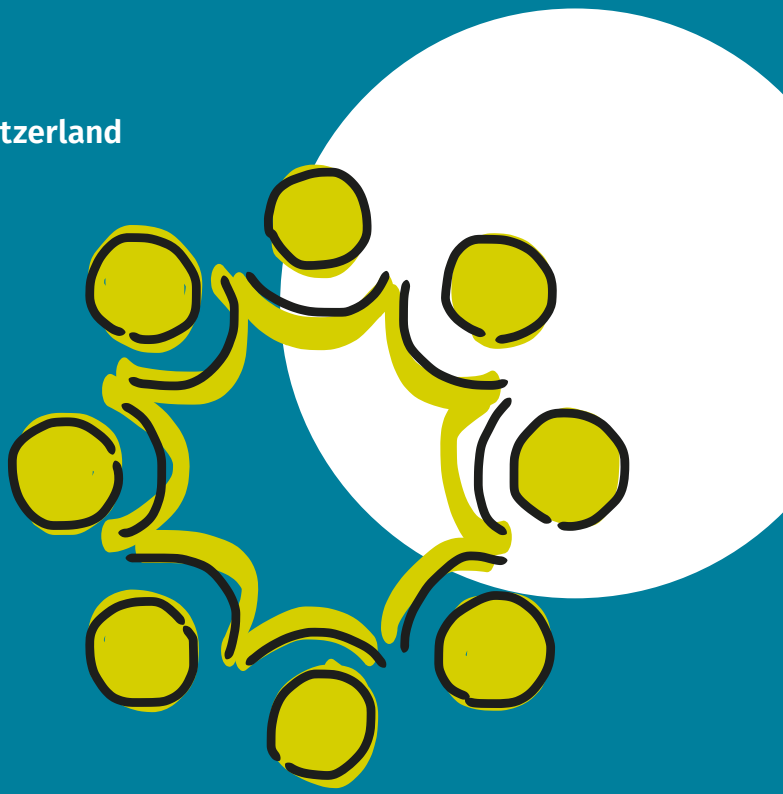
**Editors**
Andreas Wagner
Michael Granitzer
Christian Guetl
Christine Plote
Stefan Voigt

# Proceedings
# 3rd International Open Search Symposium
#ossym2021

**11-13 October 2021**
**Online, hosted by CERN, Geneva, Switzerland**



open search foundation

CERN

ISSN: 2957-4935
ISBN: 978-92-9083-633-9

Editors:

Michael Granitzer, University Passau, Germany
Christian Güetl , Graz University of Technology, Austria
Christine Plote, Open Search Foundation, Germany
Stefan Voigt, Open Search Foundation, Germany
Andreas Wagner, CERN, Geneva, Switzerland

ISSN: 2957-4935
ISBN: 978-92-9083-633-9
DOI:



This report should be cited as:

Proceedings of 3rd International Open Search Symposium #ossym2021, Online, hosted by CERN, Geneva Switzerland, 11-13 October 2021, M. Granitzer, C. Gütl, C. Plote, S. Voigt, A. Wagner (eds). https://doi.org/10.5281/zenodo.6840911

# Foreword

The Open Search Symposium - #ossym - with now its third edition, has become a vibrant meeting place for the ever-growing open search community - a community of scientists, developers, entrepreneurs, public organisations, concerned citizens and many more, to care for and thrive to develop the next generation of public, open and democratic internet search. In times when information, data, privacy, and separation of true and fake in the digital sphere becomes ever more important the open search community seeks to develop methods, tools, and technology to cooperatively crawl the web, generate substantial open web-indices and build a web search ecosystem that is governed by openness, ethics, respect of privacy as well as democratic and legal values.

With these first 'fully-fledged' proceedings of #ossym 2021 we are happy to kick-start a routine series of #ossym-proceedings aggregating and disseminating the articles and findings submitted to and presented at the yearly Open Search Symposia also in the years to come. It is great to see the quality and quantity of the contributions and we hope that these proceedings will help to raise awareness and may help to inspire even more active participation in the Open Search Initiative from all relevant scientific, societal, public, and economic domains.

We hereby kindly want to thank all the active authors, contributors and supporters of #ossym series for making the #ossym conference an interesting and dynamic place for discussing and developing the next generation of internet search - for the current and future users of an democratic, fair and open Internet – for us all.

Together, for a better net.

Andreas Wagner, Michael Granitzer, Christian Gütl, Christine Plote and Stefan Voigt

# More information

- Conference Website on CERN Indico
  https://indico.cern.ch/e/OSSYM-2021
- Open Search Community at Zenodo
  https://zenodo.org/communities/opensearch/
- Event information at the Open Search Foundation
  https://opensearchfoundation.org/ossym21

# Symposium Organisation

## Programme Committee



## Symposium Organising Committee



## Symposium Local Support

# Contents

Preface ..... I
Editorial ..... II
Copyright notice. ..... II
Citation ..... II
Foreword ..... i
More information ..... i
Symposium Organisation ..... ii
Research Track Information ..... iv
Papers ..... 1
APP-P01 - Towards Open Search Applications for the Broader Community ..... 1
ASA-P01 - Goggles: Democracy dies in Darkness, and so does The Web ..... 7
CCA-P01 - Privacy in Open Search: A Review of Challenges and Solutions ..... 15
SND-P01 - Searching on Heterogenous and Decentralized Data: A Short Review ..... 21
SND-P02 - A Proposal for Client Based User Profiles for Open Search in Large and Highly Connected Organizations ..... 27
SSE-P01 - Towards Open Domain Literature Based Discovery ..... 33
SSE-P02 - Modules for Open Search in Mathematics Teaching ..... 39
WCA-P01 - Creating a Dataset for Keyphrase Extraction in Physics Publications and Patents ..... 45
WCA-P02 - The Impact of Main Content Extraction on Near-Duplicate Detection ..... 50
Extended Abstracts ..... 56
APP-A01 - Improved Discovery and Access to Research Data in Energy Systems Analysis ..... 56
APP-A02 - Requirements for an Open Search Infrastructure from the Perspective of a Vertical Provider ..... 57
APP-A03 - Conceptual considerations for comprehensive and cooperative crawling and indexing the Web ..... 58
CCA-A01 - The Development of a Social-Media-Strategy for the Open Search Foundation Applying the Social-Media-Cycle ..... 60
CCA-A02 - The effect of search engine optimization on search results: The SEO Effect project ..... 61
FWC-A01 - Avoiding Useless Content while Crawling the Web ..... 63
FWC-A02 - Processing Crawled Data ..... 64
FWC-A03 - From web graphs to prioritizing web crawls ..... 65
SND-A01 - Neuropil – a distributed, privacy-preserving, search index structure ..... 66
SND-A02 - Indico & Citadel Search: A collaboration case study ..... 67
SSE-A01 - Open Search @ DLR - towards transparent access to web-based information in science ..... 68
WCA-A01 - Understanding Websites ..... 69
WCA-A02 - FastWARC: Optimizing Large-Scale Web Archive Analytics ..... 70
Short Abstracts ..... 71
FWC-S01 - URL Frontier: An Open-Source API and Implementation for Crawl Frontiers ..... 71
Appendix ..... 72
List of Autors ..... 72

## Research Track Information

ASA - Alternative Search Approaches
APP - Applications
CCA - Cross cutting aspects
FWC - Federated Web-Crawling
SSE - Open Search for Science and Education
SND - Search and Discovery
WCA - Web Content Analysis

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772468

# TOWARDS OPEN SEARCH APPLICATIONS FOR THE BROADER COMMUNITY

Aleksandar Bobic*[1] CERN, 1211, Meyrin, Switzerland
Melanie Platz, Saarland University, 66123, Saarbrücken, Germany
Christian Gütl, ISDS CoDiS Lab, Graz University of Technology, 8010, Graz, Austria
[1]also at ISDS CoDiS Lab, Graz University of Technology, 8010, Graz, Austria

*Abstract*

In recent years, a handful of commercial search providers dominated the worlds search ecosystem and, to a large extent, the world's information flow. Over the years, many open issues relating to such a dominating ecosystem have been identified, and multiple tools and user groups were created in an attempt to address these issues. As part of the Open Search Foundation's Applications Working Group, we aim to explore open-search-based applications which could address the identified issues. To that end, this paper aims to identify privacy issues connected to the reliance on a few commercialised search solutions, potential data sources which could be used to enhance an open search index, as well as app ideas for enhancing and leveraging an open search index. A first pilot study was performed with two diverse student groups to get an insight into user requirements related to open search.

