Proposal for an Open Web Index

Nikolaus Huss, KovarHuss GmbH Policy Advisors, Berlin
huss@suma-ev.de

Prof. Dr. Dirk Lewandowski, Hamburg University of Applied Sciences, Hamburg
dirk.lewandowski@haw-hamburg.de

Dr. Wolfgang Sander-Beuermann, SUMA-EV, Hannover
wsb@suma-ev.de

Albrecht Ude, journalist, Berlin
Albrecht@ude.de

Submitted on behalf
SUMA-EV - Verein für freien Wissenszugang
Röselerstr. 3
30159 Hannover
Germany
office@suma-ev.de

Contents

Summary

This paper describes the Open Web Index, a European-based approach for a competitive and data-protecting digital infrastructure. It aims to build a basis for genuine competition in the digital platform business. In that, the Open Web Index is a strategic approach to reduce the predominance of foreign Web indexes and search engines. The main Idea of Open Web Index is to set up a publicly funded, global, searchable index of the Web that is open to competing companies, institutions and civil society actors.

Problem statement

1. The Web would be useless without means for searching its contents.
2. As there is no central directory of the Web, private search engine companies have built large indexes of its contents.
3. There are only a few Web-scale indexes, operated by Google (U.S.), Microsoft (U.S.), Baidu (China) and Yandex (Russia).
4. Search engines are built upon the indexes. Indexes are, however, not limited to search engines, but are needed for many other services.
5. Companies operating Web-scale indexes do not allow sufficient access to their data to other parties interested.
6. The difficulties in building a Web index lie in technical issues, operating costs, Web size, and freshness.
7. Due to these difficulties, there is no Web index built by a European company (or other entity).

Proposed solution: An Open Web Index

1. The vision of the Open Web Index is to build a public library of the Web.
2. The primary technical idea is to separate the index from the services that are built on the index.
3. European companies and institutions would largely benefit from an Open Web Index, as they could build their applications on top of it.
4. Given a considerable uptake on the market, the Open Web Index would foster plurality.
5. Users would also benefit from better transparency, as the index would be open to everyone.
6. Building an Open Web Index would help with the safeguarding of the critical infrastructure, restoration of the informational sovereignty of Europe, stimulation of competition in Internet search, stimulation of the European start-up and internet economy, development of business models beyond the concept of “data for information”.
7. The Open Web Index will allow Artificial Intelligence (AI) to fulfil its full potential, as it will enable combining Web data with proprietary data.
8. The Open Web Index will help to build a multitude and great variety of Internet services, including but not limited to, search engines, maps and routing services, price comparison services, and trend analysis.

Implementation

1. Implementing the Open Web Index can only be achieved on the European level with appropriate funding.
2. For building and running the Open Web Index, a new institution in the form of a foundation should be built.
3. The European Open Web Index foundation could work together with a network of research in Europe. This network could be used to develop the structure of the Open Web Index further, as well as providing the companies and institutions that access the web index with qualified partners from science.

1 Basics and problem statement

In this section, we explain the basic concept of indexing, and the differences between an index and a search engine are pointed out. We describe the most critical obstacles to building Web indexes and give an overview of existing Web indexes. This section concludes with explaining why current indexes do not provide a solution for making the Web's contents accessible.

1.1 What is a web index?

An “index” is first and foremost a tool (a reference, an ordered directory) to find something. Everyone knows it from the last pages of non-fiction books. There, words from a book’s text are listed with page numbers: those pages are where these words /or the topics occur. In this respect, an index contains "key" words. It allows quick and easy access to the desired content.

An index can also be much more than only word-related. It may also contain information such as,

- Where which images appear,
- Where what images or graphics on which topics occur,
- Which words occur in headings (so what words might be particularly relevant),
- What other books or articles are cited,
- … And much more.

The creator of the index is free in deciding which information is included in the index to facilitate the retrieval of content in the book. The thicker the book, the more critical the index and its quality become.

The thickest “book” that has ever been written by humanity is the World Wide Web, the largest part of the Internet nowadays. Without a Web index, one could find hardly anything on the Internet. Therefore, Web indexes are of central importance: Without them, the Web would practically not be usable.

1.2 What is the difference between an index and a search engine?

Search engines do with the Web index what users do with the index of a book: They see to what words (or pictures, topics, etc.) fit to which web pages. When a user enters a query, the search engine searches its index for relevant documents.

In theory, many search engines could be built upon a single index. These search engines could still provide very different results; as a result set is always influenced by the index (i.e., what is available to the search engine), as well as the ranking algorithms (i.e., how the contents of the index are ordered in response to a user’s query).

