatekeeper to information for society. This can lead to issues such as search engine manipulation of public behavior (Epstein et al., 2017), increased advertisement, and limited access to diverse information (Barker, 2018). Moreover, the current ecosystem pressures smaller web contributors across various fields to comply with gatekeeper-imposed search engine optimization rules or risk digital invisibility. Overall, the current situation requires a change towards a more open web search ecosystem for more diverse information access that is not forced to comply with the economic interests of the few gatekeepers. Such an open web search ecosystem should also address the needs of a broad user-base ranging from very heterogeneous individual users over content producers, service providers, and organizations in general.

In this opinion paper, we argue that an Open Web Index (OWI) as first proposed by Lewandowski (2019) can form the basis for such an open web search ecosystem. We expand on this concept of an OWI and focus on the challenges and questions related to its creation and long-term sustainability. Specifically, we address three main areas:

1. Principles of an OWI: Different to Lewandowski (2019), we identify the need for an open index to be treated as open data. This allows the separation of index creation and search, thus reducing costs and increasing utility. Additionally, we discuss extensibility, collaborative creation, user content, control, and legal compliance.
2. Construction of an OWI: We outline a method for constructing and distributing such an index based on the six principles, leading to the concept of declarative search engines as a cost-effective approach to enable web-scale search.
3. Applications of an OWI: Assuming the existence of an open index, we explore potential applications and impacts, indicating economic potential and future possibilities.

In what follows, we first review the challenges underlying search engine construction and then discuss each of these main areas in turn.

## 2 | CORE CHALLENGES FOR BUILDING A WEB SEARCH ENGINE

Let us start by reviewing previous attempts at establishing alternative web search engines (see Table 1). We aim to identify the core challenges of building an OWI, namely (a) web data crawling at scale, (b) size and storage for an index, and (c) serving the index for billions of users. We focus on recent endeavors and consider only search engines that aim to cover nearly all of the Web and serve a broad user base with diverse search goals, excluding special-purpose search engines like digital libraries.

Building a fully independent search engine requires its own crawling infrastructure to feed its own index and serve it to users with their own ranking algorithm. However, the price tag of crawling and indexing the whole web can be put at around 1 or 2 years of time and well over one billion dollars in cash (Cliqz GmbH, 2019). Hence, besides Google, Microsoft's Bing, and long-established regional search engines Baidu and Yandex, few operate their own index infrastructure and their own ranking algorithms.

Qwant as European competitor has set the goal to become a fully-independent search engine, but as of now it is still using Bing to improve its rankings; whether due to technical difficulties maintaining a complete index, showing relevant results without direct user feedback, or other reasons, is unknown to the authors. YaCy is an entirely distributed search engine that avoids the technical difficulties of maintaining a central infrastructure (Herrmann et al., 2014), but it has remained more of a technical curiosity than a practical and widely used search engine. Smaller contenders with independent search indexes include GigaBlast, Mojeek, and Exalead, but they do not seem to match the search result quality of the major search engines (LibreTechTips, 2020).

Overall, the size of the indexed web is estimated at approximately 60 billion<sup>1</sup> web pages. Hence, indexing the web requires a massive investment in infrastructure. A simple full-text index of a 1.6–2.1-billion document can be built at around 20–30 TB with an additional 30 TB for holding the original cached HTML pages (Bevendorff et al., 2018)—about the size of Google's index back in 2004 (Das & Jain, 2012). Such an index contains no multimedia content, no user data, no knowledge graphs, and no recent updates.

Table 2 provides a rough estimate on a bare-minimum text index with no additional signals