The findings indicate that students with substantial technical knowledge see search result and market manipulation as the main issues. In contrast, students with less technical knowledge see ads and recommended content as issues. Furthermore, many data sources for an open index have been identified. Finally, a multitude of apps for extending as well as using an open search index was identified. These insights can be used to conduct further studies and guide the development process of new independent apps based on an open search index concept.

## INTRODUCTION

Soon after its invention, the WWW transformed into a universal information system for the broader community enabling data and information creation and sharing but also collaboration and social networking. From its early days on, the nature of the WWW required tools for searching resources and content. Thus, different types of search systems and services emerged, which can be classified as catalog-based, crawler-based, and meta-search-based systems [1–3].

A significant number of search tools and initiatives emerged, however over the years, by vertical integration, consolidation and concentration, only a few big players are still on the market and offer their services [4–6]. Users, businesses, NGOs and even governments increasingly rely on these few services and depend on access to relevant search results to satisfy their information needs but also stay visible to others [5, 7, 8]. Recent developments in the context of copyright regulations in Australia between the government and global search and social media companies have shed some light on dominant markets and high dependencies [9]. Access to content and search indices as well as to fair search results have become basic infrastructures such as water, power supply or GPS data for navigation. This recent example may illustrate just the tip of an iceberg as research groups and interest groups in several subject domains have uncovered various problems over the last years. Issues include aspects such as restricted accessibility; privacy, security, and trust; bias, information asymmetries as well as data-driven discrimination and differentiation [10–14].

Several tools and services have been initiated to overcome at least partially the above-outlined issues. These cover a wide range of features from privacy-preserving access to search services to alternative and domain-focused search to an open search index (OSI). This trend is also reflected by diverse user groups who want to get their information needs served apart from mainstream services in a transparent and unbiased way [13, 15–17]. In particular, the latter motivated us to initiate the Applications Working Group within the Open Search Foundation (OSF) [17, 18] and to explore open-search-based apps for the broader community.

This research aims to identify possible applications based on an OSI for the broader community by following the Design Science Research (DSR) approach. This paper focuses on student user groups in different subject domains to uncover their needs, extract possible features for search apps and ideas for apps built on the OSI. To this end, the remainder of this paper is organized as follows: section 2 covers background and related work, section 3 outlines the study design, which is followed by findings and discussion in section 4. Finally, section 5 introduces a first app idea based on the findings which will be implemented as a sample app within the OSF community. The main contribution of this work is the identification of multiple user needs in the context of an OSI, multiple groups of features for an open search app and various app categories which could be built on an OSI. These insights can guide future developments of open search indices and the apps using them.

## BACKGROUND AND RELATED WORK

Time Berners-Lee's first proposal of the World Wide Web (W3 or WWW) in 1989 has adoption soon by the research community, and become a global and widely used information and collaboration space in many application domains [19, 20]. Due to the concept and architecture de-

* aleksandar.bobic@cern.ch

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772468



sign, a flexible and scalable space for information and data publishers and consumers has been created. However, the decentralized architecture of the W3 implementation also yielded issues such as broken hyperlinks between resources and opened new questions such as how to effectively index and search for resources [21].

Right after the usage of the WWW by geographically dispersed user groups, the first search services and tools appeared already around 1993. One of the first was a manually compiled and managed central index of Web resources for browsing, also followed by distributed ones and enhanced with search features, which have become known as search catalogs [1, 4]. Already in the same year, the first web crawler, originally for Web statistics, appeared. Soon after, it was used to build an automatically compiled search index, first based on a few metadata and gradually enhanced towards a full-text search and is also known as crawler-based search engine [1, 2, 4]. Unlike the first central indexes, distributed crawler-based search engines appeared in the mid to late 1990s. This search architecture allows to splits organizational and computational effort, and geographically or subject-based crawling of portions of the Web were possible. In the same way, distributed indices - also geographically and/or subject-based – could be built from one or several crawler instances, and even other indices could be used. Finally, lightweight and rich user clients and interfaces can serve the users according to their needs and end-devices [22, 23]. This conceptual design of a distributed search system has also inspired us to define a conceptual architecture for researching search applications for the broader community.

Soon after the first availability of crawler-based search engines, systems appeared, making use of more than one search engine to increase the coverage of the Web index and provide a unified user interface for the clients. This type of search service has become known as a metasearch system [24–26]. Finally, search systems used by individual users indexing defined parts of the Web and local resources, have been phrased as desktop search [27].

As the Web has developed from a mainly passive consumer-based information system into an active "prosumer" (producer and consumer) one over the last two decades [20], information and communication as well as collaboration habits have changed [28, 29], new features, apps and expectations from different stakeholders have evolved and further are on the horizon.

From a business and start-up viewpoint, in addition to the main search providers, such as Google and Bing, multiple new providers were created, such as Ecosia, DuckDuckGo, Qwant, Chatnoir and more. These alternatives are mainly targeting niches on geographic and content focus as well as privacy protection. However, most of these cannot compete with the largest providers either because of a lack of funding or simply because most popular pages do not allow them to crawl their content [18].

Interest groups, NGOs and policy makers such as the European Union are focusing mainly on security (e.g. exclude harmful content) and privacy (e.g. protection of private data and behavior, right to forget) as well as unbiased information access and the support of minorities [10, 14, 30, 31]. The OSF campaigns for on open and distributed search index, unrestricted accessible and protecting privacy aspects [18].

Research communities in the field of Information Retreival (IR) but also in related fields have recently emphasized search features, methods and technologies for the next level of retrieval and web search systems, and suggested possible applications direction for the future. The research communities have identified many features in this context: conversational information seeking; support of complex and long-term information seeking; complement of search result pages by generated information objects; personalized IR but also access to personalized information; incorporating the IoT world in the retrieval process; learnable IR optimizing feature and indexing structure as well as relevance function; new and distributed IR architectures trust concepts; explainability of selection criteria and results. As additional non-functional aspects have fairness, accountability, confidentiality and transparency gained increasing attention [11, 32, 33]. Identified and suggested apps include new ways of web search including domain specific and personalized search; information discovery in social networks, medical search, and e-commerce; conversational information seeking system; personal information systems, notifications and push system; bibliometric and scientific content search; investigative journalism and learning activities [11, 32–34].

Advances in NLP, AI and IR as well as in hardware and software aspect have paved the way to innovative and future-oriented features for research and developing new and enhanced retrieval systems. Therefor new and supportive applications for various user groups are feasible in near to mid-term future. Although search applications and the application of new search features are covered in research, to our best knowledge less has been reported on the needs of the end users and the broader community. Therefore, we want to contribute with our research in gaining better insight in the needs of various user groups. In this paper, we report about our initial step of research and findings by a first selected user group.

## STUDY DESIGN

DSR is applied as part of this work to develop apps based on an OSI for the broader community. DSR typically involves creating an artefact - in our case, apps based on an OSI - and/or design theory - in our case, design principles for apps based on an OSI - to enhance and optimise the current state of practice and existing research knowledge. To identify the user needs, be able to make statements about the relevance of (envisioned or developed) apps and perform "good" DSR [35], collective case studies are performed by conducting surveys [36]. Along these lines, two cases are presented in the following.

To investigate how students view privacy issues connected to search posed by large search monopolies, investigate how

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772468

they would approach such issues and identify potential solutions which could be developed in the context of the OSF and based on the previously mentioned goal of developing apps based on an OSI for the broader community this paper aims to answer the research questions in Table 1. To answer these questions we surveyed one student group and led a discussion session with another smaller student group and later analysed their responses.

Table 1: Research questions

| | |
|---|---|
| **R1** | How do university students view privacy issues of commercial search providers? |
| **R2** | What are the data sources students would use for an open search index? |
| **R3** | What applications would students like to see developed as part of an open search index? |

### *Settings and Instruments*

The students were provided with a document that contained the following items:

- Brief introduction describing how most people and communities are dependent on a few search services
- Description of a conceptual architecture of an open search index seen in Fig. 1.
- Five tasks requiring the student to discuss:
  - Issues of depending on a few search providers
  - Identifying five or more sources that could be used to extend the open search index data
  - Identifying five or more applications that extend the search index data
  - Identifying five or more applications that use the indexed data to provide services
  - Details of two example applications they provided
- A template file with marked sections for each answer, which enables an easier analysis of students answers.

The architecture in Fig. 1 depicts an OSI (3) which is primarily built using data from public web content (1) and data from dynamic content services (2). Data source centred apps serve as an additional source of data that enriches existing index data or provides additional data. The index data can be retrieved either using the open search service or through additional retrieval centred apps that would use the data to provide additional services which are not directly related to search.