1.3 Why is it so difficult to create a web index?

There are four significant obstacles when building a Web-scale index:

(1) Technical issues: Index providers face substantial technical difficulties due to the large numbers of documents resulting from the ever-changing nature of the Web.

(2) Operating costs: To maintain a Web index, a cluster of thousands of distributed servers is needed. This results in significant costs of hardware, infrastructure, maintenance and staff.

(3) Web size: The Web is vast, and an index needs to be tasked with covering as large a part of the Web as possible. Modern indexes (like Google's) know of trillions of existing pages.[1]
(4) Freshness: A search engine needs to keep its index current, meaning it needs to update at least a part of it every minute. This is a huge task, considering the scale of a Web index.

1.4 What Web indexes already exist on the market?

Worldwide, there are only four Web indexes that are competitive regarding (1) size and (2) freshness:

(1) Google (USA)
(2) Bing (USA)
(3) Baidu (China)
(4) Yandex (Russia)

These Web indices are all privately owned. And, they are all located outside of Europe. Europe's digital economy and civil society are virtually dependent on non-European businesses. This is particularly evident concerning search engines, the cornerstone of our digital information infrastructure. If we were to apply the present situation in the digital world to the mass media, we would find ourselves with only four foreign television channels as the sole sources of information for the public. Businesses would also be dependent on these channels, as they would be the only available outlets for their advertising. Such a situation contradicts the pluralism of our Western democratic societies. Pluralism must also be reflected in a diversity of information systems.

The four relevant indexes feed all major search portals (with, e.g., Yahoo showing results from Bing).

Smaller indexes (e.g., Duck Duck Go; Qwant) exist but lack the scale to be competitive on the market. A smaller index may be suitable for many search tasks but will not be sufficient to compete in the worldwide search engine market or serve as a basis for an analytical tool in need of Web data.

Open index initiatives (e.g., Common Crawl) exist but focus on static snapshots of Web data that do not fulfil the two requirements of size and freshness stated above.

1.5 Why is the problem not solved having four commercial Web indexes?

The companies operating the four Web indexes either do not allow access to third parties at all or let its commercial partners access only a very limited subset of *results* from the index. The most significant differentiation here is access to the index vs. access to some results from the index. Google and Bing allow its partners (e.g., major search portals) access to some top results in response to a query. They are then allowed to present these results in the predetermined order. They are, however, not allowed to show the results in a different order or use the results for other purposes.
When a third party accesses results from one of these index providers, the data for each result is insufficient, containing the URL, a heading and a short description. This is mainly what a user sees when examining a search engine results page. For building applications on top of a Web index, however, companies would need access to the rich representations of Web documents that form the strength of such an index.

---

[1] Schwartz, B. Google's search knows about over 130 trillion pages. *Search Engine Land*, 2016. http://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378.

[2] Goel, S., Broder, A., Gabrilovich, E., and Pang, B. Anatomy of the long tail: Ordinary people with

2 Proposed solution: An Open Web Index

In this section, we show how an Open Web Index (OWI) can overcome the problems described in section 1 and how it can be used as the data source for a multitude of services to the benefit of commercial companies and non-commercial entities, alike.

2.1 Vision: A public library of the Web

The vision of the Open Web Index can be compared with a public library that contains all published knowledge. For instance, in Europe's national libraries, all texts published in the respective countries are collected, registered in a directory and made available to the public.

The Open Web Index is organised similarly. The index collects all content of relevant websites and makes it available to all parties interested. The OWI makes all content openly accessible, comparable with a public library. Search engine operators, database providers, and other digital service providers, businesses and civil society organisations can access its contents. The Open Web Index is a public infrastructure that can be used flexibly in a changing world.

A variety of different services can be built upon a web index and generate added value. The most common examples of services based on a web index are search engines. However, many other services, whether imaginable today or not, can be built upon the index (see details in section 2.5).

2.2 Basic technical idea: Separating index and services

We are proposing a plan for a missing part of the Web's infrastructure, namely a searchable index. The idea is to separate the infrastructure part of the search engine (the index) from the services part, thereby allowing for a multitude of services, whether existing as search engines or otherwise, to be run on a shared infrastructure. Figure 1 shows how the public infrastructure is responsible for crawling the web, for indexing its content, and for providing an interface and application programming interface (API) to the services that are built upon the index. While services are allowed to do their own further indexing to prepare documents, some advanced indexing is also provided by the open infrastructure. Furthermore, as modern search engines rely heavily on usage data, these data (most prominently search queries routed to the index) are collected and made available for reuse.