TABLE 1 Comparison of search engines

Search engine	Years active	Alexa rank	Country	Ind.	Scale	User data	Funding	Transparency
MetaGer	1996–today	64,210	DE	No	Uses Bing + Scopia	None	Ads, donations	Open source
Google	1997–today	1	USA	Yes	Own datacenters	Own traffic	Ads	Closed
Yandex	1997–today	62	RU	Yes	Own datacenters	Own traffic	Ads	Closed
Startpage.com	1998–today	1895	NL	No	Uses Google	None	Ads	Closed
Naver	1999–today		KR	Yes	Own datacenters	Own traffic	Ads	Closed
Baidu	2000–today	5	CN	Yes	Own datacenters	Own traffic	Ads	Closed
Gigablast	2002–today	19,819	USA	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		USA	Yes	Community-moderated	User contribution	Ads	Open source
DuckDuckGo	2008–today	182	USA	Hybrid	Uses Yahoo, Bing	None	Ads	Open source
Bing	2009–today	38	USA	Yes	Own datacenters	Own traffic	Ads	Closed
Ecosia	2009–today	471	DE	No	Uses Bing	None	Ads	Closed
Qwant	2013–today	7408	FR	Hybrid	Uses Bing + own index	None	Ads	Closed
Cliqz	2015–2020	52,948	DE	Yes	Own index	Human web	Ads	Mostly closed
Brave Search	2021–today							
Neeva	2021–today							
You.com	2021–today		USA	No	Uses Bing	Hybrid	Venture capital	Closed

(e.g., usage data) that adds up to 10 PB storage. Consequently, assuming integrating further usage data and metadata it is safe to assume that a minimum capacity of 50 PB has to be planned for at the low end. Considering Google's 100 PB index, these estimates are extremely conservative and the actual storage requirements may further increase. As a more practical example, the Qwant index had a size of “several hundred terabytes” with 2 PB of archival data (Qwant SAS, 2019) in 2019 and yet the search engine still sees the need for complementing their ranking with results from Bing.

Besides storage, 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 plus servers. 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 of these types of meta search engines are DuckDuckGo and Ecosia and Startpage. 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 search engines, the results will hardly outperform 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 serving as middleman 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.

### 3 | PRINCIPLES OF AN OPEN WEB INDEX

A Web Index is a fundamental data structure that enables rapid content-based access, sorting, and filtering of extensive web documents and forms the backbone of every web search engine. The quality of a web index relies on the excellence of the indexed documents augmented by additional signals such as usage data, metadata, or link

**TABLE 2** Estimated storage and computing resources extrapolated from Bevendorff et al. (2018) to 60 billion web pages (text-only)

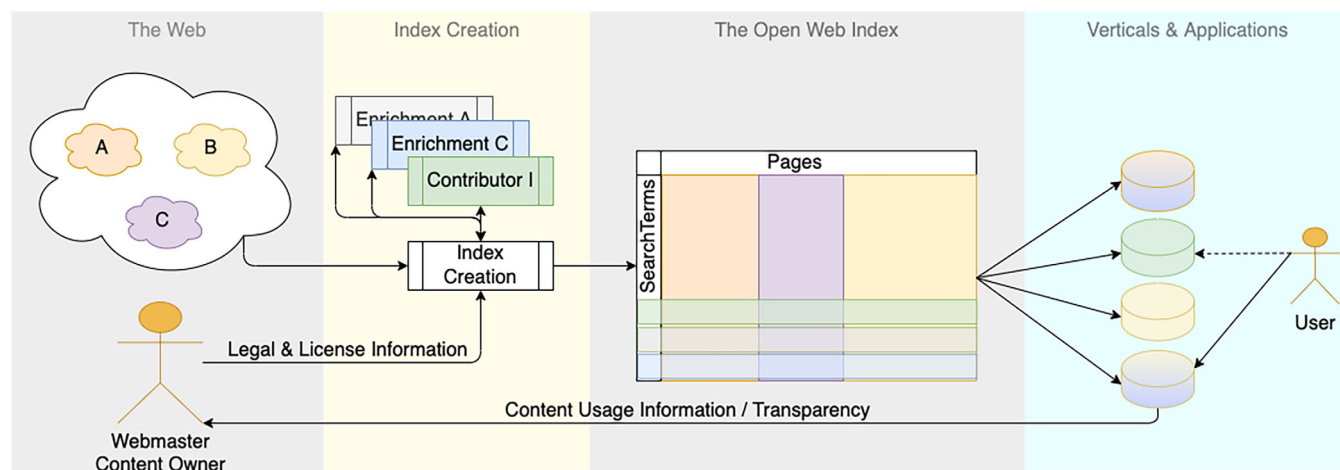
Entities/component	Technical specification
Estimate for storage raw data (replicated 3 times)	6000 TiB
Size of the Open Web Index (replicated 3 times)	2000 TiB (fast access)
Temporary storage for intermediate results	4000 TiB
Node requirements storage and analytics computations	25 Nodes à 96 cores & 256 GiB RAM
Node requirements for serving the base-index	70 Nodes à 48 cores & 256 GiB RAM

structure, which enable fine-tuning the search rankings to user requirements. Lewandowski (2019) introduced the idea of openly creating such an index, separated from the actual search services that are potentially maintained by different legal entities. The indexing infrastructure provides an API for search engine providers and manages usage data. Consequently, such an index would enable the development of several search engines with substantially reduced upfront costs.