### *Study Participants*

The first student group called the *ISR Group* was recruited from an Information Search and Retrieval course at the Graz University of Technology and included 48 students. These

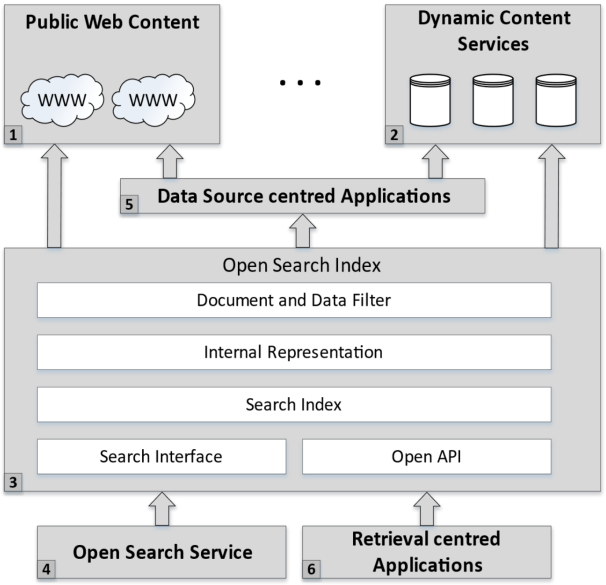



Figure 1: Simplified conceptual architecture of the OSI.

were mainly Masters students focusing on computer science. Additionally this group completed the tasks online, unobserved and without any supervision or input from a lecturer. The second group called the *TET Group* consisted of 6 students and was recruited from the University College of Teacher Education Tyrol. These students focus on primary school education with a focus on mathematics. Furthermore, this group was guided by a lecturer online.

### *Data Collection and Cleaning*

During data cleaning and analysis it was noticed that some students misunderstood questions. These answers had to be removed. Additionally some of the answers contained suggestions which are an essential part of any search app and were therefore left out.

## ISR GROUP - STUDY RESULTS

### *Reliance on a Few Search Providers*

We first identify multiple *issues of depending on a few search providers* which we divided and summarized into three groups based on their focus.

**Search result influence** can shape peoples' opinions by hiding information and changing the rank of web pages. Since commercial companies usually provide search engines, they tend to focus on profit and could therefore influence the results to benefit them financially but would not benefit users. Furthermore, they might influence search results to gain political power, potentially enabling them to gain more profits. On the other hand, authoritarian regimes might force companies to adapt their search results without the users' knowledge. Since commercial providers do not share their algorithms, they are essentially black boxes, making it hard or impossible for users to realise when censoring or the data manipulations mentioned above occur.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772468



The **market influence** group focuses primarily on the interactions between commercial search providers and other companies. Since most of the world relies only on four primary search engine providers, they dominate the market and therefore influence the regulations and their services to suit them best. Furthermore, they enable other companies to advertise in their search results. This creates an environment favouring large established companies with more money and penalising smaller companies.

Some of the issues do not fit in the above groups and are therefore listed here. Firstly, commercial search providers tend to provide their services for free and collect large amounts of personal data in exchange. Because of the above-mentioned lack of transparency, collected data can be used to influence users' opinions or sell their data for profit without the users' knowledge. Additionally, commercial search providers own large server farms, which have a negative environmental impact. Finally, due to market dominance, most people and companies rely on these commercial services to do their tasks. In case one of these services stops operating or blocks, certain entities it could disrupt the entity's source of income.

## *Data Sources for Extending the OSI*

To identify *what data sources could be used to extend the OSI data* we grouped students' suggestions based on the source type. Sources are firstly split into digital and physical resources. The physical resources include maps, books and old church archives. The digital resources are further split into private and public sources.

**Private digital sources** represent data accessible only by authenticated individuals. They are split into local data located and generated by a personal device and global data located on a protected server, a company's intranet, or even on the OSI. Examples of local sources include user device data, smart device health data and more. On the other hand, examples of global sources include search queries, indexed user data, private social content, a company's intranet, and more.

**Public digital sources** describe data publicly accessible either by crawling the Web or using an API. These sources are further split into multiple sub-groups. *Verified* sources from official institutions or individuals such as research data, statistics about a country, health data and more. *Social* sources generated through online social interactions such as user relations, social media posts, mailing list content, newsgroup content, question answering sites such as Quora[1], expert input and more. *Document* sources usually stored in larger groups on dedicated servers that may include academic documents, books and more. *Dataset* sources which are either hosted or accessible using APIs and usually contain aggregated data about a specific topic include AWS open data[2], google public datasets[3] and other publicly accessible datasets. *Media* sources such as images, video, audio content and content related to them such as subtitles, location data and more. *Device* sources generated by publicly accessible devices such as ticket machines, cameras, satellites, public IoT devices and others. *Geographic* sources which usually describe some geographic or traffic aspect include location data, land, water or air traffic data and more. *Other* sources which do not belong into any of the categories mentioned above include user interactions, online wikis, archived content, online retailer content, public data about individuals, code repositories and more.

[1] quora.com

[2] https://registry.opendata.aws/

[3] https://www.google.com/publicdata/directory

## *Applications for Extending the OSI*

Next, we investigate suggestions for applications that could *help extend the OSI with new content*. We categorise students' suggestions into two main groups.

**Apps requiring user input** are represented by two sub-groups. First, apps analysing existing content, such as generating user profiles, web page statistics, and user location tracking. Second, apps that collect user content such as web page reviews, self-reported health symptoms, submission of business opening hours, submissions of research data, and more.

**Apps processing existing data** generate new content from indexed data. They are split into three sub-groups.

*Content analysis apps*, analyse existing content to provide new insights such as plagiarism identification, information credibility estimation, content license information and more. Furthermore, these apps also focus on calculating the number of ads and tracking scripts, analysing inter-page interactions by investigating how a topic spreads among pages and analysing the similarity between pages. Finally, these apps also focus on media content by, for example, comparing similarities between videos.

*Content generation apps* generate content from existing data by adding descriptions and metadata to documents, detecting text and objects in images, summarising content or extracting statistics from datasets or financial data.

*Content modification apps* generate content by modifying existing data. These apps could, for example, remove bias, automatically translate text or adjust the language level based on the readers' experience.

## *Applications for Using the OSI*

The analysis of *app ideas for using the OSI* resulted in a diverse set of ideas that can be primarily divided into two categories.

**Search focused ideas** contain ideas for improving the search functionality. It is further divided based on search features and search targets.

*Search feature ideas* describe suggestions aiming to improve the search experience. These include search execution features such as searching using various input methods and filters, optional search using user preferences, custom-ranking algorithms, and more. Additionally, this group includes ideas for result presentation such as summarization results, filtering results with specific properties such

https://doi.org/10.5281/zenodo.5772468

as mobile-friendly pages, presenting connections between different search results and more. Students also suggested multiple ideas for the general search infrastructure such as a search API, self-hosting search indices and browser integration. Finally, general suggestions included the option to delete indexed personal data, support for question answering, optional gathering and analysis of usage data, optional usage of interaction to improve retrieveal and more.

*Search target ideas* focus on search environments, data types and datasets which could be used to perform searches. Suggestions for environments include traditional online searching, offline searching through a local device, searching inside of webpages, inside of documents and search through multiple sources and data types. Data type and dataset suggestions include text documents, media files, sources of a specific document, user reviews, research papers and more.

**App focused ideas** include suggestions which are not directly associated with search. Just like some of the previous categories this one is divided into multiple sub-categories.

*Geo focused ideas* primarily encompass features revolving around maps and traffic. Students suggested ideas include an open maps app, navigation app, live visitor data tracking and smart navigation based on crowd movement. Additionally, they also suggested that users should be able to submit their geo data.

*Analysis focused apps* contain data analysis and monitoring solutions. Idea examples include data visualisation dashboards, trend analysis tools, a data exploration and analysis tool, and more.

Other app idea examples include virtual assistants, government and education-focused apps, news and statement validity estimation apps, business-oriented apps, literature research apps, social networks and many more.

## TET GROUP - STUDY RESULTS

Due to the group focus on primary education, tasks were translated into the German language, and some questions had to be left out or reformulated to adjust the survey to the technical understanding of the group. Therefore, only data Sources for extending the OSI and apps for using the OSI were focused.

### *Data Sources for Extending the OSI*

Students suggested including high-quality teaching materials and books such as literature for children. Furthermore, they suggested including more details about the retrieved literature to decide if a product is suitable before buying it.