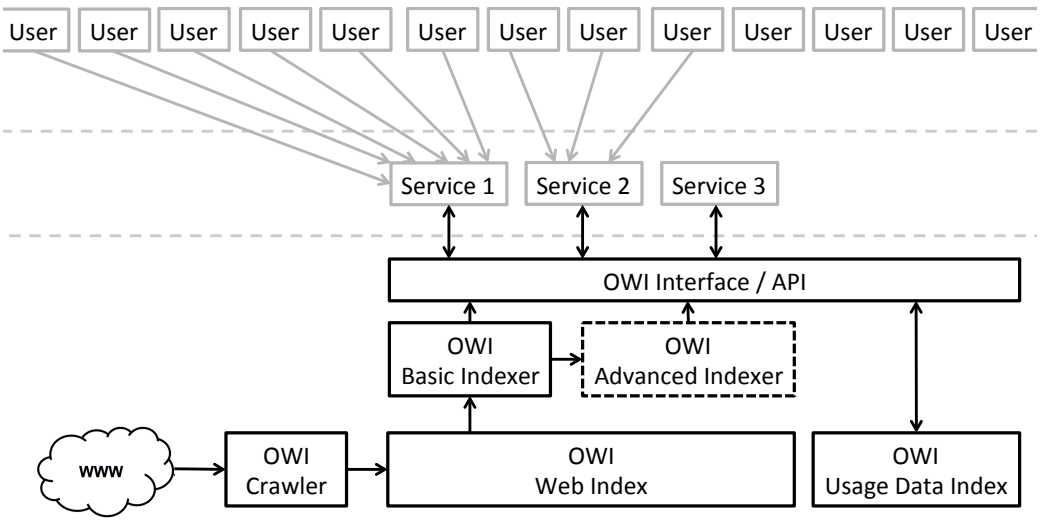

Fig. 1: Separating services from infrastructure

2.3 Benefits to European companies and institutions

An Open Web Index creates a new and open digital infrastructure for a variety of competing services to build on. On the one hand, these are familiar services like search engines and related developments (meta trading platforms, aggregation services, trend analyses, etc.). On the other hand, these can also be entirely new services whose properties and added value we are not yet able to see. It is a feature of almost any new infrastructure that it is impossible to predict what will be developed on it – an infrastructure that is free for use, however, creates the potential for disruptive, innovative leaps. When the infrastructure "Internet" was built, no one thought of Google or Facebook.

The main benefit of an Open Web Index would be for all interested parties to be able to develop their own applications without the problem of having to create their own Web index first. Building an index of considerable size is currently an impossible endeavour not only, but especially, for small and medium enterprises, as well as for non-commercial bodies. Unlike the early years of the Web, the present volume of data and growing complexity of the Internet means that even major corporations and organisations do not have the financial resources to establish such an index.

Given a considerable uptake for such an index, it would foster plurality not only in the use of Web content by developers but also in the variety of content that users get to see. In the current market situation, we are far from plurality, not only regarding the number of search engine providers but also the number of search results. In 2010, a study from Yahoo showed that while we can consider a search engine as a possible window to all of the web's content, more than 80% of all user clicks were found to go to only 10,000 different domains.[2] We can assume that these numbers are comparable for other search engines.
Given an Open Web Index, we can rightly assume that each search engine using the index would apply its own ranking function, and therefore, produce different results. Users would benefit in that they would not have to rely on only one or at best a few search engines but could choose from a variety of engines serving their different purposes. In that way, an Open Web Index would foster plurality and restrict the power of single companies dictating which content is shown to and consumed by users.

Another benefit would be that the index would be open to everyone, and therefore, would allow for investigating its transparency. However, search engines built on top of the index could still be “black boxes” in that they would not need to make their ranking functions open to anybody. Users of an Open Web Index will have unrestricted access to the stored information and are no longer at the mercy of a global corporation’s goodwill. Users of this index could be search engine operators who develop different types and forms of search engines with it. These search engines can be very diverse – different interfaces, different designs, different topics, different business models – and they will compete with each other.

However, there can also be entirely different applications than search engines. Services could find “data treasures” in the Web content, digging up this “gold” of the Internet with the Open Web Index, and providing it to humanity. Thus, for example, new insights on content correlations could arise through analysis of link structures and their clustering. Humanity's basic problems of knowledge could appear in a new light, for example, "In which categories does human thinking occur?". However, also very concrete, practical questions such as different pricing for the same product in markets worldwide could be investigated using the index. A whole new variety of entirely new tools will arise.