We build on the Open Web Index (OWI) concept from Lewandowski (2019), but argue that solely providing an API to an OWI may limit the usage of that index. For example, different search engines might require quite different ranking mechanisms, leading to large development and service costs. Furthermore, individual users or entities may only need to search a portion of the Web rather than the entire WWW, which further adds to the search costs. The significant players already provide APIs for their index, and hence an OWI would need to compete regarding costs for API call, coverage, quality, and so forth.

To advance the innovative idea of an OWI, we propose six principles for creating an OWI and present a brief conceptual overview in Figure 1:

- Distributed as open data: Instead of solely providing an API over an openly created web index, we propose to view the index as open data, that is, an openly created data structure available under some open (use) licenses. Consequently, the OWI can be sliced and diced as needed by a potential search engine provider, or be used for completely other means, including, for example, the data for training neural language models. A proper public licensing model can allow third-party contributions to the index or its parts, opening up the road for collaborative index creation.
- Open and extensible index creation: Index creation should be transparent and extensible. All pipelines used—from crawling to preprocessing to indexing—have to be open source, and their configuration needs to be exposed openly. Extensibility empowers third parties to contribute (algorithmic) components to the index creation pipelines, containing up-to-date semantic enrichment models, and providing researchers and innovators with the opportunity to explore their methods on a large scale.
- Collaborative creation: Ideally, an OWI would be created collaboratively based on a Wikipedia principle, but more focused on domain experts as in the Nupedia (Sanger, 2005). Adding domain knowledge to the index should significantly increase its quality. However, different to the case of Wikipedia, the underlying process is technically complex, which might require more technically oriented intermediaries (e.g., digital libraries, research computing centers, etc.). However, if the pipelines for index creation can be decentralized among independent computing centers, the costs for index creation could be also significantly reduced, yielding a cost-efficient, high-quality index.
- Tracking content usage, not the user: An OWI must not collect data about individual users, even if this would be an important data source for optimizing search and retrieval processes. Assuming that the OWI is used in a lot of different search engines, one could collect aggregated information about the content usage in those engines, instead of collecting information from individual users. It would be up to the search engines to collect and aggregate click data for user groups. Such aggregated, anonymized usage data could be managed in addition to the OWI as a by-product, but not necessarily fully integrated into an index.
- Control to the content owners: Content owners should be empowered to control the usage of their content in an OWI, on a more fine-grained level than is possible using current approaches like the *robots.txt* standard. This includes provision of legal information, like for example machine-readable content licensing, or, on the other side, compliance with jurisdictional requirements like GDPR. Similarly, through principle 4, web content owners will be informed on usage details of their content opening up opportunities for new business models.
- Legal compliance for content users: Due to different legal frameworks in different countries, legal uncertainties when crawling and preprocessing web data remain high, for example, regarding intellectual property and licensing rights. The current gatekeepers hold a unique position such that content owners have to waive the rights to use their content, for the possibility to be found. Providing an OWI needs processes that consider different legal frameworks, ensure legal usage of content and the exclusion of illegal content (see, e.g., Erenli et al., 2021).



**FIGURE 1** Illustration of key concept of an OWI as basis for an open, extensible, transparent, and legally sound web search ecosystem. Colors depict different parts of the Web and contributions from different third parties in building a web index

We argue that these properties will allow the creation of an open, extensible, transparent and legally sound web search ecosystem which would yield to new business models and empowers end-users with different models of web search. It separates the role of the web index creator and curator from the role of the search engine provider, which reduces costs and means of governance when utilizing web data.

## 4 | BUILDING AND DISTRIBUTING THE OPEN WEB INDEX

The creation of the OWI follows the traditional pipeline of crawling, enrichment, and indexing, but over decentralized computing centers. In this section, we will discuss critical aspects of crawling and semantic enrichment to fulfill the outlined principles. Central to using the OWI as Open Data is the concept of Declarative Search Engines, which enables the use of partial indices, including means for index merging and splitting, for easy creation and deployment of search engines.