### *Applications using the OSI*

Students would like to adjust search results using search result filtering, restricting searching to a location and removing ads and networking with, e.g. Facebook. Additionally, they would find it helpful to select the results' language and show the results in their original language (with the optional possibility to translate page afterwards).

This group focused primarily on ads and dynamically recommended content issues. On the other hand, the ISR group mainly focused on privacy-related issues. A lack of technical understanding could explain this difference. It could also further indicate that the wider public should be more informed about tracking and data privacy issues.

## CONCLUSION AND FUTURE WORK

This paper firstly describes the open issues of relying on a few large commercial search providers and introduces open search-based apps. As part of this work, a study is conducted to identify what students view as issues of relying on a few large commercial search providers, what sources they would use to enhance an open index and which open index-based apps they would like to use. A large group of students focusing on IR was surveyed, and a smaller group focusing on teacher education participated in a discussion led by their lecturer. Their answers include issues concerning privacy, dynamic content suggestions and ads. Furthermore, many data sources and data source groups were suggested to enhance an OSI. Finally, multiple app categories that could either enhance an OSI or use it to offer services were identified. The study results can be used to guide further studies and search-based app development.

To broaden the insights of the current study, strengthen our understanding of user needs and identify further issues relating to commercial search giants, we plan on carrying out a study with a more extensive scope in the future. Additionally, as part of the OSF Applications Working Group, we plan on developing an initial prototype based on the insights gathered in this study.

## ACKNOWLEDGMENTS

We would like to thank other members of the OSF Applications Working Group for their contributions to the discussions in our meetings and to the two student groups for participating in the study.

## REFERENCES

[1] Tom Seymour, Dean Frantsvog, and Satheesh Kumar. History of search engines. *International Journal of Management & Information Systems (IJMIS)*, 15(4):47, sep 2011.

[2] Saeid Asadi and Hamid R Jamali. Shifts in search engine development: A review of past, present and future trends in research on search engines. *Webology*, 1(2), 2004.

[3] C. Mic Bowman, Peter B. Danzig, Udi Manber, and Michael F. Schwartz. Scalable internet resource discovery: Research problems and approaches. *Commun. ACM*, 37(8):98–ff., August 1994.

[4] E. Van Couvering. The history of the internet search engine: Navigational media and the traffic commodity. volume 14, pages 177–206. 2008.

[5] Patrick Barwise and Leo Watkins. The evolution of digital dominance. *Digital dominance: The power of Google, Amazon, Facebook, and Apple*, pages 21–49, 2018.

https://doi.org/10.5281/zenodo.5772468




[6] Edward Iacobucci and Francesco Ducci. The google search case in europe: Tying and the single monopoly profit theorem in two-sided markets. *European Journal of Law and Economics*, 47(1):15–42, 2019.

[7] Shuang Li, Xiang Lan, Yuezhi Zhou, and Yaoxue Zhang. Exploring and understanding web search behavior with human activities. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE, 2017.

[8] Aleksandar Grubor and Olja Jakša. Internet marketing as a business necessity. *Interdisciplinary Description of Complex Systems: INDECS*, 16(2):265–274, 2018.

[9] Bridget Judd. Google is threatening to pull its search engine from australia. so what does that mean for you? - abc news, 01 2021. https://www.abc.net.au/news/2021-01-25/google-may-pull-search-engine-from-australia-what-happens-next/13086712.

[10] Orla Lynskey. The power of providence: the role of platforms in leveraging the legibility of users to accentuate inequality. pages 177–201, 2018.

[11] J Shane Culpepper, Fernando Diaz, and Mark 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.

[12] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. I vote for—how search informs our choice of candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, 22:320–341, 2018.

[13] Anas El-Ansari, Abderrahim Beni-Hssane, Mostafa Saadi, and Mohamed El Fissaoui. Papir: privacy-aware personalized information retrieval. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–17, 2021.

[14] Fernando Galindo and Javier Garcia-Marco. Freedom and the internet: empowering citizens and addressing the transparency gap in search engines. *European journal of law and technology*, 8(2):1–18, 2017.

[15] Wikipedia. List of search engines. `https://en.wikipedia.org/wiki/List_of_search_engines`, 2021.

[16] Wikipedia. Outline of search engines. `https://en.wikipedia.org/wiki/Outline_of_search_engines`, 2021.

[17] OpenSearchFoundation. Open search foundataion official website. `https://opensearchfoundation.org/en/open-search-foundation-home`, 2021.

[18] Daisuke Wakabayashi. Google dominates thanks to an unrivaled view of the web - the new york times. `https://www.nytimes.com/2020/12/14/technology/how-google-dominates.html`. (Accessed on 17/06/2021).

[19] CERN. A short history of the web. `https://home.cern/science/computing/birth-web/short-history-web`, 2021.

[20] Karwan Jacksi and Shakir M Abass. Development history of the world wide web. *Int. J. Sci. Technol. Res*, 8(9):75–79, 2019.

[21] Tim Berners-Lee, Robert Cailliau, Jean-François Groff, and Bernd Pollermann. World-wide web: the information universe. *Internet Research*, pages 52–58, January 1992.

[22] C Mic Bowman, Peter B Danzig, Darren R Hardy, Udi Manber, and Michael F Schwartz. The harvest information discovery and access system. *Computer networks and ISDN Systems*, 28(1-2):119–125, 1995.

[23] Keith Andrews, Christian Gutl, Josef Moser, Vedran Sabol, and Wilfried Lackner. Search result visualisation with xfind. In *Proceedings Second International Workshop on User Interfaces in Data Intensive Systems. UIDIS 2001*, pages 50–58. IEEE, 2001.

[24] M Manoj and Jacob Elizabeth. Information retrieval on internet using meta-search engines: A review. *Journal of scientific & industrial research*, 67(10):739–746, 2008.

[25] Susan Gauch, Guijun Wang, and Mario Gomez. Profusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computer Science*, 2(9):637–649, 1996.

[26] Erik Selberg and Oren Etzioni. The metacrawler architecture for resource aggregation on the web. *IEEE expert*, 12(1):11–14, 1997.

[27] Bernard Cole. Search engines tackle the desktop. *Computer*, 38(3):14–17, 2005.

[28] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. Not quite the average: An empirical study of web use. *ACM Transactions on the Web (TWEB)*, 2(1):1–31, 2008.

[29] Yupeng Jiang, Yingying Zhao, and Hui Xu. Analysis and portrait of college students' online behavior habits. In *Journal of Physics: Conference Series*, volume 1302, page 022007. IOP Publishing, 2019.

[30] Lucas D Introna and Helen Nissenbaum. Shaping the web: Why the politics of search engines matters. *The Information Society*, 16(3), 2006.

[31] Jahna Otterbacher. Addressing social bias in information retrieval. *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 121–127.

[32] Iván Cantador, Massimo Melucci, Max Chevalier, and Josiane Mothe. Circle 2020-the first joint conference of the information retrieval communities in europe. In *Joint Conference of the Information Retrieval Communities inEurope (CIRCLE 2020)*, volume 2621, 2020.

[33] Zhumin Chen, Xueqi Cheng, Shoubin Dong, Zhicheng Dou, Jiafeng Guo, Xuanjing Huang, Yanyan Lan, Chenliang Li, Ru Li, Tie-Yan Liu, et al. Information retrieval: a view from the chinese ir community. *Frontiers of Computer Science*, 15(1):1–15, 2021.

[34] Yvonne Kammerer, Saskia Brand-Gruwel, and Halszka Jarodzka. The future of learning by searching the web: mobile, social, and multimodal. 6:81–91, 2018.

[35] Alan R Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.

[36] Robert K Yin. *Case study research and applications: Design and methods*. Sage publications, 2017.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044986

# GOGGLES: DEMOCRACY DIES IN DARKNESS, AND SO DOES THE WEB

R. Berson, S. Sathyanarayana, A. Karaj, E. Larsson, and J.M. Pujol
Brave Search, Munich, Germany
search@brave.com

*Abstract*

This paper proposes an open and collaborative system by which a community, or a single user, can create sets of rules and filters, called *Goggles*, to define the space which a search engine can pull results from. Instead of a single ranking algorithm, we could have as many as needed, overcoming the biases that a single actor (the search engine) embeds into the results. Transparency and openness, all desirable qualities, will become accessible through the deep re-ranking capabilities *Goggles* would enable. Such system would be made possible by the availability of a host search engine, providing the index and infrastructure, which are unlikely to be replicated without major development and infrastructure costs. Besides the system proposal and the definition of the *Goggle language*, we also provide an extensive evaluation of the performance to demonstrate the feasibility of the approach. Last but not the least, we commit the upcoming Brave search engine to this effort and encourage other search engine providers to join the proposal.