---

[2] Goel, S., Broder, A., Gabrilovich, E., and Pang, B. Anatomy of the long tail: Ordinary people with extraordinary tastes. *Proceedings of the third ACM international conference on Web search and data mining*, ACM (2010), 201–210.

2.4 How is an Open Web Index related to Europe's goals?

Building an Open Web Index will help with, but is not limited to, the following:

- Safeguarding of the critical infrastructure
- Restoration of the informational sovereignty of Europe
- Stimulation of competition in internet search
- Stimulation of the European start-up and internet economy
- Development of business models beyond the concept of "data for information"

2.4.1 Safeguarding Europe's critical infrastructure

The Open Web Index is part of a crucial infrastructure: Only content that is found on the Internet is available in emergencies. The difficult negotiations between the European Commission and Google demonstrate that it is time for Europe to have an independent digital infrastructure even under challenging circumstances. The following principle applies: In a situation of crisis the only infrastructure which is secure is the one which is subject to one's own jurisdiction.

2.4.2 Restoration of the informational sovereignty of Europe

Discussions about civil rights, informational self-determination, data protection and the right to privacy in digital society differ in various cultural areas. However, despite all differences between European countries, the European identity differs considerably from the Chinese, Russian, and even from the American way of handling data protection and privacy. The establishment of civil rights on the Internet depends on whether the legal view can be enforced. It requires the existence of a fundamental digital infrastructure. The concept of the Open Web Index corresponds with the features typical of the Internet-based "platform economy". It provides a digital infrastructure service within the European jurisdiction but leaves it to businesses to compete for the best solution and enter into permanent competition for innovation.
Informational sovereignty can only be achieved from a powerful network infrastructure.

2.4.3 Stimulation of competition in Internet search

The use of only one single search engine narrows the concepts of searching and retrieving information on the Internet. Each presentation of search results is an (algorithmic) interpretation of web content – while many interpretations are possible, the use of just one search engine narrows the opportunities to one single interpretation. The Open Web Index helps to avoid the danger of a so-called filter bubble.
Search applications for the Web can only be built on the basis of a comprehensive and regularly updated web index. New concepts like the semantic indexing of the entire Web and "data-saving concepts" for search or approaches for the disclosure of the identity-driven by users on the Internet require an openly designed network infrastructure; the core is an open system for the indexing of knowledge available on the Web.

2.4.4 Stimulation of the European start-up and Internet economy

The restriction of access rights to currently available search engine APIs is increasingly turning out to be a significant obstacle for an open and diverse network society (in economy, in research, as well as in civil society). Open and equal access to the interfaces of the Open Web Index enables the implementation of new and promising concepts beyond the narrowed view of current Internet economy with increasingly rigid oligopolies.

2.4.5 Development of business models beyond the concept of "data for information"

Popular services like Google and Facebook are not free to their users; instead, they apply the concept of "data for information". Data represent the collection of information about the preferences and behaviour of users and the exploitation of data in the interest of the company. The information imbalance between a few globally operating Internet companies and citizens who have been reduced to being Internet users for a long time now is reduced.

The development of alternative services that are not based on the “voluntary compulsory” disclosure of information is long overdue in a democratic society – and requires the appropriate information basis.

2.5 The Open Web Index as a platform for Artificial Intelligence (AI)

Artificial Intelligence (AI) is seen as the solution to many of today’s problems. Providing AI allows, for instance, building services that can assist users in their daily goals, services that can provide intelligent analysis of vast data collections, and services that can make smart decisions based on reasoning.
All Artificial Intelligence applications, however, need data as their “fuel”. The lack of an open and large-scale Web data resource leads to AI application not fulfilling their potential. The Open Web Index would provide the solution for applying AI to Web data and, more importantly, for allowing to combine this Web data with proprietary data. Such combinations will allow for reaching the full potential of AI.

2.6 What can the OWI be used for besides serving as a source for search engines?

A web index puts unstructured and heterogeneous data of the Internet into a structured shape. This enables setting up Internet services, which process this data for different kinds of usage.