### 4.1 | Coordinated and legally compliant crawling

Web crawling is the process of navigating the graph structure of the Web for discovering and fetching Web data. It is the predominant method for web search engines to gather content for their index.

While webmasters and content owners have some control over the crawling process via de facto standards like *robots.txt* (Koster et al., 2022) and *sitemaps*,<sup>2</sup> further control and steering mechanisms remain purely

proprietary in the hands of gatekeepers like Google's Search Console. At first glance, these services appear to give webmasters control over the usage of their content. However, a deeper look reveals that they basically outsource complex technical challenges to webmasters. The data collected is not opened up for other crawlers to use, thereby creating a “vendor lock-in” as webmasters will want to optimize their web pages only for the most important search engine. Additionally, webmasters do not have full control over usage rights, licenses, or data protection information—an aspect particularly important when considering the use in AI tools, which might not be in the users interest. This demonstrates that webmaster services are oriented towards optimizing the particular search engine (and the generated revenue), rather than serving as a general support tool for improving web crawling or enabling full legal control.

An open crawling pipeline for an OWI should address these issues by pursuing two major goals. First, reduce independent crawling efforts by either opening up own crawls in the form of WARC files (Mohr et al., 2008), comparable to the CommonCrawl data, or by coordinating ongoing independent crawling efforts. Second, open up webmaster data, per website, such that content owners and managers can express legal constraints on how their content is used and also track content usage across the different steps of a search engine. With the advancement of large language models, giving the control back to the content owner, without limiting content dissemination, becomes even more important.

### 4.2 | Preprocessing and semantic enrichment

Many analysis tasks attempt to segment web pages and remove boilerplate elements, leaving only the remaining

“main content” for further analysis; yet, this depends on the application. The final step usually involves the semantic analysis of the main content and the extraction of semantic data. Examples are the extraction of geo-references and named entities, and entity linking, the task of identifying those entities in an already existing knowledge graph.

Three core requirements would be essential when building up preprocessing and semantic enrichment pipelines (see, e.g., Wachsmuth, 2015 for details on such pipelines): First is efficiency in order to be able to process data on a petabyte scale and to keep up with the ever changing nature of the WWW. Second, basic enrichment services of web data, like for example metadata extraction through microformat parsing (Khare & Çelik, 2006) or named entity extraction/linking (van Hulst et al., 2020) or web-genre classification (Lex et al., 2010) for partitioning the index. Third, extensibility for advanced, special purpose semantic enrichment techniques on demand like for example annotating information quality via “information nutrition labels” (Fuhr et al., 2017) or for detecting hate speech in web pages (Caselli et al., 2020). While efficiency and basic enrichment deliver the core functionality to deliver state-of-the-art search at scale, extensibility will be essential to give researchers and innovators room to develop new concepts, like for example conversational search through large language models.

### 4.3 | Declarative search engines

Since the OWI is intended to accommodate a variety of search engines, it must also be easy for search engine designers to use (parts of) the index for their own purposes. To this end, we propose the OWI to become a (distributed) information system similar to the well-known Docker hub. Instead of virtual machines, though, the OWI would contain prebuilt indexes that would be readily usable. To define downstream search engines using the OWI, we argue to introduce an Open Web Search Engine Hub (OWSE-Hub). Similar to the OWI, the OWSE-Hub forms a web-based information system comparable to the Docker hub, but it will contain complete search engine stacks to enable the fast and easy creation of new search verticals.

A potential architecture of the envisioned OWSE-Hub is sketched in Figure 2. Using the OWSE-Hub, users can declaratively define their own search configurations (similar to academic work like Cornacchia & de Vries, 2006 or Kamphuis & de Vries, 2021), or the declarative formulation of search engines in PyTerrier (Macdonald et al., 2021). Users can “pull” predefined specifications from the OWSE-Hub, use those to “build” their own

custom search engines, and “push” the most useful ones to share these with others. This flexible setup allows for the creation of a wide variety of search engines, not only for commercial usage but also for personal and corporate search, and allowing both centralized and federated search setups.