## MOTIVATION

Democracy dies in darkness, a line recently adopted by the Washington Post as their slogan, warns us that unless people are informed with facts and truth, no true democracy is possible. Those who benefit from darkness have always tried to control media in order to control and manipulate public opinion with propaganda. Until recently, propaganda has been the exclusive domain of nation-states or state-sponsored actors through mass media [19]. With the mass popularization of the Web in the last two decades and the subsequent privatization of it by big platforms like Google, YouTube and Facebook, the paradigm has changed. Propaganda is no longer a tool of an elite, but it has been commoditized to the extent that it is as accessible as advertisement, becoming a weapon that too many actors have access to. One must appreciate the irony that those most vocal about the risks of propaganda are those who controlled it in the past. Nevertheless, the risk of fake-news—a neologism created to mitigate cognitive dissonance—cannot be ignored [36, 5, 30, 6, 33]. It is dangerous for a society if people living in it cannot distinguish between facts, opinions and outright misinformation. Although this danger has always existed, today the situation is dire if only because quantitative becomes qualitative and although all information is theoretically available, in practical terms it is not.

### *A Single Point of Failure*

Like never before, all the information (and misinformation) of the world is available upon request. But the way to access this information has narrowed to become a quasi-monopoly. The abundance of information has led to a significant transfer of power from creators to aggregators. Access to information has been monopolized by companies like Google and Facebook [23]. While everything is theoretically still retrievable, in practice we are looking at the world through the biases of a few providers, who act, unintentionally or not, as gatekeepers. Akin to the thought experiment about the tree falling in the forest [3], if a page is not listed on Google's results page or in the Facebook feed, does it really exist?

The biases of Google and Facebook, whether algorithmic, data induced, commercial or political dictate what version of the world we get to see. Reality becomes what the models we are fed depict it to be [25]. And a reality defined by Google's search ranking algorithm, is one that does not and cannot capture the intricacies and variety of human knowledge and opinion.

Traditionally, the role of media was to serve as the middleman separating the chaff from the grain, of course with their respective biases. Journalists and editors were the curators and the publishing house was responsible by reputation and by law. Furthermore, every country had tens or hundreds of, to a certain degree, independent firms. Media consolidation in the 90s somewhat killed the field [37], reducing the number of firms able to filter information. But the real impact came with the consolidation of the big Internet platforms, basically Google and Facebook. The role of curation has been eliminated as the majority of value is captured by the platforms so it is no longer economically viable [10, 18, 26]. With fewer and weaker intermediaries, we also reduce the amount of independent points of views or windows to the world.

We have been forced to trust that the worldview of a few internet platforms is non-partisan while it clearly cannot be. The public space has been privatised by a handful of private corporations. Such concentration of access to information is a single point of failure, and it has failed.

## PROPOSAL

Let us start with a disclaimer; there is no technical solution that solves the aforementioned problem once and for all. The issues derived from monopolies are well understood and fall well beyond the reach of any technical solution.

However, what we could do, is to acknowledge that market dynamics coupled with freemium models tend to produce a winner-takes-all scenario [4], the prelude of monopolies. Under these market constraints, we propose to increase the

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044986



number of options, windows through which reality is made sense of. While it would be desirable to achieve that goal through independent actors (platforms), in lieu of that we can achieve the same effect within the same platform. The proposal presented in this paper can be portrayed as a fail-safe to prevent any platform from becoming a single window to the world. If Brave or any other company were to displace Google, the ranking algorithm would still be the one dictating the way the world is perceived. We would have changed actors, but the problem would remain.

In this paper we introduce ***Goggles***, which is meant to provide people with a way to access information according to their **explicit biases**. In layman's terms, to put *Goggles* on, to see a different version of reality.

Search engines are free to incorporate user-defined *Goggles*, specified in an open language drafted in Section , and modify their ranking so that the user's explicit preferences take precedence over the ranking of the search engine itself.

Such system would have the potential to pierce a hole in the single-window effect produced by the search engine's ranking algorithms. In a way, it is opening the ranking algorithm to the people using the search engine.

*Goggles* go beyond personalization. As a matter of fact, they are orthogonal. The rationale is not to customize the ranking according to the implicit interests of the user, but to offer a mechanism to define multiple rankings, plural, open and explicit, for only if it is so, can it be trusted. The benefit for the users is that they would be empowered to explore multiple realities in a straight-forward way. The point is to offer people the freedom to choose their own biases while being conscious of them. The benefit for the content creators is that they have multiple options to expose their content, by increasing their potential audience, which will reduce the need to optimize for the single set of biases implicitly encoded in the search engine's ranking [17].

The point is not to create an even stronger echo-bubble, which is what happens under personalization. Rather, the aim is to promote plurality and let people proactively and consciously choose. Confirmation bias exists; people tend to only acknowledge information that fits their own bias [27]. However, a large fraction of people are interested in exploring alternative viewpoints [14]. Current platforms, however, do not facilitate such exploration process [22], seeking alternative options (*for better or worse*) implies a cost. The costlier it is, the less likely it becomes for people to break from the single-window effect exacerbated by the ranking algorithms.

It is also not the point of *Goggles* to mitigate the fake-news phenomena, at least not directly. While having more plurality opens the space for wacky theories, it also opens the space for rational and informed ones. The way to fight fake news is to rebate them, not to ban or bury them [11]. Otherwise we will have no instrument left to control those who decide what qualifies as fake [1].

We envision a scenario where a community of people create and curate *Goggles* like,

- **"Tech Blogs".** Imagine searching through a collection of personal and company blogs curated by the community.
- **"Product Reviews without commercial intent".** Get rid of all sites with price comparisons, affiliate links, etc. Basically, to browse over product descriptions and reviews.
- **"Independent Media for any country".** Would demote major newspaper and promote minor outlets.
- **"Exclude top 1000 domains".** Would remove results from most popular domains on the Web to surface less prominent ones.
- **"Recipe search that my mom likes".** Only searches recipes on tasteofhome.com, nowhere else being considered, would become a site search.
- **"Nature lovers in the Pyrenees".** An extremely curated list of high-quality sites for hiking/trekking in the area. Excluding the more generic sites not specialized in that area.
- **"Wikipedia / Reddit / <Any site> search".** Site search is just an instance of what *Goggles* can be. The other way round also works; results that exclude results from a given site (e.g. Facebook).
- *We recently observed the tech community discussing the shortcomings of search engines [9], particularly in surfacing content by some spaces in the web. It was exciting to see how almost all the use-cases in the discussion could be addressed by* Goggles.[2]

Each of these *Goggles* is fully owned, controlled and maintained by its creators according to their own terms and services. *Goggles* can be shared, extended, and modified to fit anyone's particular needs. The most likely scenario, however, is that the great majority of users will rely on *Goggles* maintained by others because of their coverage, quality, and most importantly, because of the trust of the maintainers' integrity. Trust is an important aspect of *Goggles*. There is no way to guarantee that a particular Goggle fulfils its promise, but any Goggle can be forked, and their users vote with their feet. The fact that the list of rules composing a Goggle is open and can be copied/extended by anyone will prevent the creation of a lock-in by the original authors/creators, mimicking the ecosystem lock-in of the likes of Apple, Google and Facebook [28]. Of course, for such system to work, people must trust that the search engine serving as host applies the rules defined by the Goggle against their index without alteration. Besides the language definition, which must be

[1] *Quis custodiet ipsos custodes?* Who will guard the guards themselves?

[2] Note that *Goggles* project started late 2019 but was put on hold due to the shutdown of the Cliqz search engine. Happily, the project will continue as part of Brave from 2021 onward.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044986



standard to allow integration with the search/retrieval algorithms, a search engine should stay out of the *Goggles* ecosystem to maximize trust and variety.

The contributions of this paper are:

1. To propose the concept of *Goggles* for open/collaborative ranking. Note that the proposal/definition alone, is not entirely novel (as will be discussed in the Background Section ).
2. To define the *Goggles* language, which allows people to define their own ranking preferences in a simple way, using a grammar inspired by the ad-blocking community (proven to be both easy to write and maintain and to be expressive enough.)
3. The commitment that the Brave search engine will implement and apply user-defined *Goggles*. Which means modifications on the ranking algorithms (details in Section ). We encourage other search engines to follow. *Goggles* is in no way owned by or exclusive to Brave search engine. It belongs only to its creators and users.
4. To show that search engines can serve an additional role to the community by exposing their infrastructure and index. Allowing public and open access to such privileged resources.