Some examples for Internet services (other than search engines) are:

- Maps and routing services could amend locations with data from the Open Web Index. For instance, additional data such as descriptions of products and services, opening hours, and items in stock in real-time could be added using the OWI to amend descriptions of hotels, restaurants, shops, etc. This data could be divided into selectable tiers with further background information, such as history of cities and sights or prices of hotels, restaurants and shops.
- Price comparison and meta-online shops could obtain their data directly from the OWI, thus making the process of implementing them easier, more complete and transparent.
- The OWI would enable analysing link structures more easily. Questions such as “Who links to whom?”, “Where do link clusters form?”, “Which meaning do they have”? could be answered. This information could then be processed by data mining tools, e.g., for classification/categorisation. Applying the OWI in this ways would help to better understand the Web as an entity.
- The OWI data could be used for identifying trends of any kind. Questions in this area include, but are not limited to, “What topics are users interested in?“, “Is it possible to detect/predict political changes?“.
- Identifying trends can, on the one side, be done by using the extracted data and on the other side by analysing changes of content on websites: "What are newly emerging websites about?",
- The OWI would allow developers to develop methods for combating fake news and other untrustworthy information on a large scale. It would provide data to overcome the limitations of current approaches.
- The OWI would allow researchers from academia and industry alike to access an abundant amount of open data, instead of having to rely on datasets either provided by industry or rather small datasets collected for single projects by the researchers themselves.

3 Implementation

The Open Web Index is a large project for extending Europe's public digital infrastructure. As said, building such an infrastructure by large goes beyond the capacities of a single company or institution. It can only be achieved on the European level with appropriate funding.

For building and running the Open Web Index, a new institution in the form of a foundation should be built. A foundation is preferable over other institutional structures, as it will not be under the direct and short-term influence of political institutions. This will, on the one hand, help to achieve the long-term goal of building and running the index free from direct intervention. On the other hand, this will foster trust in the Open Web Index, as users can be sure that the data is not only reliable but also free from second-party influence.

A successful model of a foundation is the German "Stiftung Warentest", whose aim is helping consumers by providing impartial and objective information based on the results of comparative investigations of goods and services. This foundation played a crucial role in the development of the market economy.

A European Open Web Index foundation could work together with a network of research facilities active in the field of information technology, information behaviour and use, and knowledge management in Europe. This network could also be used to develop the structure of the Open Web Index further, as well as providing the companies and institutions that access the web index with qualified partners from science. This, in turn, would reduce the asymmetry of the research potential of European and North American Internet companies.

4 FAQs

4.1 How do you avoid the construction of a governmental monopoly of the OWI?

From the outset, the organisational and legal form of the OWI needs to be designed independently from governmental influence. Since the OWI does not offer access to end-users, a monopoly situation as the current one is impossible.

4.2 How will the OWI be accessible for end users / who will be allowed to use the OWI?

The OWI will not be accessible for end-users but for companies or organisations which use the OWI for developing and providing Internet services. It follows a B2B business model. Every company or institution that acknowledges the general terms and conditions is allowed to use the OWI.

4.3 Is the OWI intended to become a "European Google"?

NO! The OWI provides a data infrastructure that can be used by other companies or organisations to create new Internet services.

4.4 How does the OWI ensure privacy?

Every client of the OWI has to accept the general terms and conditions. The general terms and conditions of the OWI will define data privacy, strictly following the European General Data Protection Regulation. Every client has to obey to these regulations. In case a client disregards these rules it will lose his access to the OWI.

4.5 How can requests to the OWI be handled in real-time?

When a client of the OWI makes a request to the OWI that returns millions of results, fast processing is vital. The datasets of the OWI need to contain a pre-ranking, which can also be influenced by clients of the OWI. Usually, a request that produces millions of results would only return, e.g., 1000 results. It is, however, possible to request the complete set of results, although each client is limited in the number of results they can receive. Every request exceeding this limit would become subject to a charge and in some circumstances pricey for the client. A side effect of this is that the Open Web Index would lose attractiveness for spammers.

4.6 What about alternative approaches?

Some alternative solutions have been proposed for fostering plurality on the search engine market. All new search engine providers face the problem of having to build their own index, which is a very costly undertaking. Furthermore, what would be gained if we had one or two, even three more search engines on the market? The problem lies not in having a few more search engines, but in providing real search plurality.

The second line of argumentation says that Google should be forced to provide fair and unbiased results. However, as ranking results is always based on interpretations (and human assumptions inherent in the ranking algorithms), there is no such thing as an unbiased result set. Only a multitude of different algorithmic interpretations can help bring about search plurality.

The third line of argumentation calls for Google to open its index to third parties. Then, it would be possible to build (search) applications on top of Google's index. However, the control over the index – and over what third parties would be able to get from the index – would still lie in the hands of a private company, the index would still not be transparent, and there would still be no influence on how the index is composed.

The fourth, and already widely discussed solution, is building a publicly funded search engine as an alternative to the commercial enterprises. However, this again would only add one more search engine to the market, instead of fostering plurality.