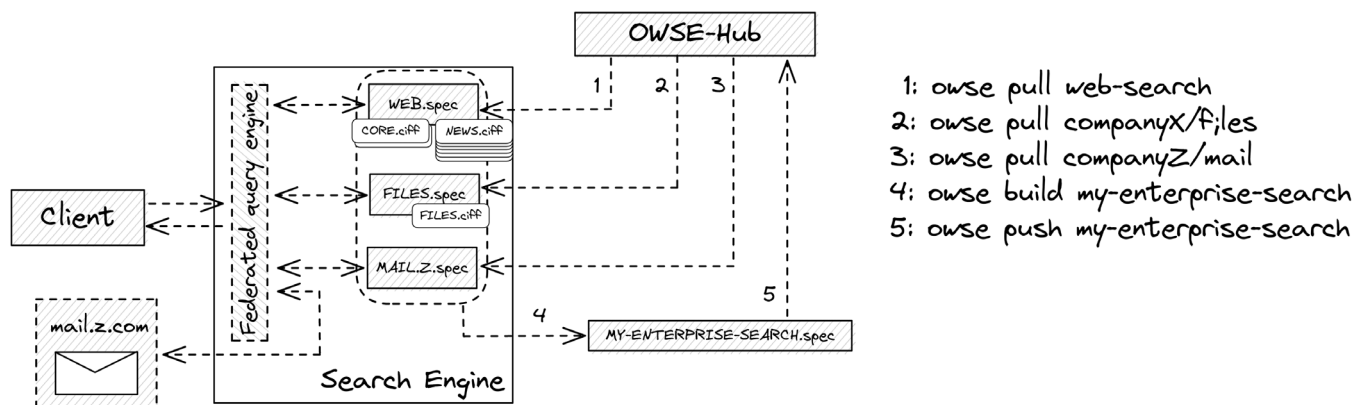
## 5 | IMPACT AND APPLICATIONS OF AN OPEN WEB INDEX

The purpose of an Open Web Index (OWI) is not to compete with dominant search engines like Google, but to provide the foundation for a competitive search engine ecosystem. This open ecosystem would involve players from various sectors and fields, ensuring diverse access to information and providing users with more choice. We believe that an OWI could have a similarly positive impact on the search engine industry as open source and open hardware have had on software and hardware development (Blind et al., 2021).

### 5.1 | Search verticals

Contrary to general purpose search engines like Google, vertical search engines serve specific domains or purposes and offer opportunities for optimized search and retrieval strategies. Current popular vertical search solutions are mostly commercially focused or integrated into enterprises' business models, such as Amazon's product search, LinkedIn's people search, or Booking.com's hotel search. An OWI could provide additional web content to these vertical search engines or even become the primary source for such engines, even in niche markets.

In addition to topical verticals, search engines could also focus on different vertical incentives such as transparency and explainability in web search, user control over the ranking process, privacy, legally compliant web search, or new AI-based retrieval models. The main impact of an OWI is to create an ecosystem that gives users true choice, from individuals to companies interested in search. By reducing the investment costs in creating a web index, an OWI could enable web-scale search verticals, providing true alternatives. An increase in vertical search engines could lead to a shift from a one-search-engine-to-all-users relationship to a many-search-engines-to-many-users relationship, where search engines would attract users through the selection of relevant content and service optimization. Individual user preferences would be less relevant, as they would already be expressed by choosing a particular search engine. While it is difficult to quantify the exact impact, a many-



**FIGURE 2** General architecture of the OWSE-Hub. The OWSE-Hub contains specifications for declarative search engines. Users can (1–3) pull search engine stacks, (4) build their own specifications for a (composite) search engine, and (5) push specifications to share with others

to-many ecosystem appears to hold greater potential for a fair and high-quality information space than a simple one-to-many ecosystem. While such an information space may not be completely unbiased, users can at least consciously choose their preferred bias.

## 5.2 | Innovative search paradigms and web-data centered applications

Beyond search verticals and federated search, an OWI would enable the exploration of new search paradigms and interfaces on a large scale. Developing new paradigms and corresponding user interfaces on a large scale and with recent, up-to-date data allows for targeting real users and generating feedback beyond single research prototypes. Search paradigms could include conversational search (Anand et al., 2020), (temporal) argumentation search (Potthast et al., 2019), and human-centric search (search over private and public data collections). Similarly, new user interfaces could explore more data-centric visualization paradigms (Höfler et al., 2014) or new query navigation techniques (Seifert et al., 2017).