Let us emphasize once again that this proposal, *Goggles*, does not fix the problems of misinformation, echo-chambers, confirmation biases, etc. These problems are very human in nature, and no technology can solve them. At most, it can only exacerbate or mitigate them, the latter being the case of the system presented in this paper. What we propose in this paper is a way to decrease the single-window effect created by the search engines such as Google, Bing, and of course, Brave. By opening the ranking from one(s) to many we open the possibility of having many different rankings, serving different biases and intents. Needless to say, that search engines must collaborate on that effort by providing the infrastructure and index to back it up.

*Goggles* intends to offer multiple perspectives to the same query and to be explicit about it. So that people choosing liberal media *Goggles* are free to do so, but this is a conscious and deliberate choice. If they want, they can explore the opposite *Goggles* to expand their perspective. Something as simple as this is not easy, as systems are not designed to that purpose [7, 32]. Allow us to stress that the biases embedded on a Goggle do not need to be "positive". There will be *Goggles* created by creationists, anti-vaccination supporters or flat-earthers. However, the biases will be explicit, and therefore, the choice is a conscious one. We do not anticipate any need for censorship in the context of *Goggles*. Clearly illegal and sensitive content like child pornography or extreme violence should already be filtered out by the host search engine at the index layer. Consequently, such content should not be surfaced by any Goggle.

We would like to stress out that biases do not need to exist only on highly polarizing issues such as politics, religion, language, etc. Non-partisan topics like strong localization, advertisement or commercial intent removal are likely to have a strong presence. *Goggles* can just be ways to increase plurality and open niches for content that is otherwise buried under the rule of a single source of ranking.

## BACKGROUND

To the best of our knowledge *Goggles* is the first attempt to open up the ranking component of a search engine to the community.

Perhaps the most related system to *Goggles* is personalization [24], the ability to alter ranking according to the user's interests or intents. Note that this comparison, although reasonable, is deceptive. Personalization, outside the realm of faceted search [34, 2], is not actionable for the user, at most they can opt out from it. The aim of *Goggles* is not to have a single ranking fitting better the user's interests, but to offer users a wide range of possible rankings and let them choose. The same rationale applies to rankings subjected to locales, either language or geography.

We mention faceted search, which shares with *Goggles* that ability to provide external information to the query to help the search engine refine the results the user was looking for. In the case of faceted search, the user does not provide an external rule for ranking, but additional metadata, typically in a structured form. For instance, named entities, reference codes, dates, etc. Information provided by the user to facilitate the retrieval. This approach is useful on many verticals like flights, trips, books, movies, products, but is not the most convenient for general purpose, as it demands from the user a) knowledge of the domain, and b) extra burden on the input query. *Goggles* also imposes these constraints at creation time, but not while using them. Thus, the extra effort is not paid by the end-user but by the Goggle's creator/maintainer.

*Goggles* also share similarities with collaborative efforts for content discovery and classification, for instance, social bookmarks systems [20, 29] or curated lists [31]. However, such systems are designed for sharing and not suitable for search both because of the limited coverage and the lack of a proper search infrastructure.

Another area where *Goggles*' contribution is relevant is algorithmic transparency. We are not aiming to make the Brave search engine ranking transparent, but rather to allow people to modify and alter it *a posteriori*. Transparency of the ranking would provide explainability and accountability for the results and it would help to detect unfairness or illegitimate biases (e.g. gender, race, religion). We could achieve similar results with *Goggles*, but in an indirect manner. Note that full transparency on the ranking (the main ranking algorithm that is) would introduce challenging problems. Intellectual property aside, which is not a small thing, we would further open the search engine to the harmful effects of SEO (search engine optimization). SEO, especially when invasive, is one of the biggest headaches search engines have, giving access to the particularities of the main

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044986



ranking would immediately result in a boost of those sites that rely on SEO to be on top, which are usually not the ones with the best content.

A similar argument can be made on the topic of open search. This proposal does not open the full search engine, but it provides the ability to modify the most important constituent, the results. Building, maintaining and operating a search engine is neither easy nor cheap. Something along the lines of our proposal could become a suitable middle ground. Traditional search engines could act as hosts, providing their index and computational resources. The final ranking, however, could be driven by a community of people maintaining a large and open collection of *Goggles*.

The underlying idea behind *Goggles* is simple, borderline trivial. As a matter of fact, related concepts have been proposed in the past [12], however, unless it is coupled with a search engine infrastructure, the chances of success are small. Custom rerankers are only one side of *Goggles*. Performing a rerank, depends both on the rules of reranking but also on the original result-set where the rules will be applied. Hence, the effectiveness of the system is predicated on obtaining a large set of results on which the rules can be applied. Without the active collaboration of a search engine provider, such large result-set is not available. Top 10 results or top 50 in the case of Bing API [13] are not nearly big enough. Of course, scraping is always a possibility, but latency will become an unsolvable issue. It would take a few seconds to scrape the first 100 results out of a search engine, if we manage to not get blocked. And still, a result-set of 100 results, while better than 10, is still way too small. The only way to efficiently implement something like *Goggles* is with the collaboration of a search engine which allows the user to send a custom re-ranking function to be applied to the first set of results (typically in the tens of thousands) rather than on the final steps where the candidate result-set has already been reduced enough to have a poor overlap with the user custom re-ranking. In Section  we briefly describe how the *Goggles* language is applied to Brave's search ranking algorithm.

## INTEGRATING WITH EXISTING SEARCH ENGINES

Modern search engines have strict latency requirements, usually less than a second, in which they need to respond to the user query. A common way to architect a search engine to address this issue is to split the process into multiple phases. The recall phase involves matching the user query against billions of (in some cases, a lot more) pages with simple features to help reduce a candidate set to a reasonable size for further processing, typically in the order of few thousands. Subsequent phases, usually known as precision phases, narrow down the candidate set using a stack of increasingly sophisticated and costly models. The last phase of this process, the ranking, involves a very small candidate-set and is the one responsible for the final ordering of results given to the user.

The effectiveness of *Goggles* increases the earlier they are integrated into the search process so that more pages can be subjected to the rules being applied. Consider the Goggle "Filter out the results from the top 1000 domains on the internet", which could be an interesting way to explore the internet. Applying this on the final result set for most queries would lead to very few results, if any, due to the inherent bias in most search engines to surface content from popular domains. The rules defined by *Goggles* are better applied to the largest candidate-set possible, so that the intersection between candidates and rules to be applied is not empty. Only when intersection is large enough, will the re-ranking introduced by *Goggles* be noticeable.

Deep integration between *Goggles* and the host search engine is needed for the system to work. However, such integration poses different issues: 1) *Efficiency:* applying the rules against all elements of the candidate set (typically URLs) has to be extremely fast to minimize the overhead. In the following section we will present our solution to this issue. And 2) *Independence:* the host search engine needs to have total control over their index. This trait is given on search engines running their own fully-fledged index, e.g. Google, Bing, Yandex, Baidu and Brave. However, other search engines that rely totally or in part on external indexes might not have the ability to pull a large enough candidate-set to perform the user re-rank defined on his Goggle. DuckDuckGo, Qwant and Ecosia, which rely on the Bing API, are limited to whatever the API offers.

In this paper we lay down the language and the supporting matching engine, however, integrating such system into the code of a large-scale search engine is non-trivial. We commit Brave search engine to do so, to be a host for *Goggles*. We believe and welcome other search engines to also be hosts, after all, the more choices of *Goggles* and of hosts search engines, the better.

## LANGUAGE FOR *GOGGLES*

For the purpose of *Goggles*, we created a DSL (Domain Specific Language) which will allow users to express rules able to capture flexible filtering logic applied on a large set of search results. This DSL needed to be plain text and self-contained to ease hosting and sharing, flexible enough to express fine-grained filtering logic of URLs and page features, yet sufficiently constrained so that filtering can be implemented in a very efficient way (as mentioned previously, this system needs to be able to match thousands of candidate results against thousands of rules for each user query, without impacting latency in a perceivable way). Finally, it needed to be accessible enough so that even people without a technical background could quickly grasp its syntax and write rules, which would also encourage collaboration around the creation and curation of *Goggles* (e.g. communities).

After considering all these requirements, we realized that we could leverage prior work, addressing a totally different use-case but sharing similar challenges. We decided to base our DSL upon a subset of the syntax used by con-

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044986



tent blockers to perform "network filtering" (i.e. ads- and trackers-blocking): the so-called "AdblockPlus filters syntax" [1, 21]. This language already proved in the past that it, 1) allows to express logic to target URLs in a powerful way, 2) can be implemented extremely efficiently [38], and 3) is friendly to contributors and gave rise to numerous communities maintaining lists with a robust open collaboration model [16, 15, 35].

The language is also already widely documented, is flexible enough to allow custom extensions while maintaining backward compatibility (e.g. new options can be added without breaking other engines). This last point is especially important since we hope that other search engines will follow suit and also adopt support for *Goggles*. It was observed in the content-blocking communities that, in practice, maintainers have an incentive to keep compatibility with a maximum number of engines, and will thus use the features which are widely supported in priority (common denominator) and rely on engine-specific features only if they cannot do otherwise; this allows some flexibility for engines implementing custom extensions to the language.

We now give a brief overview of this language, the draft spec of which will be hosted publicly and open for participation in the future.

A list of filters, or Goggle, is a self-contained text file where each line can contain a filter (empty lines or comments—line starting with a '!' character—are ignored). Ranking of search results will be altered based on the filters contained in the file. Each filter is composed of two parts: a *trigger* and an *action*, separated by a *$* character: *<trigger>$<action>*. The trigger part is a pattern which needs to match a result candidate. It can leverage the following features:

- **Plain Patterns**—allow targeting a URL (or another result attribute like its title) based on a string of characters which it should contain. The filter "/coronavirus-" would trigger on any URL containing this specific string of characters (e.g. https://example.com/coronavirus-update.html).
- **Wildcard Patterns**—extend plain patterns with globbing capabilities: the special symbol "*" can be used to match any number of characters. Filter "/health/*/coronavirus-" would match any URL containing the substring "/health/", followed by zero or more characters, then "/coronavirus-" (e.g. https://example.com/health/2020/coronavirus-update.html).
- **Left and Right Anchors**—introduce a special "|" character which, when appearing at the start or end of a filter, forces a pattern to match the beginning or end of a URL. Filters "|https://" and ".html|" would match URLs starting with *https://* or ending with *.html*, respectively.

Each filter can also be annotated with additional options (following the "$" character). Multiple options can be specified at the same time, and separated by comas. We leverage this syntax to add ways to further fine-tune the behaviour of *Goggles*; either to specify which features of a result candidate should be considered (i.e. *target*), or how the ranking should be affected (i.e. *action*). For example:

- **$boost=XX**—is used to alter the ranking of specific results by *XX* (e.g. *$boost=1* would not alter the ranking, while *$boost=2* would make a result two times more important).
- **$discard**—completely drops candidates from the list of results.
- Filtering based on specific attributes of the result page can be achieved with:
  - **$lang=XX**—to target the language.
  - **$inurl**—to target the URL.
  - **$inquery**—to target queries leading to a candidate.
  - **$intitle**—to target the title.
  - **$indescription**—to target the description.
  - **$intext**—to target the full content.

Last but not the least, these features can all be combined to form complex filters. For example, the filter: */news/*/covid.html|$inurl*, would match candidates based on their URL.

This description is by no mean complete or final, and we will release a specification of the language once it is stabilized.

### *Protocol*

To allow users and communities to create and curate *Goggles* over time, we propose the following protocol, inspired by the most successful filters maintainers from content-blocking communities.

We propose two modes operations for maintainers: 1) A development setup implemented as a Web User Interface which allows to quickly get feedback over newly created filters, by showing which results end up in the final result set in real time. This setup is intended to speed-up the process of creating filters, reducing friction and offering a seamless workflow. The resulting filters can then be hosted publicly on a platform such as GitHub and made available to a wider public. And 2) The production setup which is directly integrated into any search engine prepared to be a host for *Goggles*. The end user could specify a link (or identifier) to the *Goggle* in the form of network accessible URI. The search backend is then responsible for fetching the Goggle definition from the URI (or use a cached version of it), compiling it to an efficient representation optimized for matching speed, and applying it at the recall-phase to the search results to produce a resulting candidate set.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044986



### *Privacy Considerations*

It is important to consider the potential privacy implications of sending a *Goggles* URIs together with the query. The URI can become a unique user identifier, especially for those people using non-popular *Goggles*. Therefore, there is a risk of a host search engine building a partially complete user profile in some circumstances. This should not be a problem for all host search engines, though; Google and Bing for instance, link all queries to the users' accounts and consider it a desirable feature. However, for privacy preserving search engines like Brave, this becomes a hurdle.

Note, however, that the URI only doubles as a user identifier under certain conditions: 1) when a user is consistently using it for all queries, and 2) when the URI is only used by that user (or a very small group of users). None of these conditions should be the default *modus operandi* of *Goggles*. We would expect *Goggles* to be used only for a fraction of queries. Also, we expect users to rely on multiple *Goggles* for different tasks. And finally, we expect a great majority of users to rely on popular *Goggles*, for which the URI is not a valid user identifier. Reality, however, does not need to conform with expectations. We should provide an additionally mechanisms to protect privacy for those niche cases. One proposal would be to allow sending multiple *Goggles* URIs on a single query, so that the true *Goggle* is obfuscated on a larger set. The host search engine would return results for all the *Goggles* and on the client-side the results for the padding *Goggles* would be dropped. This approach, however, imposes a serious overhead on the host search engine. The final solution to this problem is left for future work.

## PERFORMANCE EVALUATION

As previously discussed, *Goggles* can only shine when applied to a very large candidate set of results (thousands of URLs). For this reason, the filtering logic can only take place in the search backend, during the recall phase. Consequently, we operate under a very tight time budget (few milliseconds) to ensure that the overall search latency is still acceptable and that the backend remains able to handle many concurrent requests from users.

To assess the viability of *Goggles* from a performance perspective, we first implemented a prototype leveraging our in-house high-performance JavaScript content blocking library [8], then a custom Rust re-implementation of a similar engine, tuned for performance. The following figures were obtained by sampling 10k results with query "coronavirus" from our search index. The filters used were a selection of 1000 domains from the most popular domains, which we use as a "trustworthy list of domains"-*Goggle*. We run the measurements with varying number of URLs and filters to get insights into how the total time evolves as a function of the input size. Results are summarized in Table 1. These measurements were performed using our Rust prototype, compiled with rustc 1.43.1, on a reasonably fast ultrabook CPU (i7 U6600) using two cores (4 logical threads using hyper-threading).

| Number of URLs | Number of filters | Time (ms) |
|---|---|---|
| 1000 | 1 | 0.17 |
| 1000 | 10 | 0.20 |
| 1000 | 100 | 0.24 |
| 1000 | 1000 | 0.33 |
| 10000 | 1 | 1.56 |
| 10000 | 10 | 1.78 |
| 10000 | 100 | 2.08 |
| 10000 | 1000 | 3.10 |

Table 1: Summary of the performance evaluation (time in milliseconds) for different number of URLs and filters.

From these results we can conclude that our initial Rust prototype is already delivering good performance on a reasonably large set of candidate URLs (note that recall phase is typically sharded across multiple servers, so the aggregated candidate set could be much larger). The figures obtained from our reference implementation give us confidence about the feasibility of the approach, even on the rare case of a single server. Secondly, we observe that the processing time per-request is almost constant thanks to the efficient dispatching data-structure used in the filtering engine [38]; this shows that *Goggles* could be handling many more filters while still meeting our time budget; the runtime being almost exclusively impacted by the number of URLs in the initial result-set (assuming the filtering runs on a single CPU). Digging further, we observed that pre-processing of URLs, which consists of extracting the hostnames as well as tokenizing the URL, is the current bottleneck with a total of 70% of the overall time spent, whereas looking up filters from the index only takes around 10% of the total time. This shows that we could improve the performance drastically by focusing our effort on these two functions.

## CONCLUSION

We believe that the system/framework proposed in this paper would be beneficial to maintain a healthier Web. *Goggles* would foster openness and diversity thanks to the community maintenance and ownership. The later being very important as the added value created should not exclusively be in control of the host search engine, or else we might end up on the current status-quo. Besides, community *Goggles* also requires the active participation on a host search engine, which would provide access to its index and infrastructure. We are happy to commit Brave search to this endeavor, as we did with the now defunct Cliqz search[3].

Needless to say that *Goggles* will be open to any other search engine or institution that is enticed by this proposal.