Successful paradigms and user interfaces based on an OWI would provide users with choices on how to obtain information on the web beyond lists of web pages. This includes neural language models (Bengio, 2008) and knowledge graphs (Hogan et al., 2022). Large language models such as GPT-4 (OpenAI, 2023) require web data at scale and in quality, making an OWI an essential resource for training these models and applying corresponding conversational search engines. Without an OWI, the gap between a few gatekeepers and the rest will widen.

Overall, application scenarios and possibilities for an OWI are plentiful. The OWI can open up an

ecosystem involving stakeholders from different sectors (e.g., industry, science, NGOs, policy makers, and associations like GAIA-X or the EOSC) as well as interdisciplinary stakeholders with very different expertise, for example, technical, ethical, and legal expertise. It will lead to interesting new opportunities by opening up the black-box currently under control of a few commercial entities.

## 6 | CONCLUSION

We believe that the web search engine ecosystem needs to become more open and diverse in order to offer users real choice for free and transparent information access. As we have argued, an Open Web Index could be the foundation for such an opening, but needs to be developed as open data product rather than hidden behind an API. This allows to decentralize index creation and separate it from the search service itself. The former requires an indexing pipeline that is distributed over independent computing centers serving their own clientele, while the latter requires—from our point of view—a declarative approach to search engines and a way for distributing index updates instead of a central search service.

Considering an OWI as Open Data enables cost-reduction and collaboration by separating index creation from index provision. Index creation costs can be reduced through cooperation between independent organizations, as outlined in Principles 1 to 3. This may benefit special purpose organizations such as libraries and archives, which can reduce costs while still providing high-quality collections, as well as computing centers at research organizations. Furthermore, the Declarative Search Engine approach allows for outsourcing index serving costs, separating index creation from serving the index. This can

result in close-to-zero costs for serving a small index, with moderate costs for basic search facilities while still enabling a full blown web search engine. Paired with custom-tailored, domain specific language models this could yield powerful new search capabilities.

The proposed concept will be prototyped in the recently started, EU-funded research project OpenWebSearch.eu. We hope to deliver a first stepping stone towards an Open Web Index and its associated effects.

## ACKNOWLEDGMENTS

This work is part of the OpenWebSearch.eu project. The OpenWebSearch.eu Project is funded by the EU under the GA 101070014 and we thank the EU for their support. We also want to thank the anonymous reviewers for their valuable feedback and comments.

## ORCID

Michael Granitzer  <https://orcid.org/0000-0003-3566-5507>

Stefan Voigt  <https://orcid.org/0000-0002-5908-331X>

Noor Afshan Fathima  <https://orcid.org/0009-0008-9707-6453>

Christian Guetl  <https://orcid.org/0000-0001-9589-1966>

Tobias Hecking  <https://orcid.org/0000-0003-0833-7989>

Gijs Hendriksen  <https://orcid.org/0000-0003-0945-3148>

Djoerd Hiemstra  <https://orcid.org/0000-0003-4967-2900>

Jelena Mitrović  <https://orcid.org/0000-0003-3220-8749>

Alexander Nussbaumer  <https://orcid.org/0000-0002-4692-5741>

Per Öster  <https://orcid.org/0000-0001-5836-8850>

Martin Potthast  <https://orcid.org/0000-0003-2451-0665>

Benno Stein  <https://orcid.org/0000-0001-9033-2217>

Arjen P. de Vries  <https://orcid.org/0000-0002-2888-4202>

Saber Zerhoubi  <https://orcid.org/0000-0003-2259-0462>

## ENDNOTES

<sup>1</sup> <https://www.worldwidewebsite.com/>

<sup>2</sup> <https://www.sitemaps.org/protocol.html>, last accessed November 25, 2022.

## REFERENCES

- Anand, A., Cavedon, L., Hagen, M., Joho, H., Sanderson, M., & Stein, B. (2020). Conversational search—A report from dagstuhl seminar 19461. arXiv:2005.08658.
- Baeza-Yates, R., & Maarek, Y. (2012). Usage data in web search: Benefits and limitations. In L. Calderón-Benavides, C. N. González-Caro, E. Chávez, & N. Ziviani (Eds.), *String processing and information retrieval—19th international symposium, SPIRE 2012, Cartagena de Indias, Colombia, October 21–25, 2012* (p. 16). Springer.
- Barker, R. (2018). Trapped in the filter bubble? Exploring the influence of google search on the creative process. *Journal of Interactive Advertising*, 18(2), 85–95.
- Bengio, Y. (2008). Neural net language models. *Scholarpedia*, 3(1), 3881. <https://doi.org/10.4249/scholarpedia.3881>
- Bevendorff, J., Stein, B., Hagen, M., & Potthast, M. (2018). Elastic chatnoir: Search engine for the clueweb and the common crawl. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval—40th European conference on IR research, ECIR 2018, Grenoble, France, March 26–29, 2018* (pp. 820–824). Springer.
- Blind, K., Böhm, M., Grzegorzewska, P., Katz, A., Muto, S., Pättsch, S., & Schubert, T. (2021). *The impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy* (Final study report) (Vol. 10). European Commission.
- Caselli, T., Basile, V., Mitrovic, J., & Granitzer, M. (2020). Hatebert: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th workshop on online abuse and harms (WOAH 2021)*. Association for Computational Linguistics.
- Cliqz GmbH. (2019). *A new search engine: Cliqz journey*. Author. Retrieved from <https://web.archive.org/web/20191205214556/> and <https://0x65.dev/blog/2019-12-05/a-new-search-engine.html>
- Cornacchia, R., & de Vries, A. P. (2006). A declarative DB-powered approach to IR. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in information retrieval, 28th European conference on IR research, ECIR 2006, London, UK, April 10–12, 2006* (pp. 543–547). Springer.
- Das, A., & Jain, A. (2012). Indexing the world wide web: The journey so far. In *Next generation search engines: Advanced models for information retrieval* (pp. 1–28). IGI Global.
- Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017). Suppressing the search engine manipulation effect (SEME). In *Proceedings of the ACM on human-computer interaction* (Vol. 1, pp. 1–22). Association for Computing Machinery.
- Erenli, K., Geminn, C., & Pfeiffer, L. (2021). Legal challenges of an open web index. *International Cybersecurity Law Review*, 2(1), 183–194. <https://doi.org/10.1365/s43439-021-00017-8>
- Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Järvelin, K., Jones, R., Liu, Y., Mothe, J., Nejdil, W., Peters, I., & Stein, B. (2017). An information nutritional label for online documents. *SIGIR Forum*, 51(3), 46–66. <https://doi.org/10.1145/3190580.3190588>
- Hagiu, A., & Wright, J. (2020). When data creates competitive advantage. *Harvard Business Review*, 98(1), 94–101.
- Herrmann, M., Zhang, R., Ning, K.-C., Diaz, C., & Preneel, B. (2014). Censorship-resistant and privacy-preserving distributed web search. In *14th IEEE international conference on peer-to-peer computing* (pp. 1–10). IEEE Press. <https://doi.org/10.1109/P2P.2014.6934312>
- Höfler, P., Granitzer, M., Veas, E. E., & Seifert, C. (2014). Linked data query wizard: A novel interface for accessing SPARQL endpoints. In C. Bizer, T. Heath, S. Auer, & T. Berners-Lee (Eds.), *Proceedings of the workshop on linked data on the web co-located with the 23rd international world wide web conference (WWW 2014), Seoul Korea, April 8, 2014*. CEUR-WS.org. Retrieved from [https://ceur-ws.org/Vol-1184/ldow2014%5C\\_paper%5C\\_06.pdf](https://ceur-ws.org/Vol-1184/ldow2014%5C_paper%5C_06.pdf)
- Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J. F., Staab, S., & Zimmermann, A. (2022). Knowledge graphs. In *ACM*



- computing surveys* (Vol. 54, pp. 1–37). Association for Computing Machinery. <https://doi.org/10.1145/3447772>
- Kamphuis, C., & de Vries, A. P. (2021). GeeseDB: A python graph engine for exploration and search. In O. Alonso, S. Marchesin, M. Najork, & G. Silvello (Eds.), *Proceedings of the second international conference on design of experimental search & information retrieval systems, Padova, Italy, September 15–18, 2021* (pp. 10–18). CEUR-WS.org. Retrieved from <https://ceur-ws.org/Vol-2950/paper-11.pdf>
- Khare, R., & Çelik, T. (2006). Microformats: A pragmatic path to the semantic web. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, & M. Dahlin (Eds.), *Proceedings of the 15th international conference on world wide web, WWW 2006, Edinburgh, Scotland, UK, May 23–26, 2006* (pp. 865–866). Association for Computing Machinery. <https://doi.org/10.1145/1135777.1135917>
- Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022). *Robots exclusion protocol RFC 9309* (pp. 1–12). IETF. <https://doi.org/10.17487/RFC9309>
- Lewandowski, D. (2015). Living in a world of biased search engines. *Online Information Review*, 39, 3. <https://doi.org/10.1108/OIR-03-2015-0089>
- Lewandowski, D. (2019). The web is missing an essential part of infrastructure: An open web index. *Communications of the ACM*, 62(4), 24. <https://doi.org/10.1145/3312479>
- Lex, E., Juffinger, A., & Granitzer, M. (2010). A comparison of stylistic and lexical features for web genre classification and emotion classification in blogs. In A. M. Tjoa & R. R. Wagner (Eds.), *Database and expert systems applications, dexa, international workshops, Bilbao, Spain, August 30–September 3, 2010* (pp. 10–14). IEEE Computer Society. <https://doi.org/10.1109/DEXA.2010.24>
- LibreTechTips. (2020). *Detailed tests of search engines: Google, Startpage, Bing, DuckDuckGo, metaGer, Ecosia, Swisscows, Searx, Qwant, Yandex, and Mojeek*. Author. Retrieved from <https://libretechtips.gitlab.io/detailed-tests-of-search-engines-google-startpagebing-duckduckgo-metager-ecosia-swisscows-searx-quant-yandex-and-mojeek/>
- Macdonald, C., Tonello, N., MacAvaney, S., & Ounis, I. (2021). Pyterrier: Declarative experimentation in python from BM25 to dense retrieval. In G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, & H. Tong (Eds.), *CIKM '21: The 30th ACM international conference on information and knowledge management, virtual event, Queensland, Australia, November 1–5, 2021* (pp. 4526–4533). Association for Computing Machinery. <https://doi.org/10.1145/3459637.3482013>
- Mohr, G., Kunze, J., & Stack, M. (2008). *The WARC file format 1.0 (ISO 28500)*. eScholarship.org. Retrieved from <https://escholarship.org/content/qt9nh616wd/qt9nh616wd.pdf>
- OpenAI. (2023). Gpt-4 technical report. arXiv:2303.08774.
- Potthast, M., Gienapp, L., Euchner, F., Heilenkötter, N., Weidmann, N., Wachsmuth, H., Stein, B., & Hagen, M. (2019). Argument search: Assessing argument relevance. In B. Piwo-warski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, & F. Scholer (Eds.), *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2019, Paris, France, July 21–25, 2019* (pp. 1117–1120). Association for Computing Machinery. <https://doi.org/10.1145/3331184.3331327>
- Qwant SAS. (2019). *How Microsoft tools strengthen Qwant*. Author. Retrieved from <https://betterweb.qwant.com/en/how-microsoft-tools-strengthen-quant/>
- Sanger, L. (2005). The early history of nupedia and wikipedia: A memoir. *Open Sources*, 2, 307–338.
- Seifert, C., Schlötterer, J., & Granitzer, M. (2017). Querycrumbs: A compact visualization for navigating the search query history. In *21st international conference information visualisation, IV 2017, London, UK, July 11–14, 2017* (pp. 35–44). IEEE Press. <https://doi.org/10.1109/IV.2017.23>
- van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., & de Vries, A. P. (2020). REL: An entity linker standing on the shoulders of giants. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, & Y. Liu (Eds.), *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2020, virtual event, China, July 25–30, 2020* (pp. 2197–2200). Association for Computing Machinery. <https://doi.org/10.1145/3397271.3401416>
- Wachsmuth, H. (2015). *Text analysis pipelines—Towards ad-hoc large-scale text mining*. Springer. <https://doi.org/10.1007/978-3-319-25741-9>

**How to cite this article:** Granitzer, M., Voigt, S., Fathima, N. A., Golasowski, M., Guetl, C., Hecking, T., Hendriksen, G., Hiemstra, D., Martinovič, J., Mitrović, J., Mlakar, I., Moiras, S., Nussbaumer, A., Öster, P., Potthast, M., Srdič, M. S., Megi, S., Slaninová, K., Stein, B., ... Zerhoubi, S. (2023). Impact and development of an Open Web Index for open web search. *Journal of the Association for Information Science and Technology*, 1–9. <https://doi.org/10.1002/asi.24818>