## ACKNOWLEDGEMENTS

*Goggles* would not have been possible without the contribution of Cliqz search engine, which was shutdown in

[3] See Acknowledgments

https://doi.org/10.5281/zenodo.6044986

early 2020. Fortunately, some Cliqz team members and core intellectual property are now part of Brave, who is happy to continue, and extend, the mission of building an alternative search outside of Big Tech.

## REFERENCES


[1] AdblockPlus. *AdblockPlus filters explained*. URL: https://adblockplus.org/filter-cheatsheet.

[2] Ori Ben-Yitzhak et al. "Beyond basic faceted search". In: *Proceedings of the 2008 international conference on web search and data mining*. 2008, pp. 33–44.

[3] George Berkeley. *A treatise concerning the principles of human knowledge*. JB Lippincott & Company, 1881.

[4] K Boudreau, Lars Bo Jeppesen, and Milan Miric. "Competing on free(mium): digital competition with network effects". In: *SSRN Electron. J* 10 (2019).

[5] Pew Research Center. *Many Americans Believe Fake News Is Sowing Confusion*. 2016. URL: https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/.

[6] Simone Chambers. "Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere?" In: *Political Studies* 69.1 (2021), pp. 147–163. DOI: 10.1177/0032321719890811. eprint: https://doi.org/10.1177/0032321719890811. URL: https://doi.org/10.1177/0032321719890811.

[7] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2018, pp. 224–232. ISBN: 9781450359016. DOI: 10.1145/3240323.3240370. URL: https://doi.org/10.1145/3240323.3240370.

[8] Cliqz. *Efficient embeddable adblocker library*. URL: https://github.com/cliqz-oss/adblocker.

[9] Hacker News Community. *We can do better than DuckDuckGo*. 2020. URL: https://news.ycombinator.com/item?id=25129431 (visited on 2020).

[10] Australian Competition and Consumer Commission. *Digital Platforms Inquiry*. 2019. URL: https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report.

[11] Saoirse De Paor and Bahareh Heravi. "Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news". In: *The Journal of Academic Librarianship* 46.5 (2020), p. 102218.

[12] Google Developers. *Custom Ranking*. 2020. URL: https://developers.google.com/custom-search/docs/ranking.

[13] Microsoft Docs. *Custom Search API v7 reference*. 2017. URL: https://docs.microsoft.com/en-us/rest/api/cognitiveservices-bingsearch/bing-custom-search-api-v7-reference.

[14] Elizabeth Dubois and Grant Blank. "The echo chamber is overstated: the moderating effect of political interest and diverse media". In: *Information, Communication & Society* 21.5 (2018), pp. 729–745. DOI: 10.1080/1369118X.2018.1428656. eprint: https://doi.org/10.1080/1369118X.2018.1428656. URL: https://doi.org/10.1080/1369118X.2018.1428656.

[15] EasyList. *EasyList filter subscription*. URL: https://github.com/easylist/easylist.

[16] EasyList. *EasyList forum*. URL: https://forums.lanik.us/.

[17] Eric Goldman. "Search engine bias and the demise of search engine utopianism". In: *Yale JL & Tech.* 8 (2005), p. 188.

[18] Ruth A. Harper. "The Social Media Revolution: Exploring the Impact on Journalism and News Media Organizations." In: *Inquires Journal* (2010).

[19] Edward S. Herman and Noam Chomsky. *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon Book, 1988.

[20] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. "Can social bookmarking improve web search?" In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. 2008, pp. 195–206.

[21] Raymond Hill. *uBlock Origin static filter syntax*. URL: https://github.com/gorhill/uBlock/wiki/Static-filter-syntax.

[22] Ryan Holmes. *The Problem Isn't Fake News, It's Bad Algorithms—Here's Why*. 2016. URL: https://observer.com/2016/12/the-problem-isnt-fake-news-its-bad-algorithms-heres-why/.

[23] House Committee on Judiciary. *Investigation of Competition in Digital Markets*. 2020. URL: https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf (visited on 2020).

https://doi.org/10.5281/zenodo.6044986



[24] Stephen Lawrence. *Personalization of web search*. US Patent App. 10/676,711. Mar. 2005.

[25] Leonard Mlodinow and Stephen Hawking. *The grand design*. Random House, 2010.

[26] Efrat Nechushtai. "Could digital platforms capture the media through infrastructure?" In: *Journalism* 19.8 (2018), pp. 1043–1058. DOI: 10.1177/1464884917725163. eprint: https://doi.org/10.1177/1464884917725163. URL: https://doi.org/10.1177/1464884917725163.

[27] Raymond S Nickerson. "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of general psychology* 2.2 (1998), pp. 175–220.

[28] Jean-Christophe Plantin et al. "Infrastructure studies meet platform studies in the age of Google and Facebook". In: *New Media & Society* 20.1 (2018), pp. 293–310.

[29] Christian Bauckhage Robert Wetzker Carsten Zimmermann. "Analyzing social bookmarking systems: A del. icio. us cookbook". In: 2008.

[30] Dietram A. Scheufele and Nicole M. Krause. "Science audiences, misinformation, and fake news". In: *Proceedings of the National Academy of Sciences* 116.16 (2019), pp. 7662–7669. ISSN: 0027-8424. DOI: 10.1073/pnas.1805871115. eprint: https://www.pnas.org/content/116/16/7662.full.pdf. URL: https://www.pnas.org/content/116/16/7662.

[31] Sindre Sorhus. *Awesome lists about all kinds of interesting topics*. URL: https://github.com/sindresorhus/awesome.

[32] Miriam Steiner et al. "Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries". In: *Information, Communication & Society* 0.0 (2020), pp. 1–25. DOI: 10.1080/1369118X.2020.1776367. eprint: https://doi.org/10.1080/1369118X.2020.1776367. URL: https://doi.org/10.1080/1369118X.2020.1776367.

[33] Joshua Tucker et al. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature". In: *SSRN Electronic Journal* (Jan. 2018).

[34] Daniel Tunkelang. "Faceted search". In: *Synthesis lectures on information concepts, retrieval, and services* 1.1 (2009), pp. 1–80.

[35] uBlockOrigin. *Resources for uBlock Origin, uMatrix: static filter lists.* URL: https://github.com/uBlockOrigin/uAssets.

[36] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151. ISSN: 0036-8075. DOI: 10.1126/science.aap9559. eprint: https://science.sciencemag.org/content/359/6380/1146.full.pdf. URL: https://science.sciencemag.org/content/359/6380/1146.

[37] Paul Wellstone. "Growing media consolidation must be examined to preserve our democracy". In: *Fed. Comm. LJ* 52 (1999), p. 551.

[38] WhoTracks.Me. *Adblockers Performance Study*. 2019. URL: https://whotracks.me/blog/adblockers%5C_performance%5C_study.html.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5887680

# PRIVACY IN OPEN SEARCH: A REVIEW OF CHALLENGES AND SOLUTIONS

Samuel Sousa*[1], Roman Kern[1], Know-Center GmbH, Graz, Austria
[1]also at ISDS, Graz University of Technology, Graz, Austria
Christian Guetl, ISDS, Graz University of Technology, Graz, Austria

*Abstract*

Privacy is of worldwide concern regarding activities and processes that include sensitive data. For this reason, many countries and territories have been recently approving regulations controlling the extent to which organizations may exploit data provided by people. Artificial intelligence areas, such as machine learning and natural language processing, have already successfully employed privacy-preserving mechanisms in order to safeguard data privacy in a vast number of applications. Information retrieval (IR) is likewise prone to privacy threats, such as attacks and unintended disclosures of documents and search history, which may cripple the security of users and be penalized by data protection laws. This work aims at highlighting and discussing open challenges for privacy in the recent literature of IR, focusing on tasks featuring user-generated text data. Our contribution is threefold: firstly, we present an overview of privacy threats to IR tasks; secondly, we discuss applicable privacy-preserving mechanisms which may be employed in solutions to restrain privacy hazards; finally, we bring insights on the tradeoffs between privacy preservation and utility performance for IR tasks.

***Index terms—*** Privacy, information retrieval, personal information, open search.

## INTRODUCTION

Data is the cornerstone of many research fields, as well as the source of information used by professionals, such as journalists, statisticians, and police makers [1]. At the pace of the Internet popularization, the number of data publishers rose to the thousands [2], encompassing university library files, government open data, Wikipedia articles, social media platforms, commercial data providers, digital data markets, scientific data repositories, among many others. Accessing these data sources is crucial for solving the reproducibility issues of scientific results and improving the means for data journalists to obtain reliable information [1]. Furthermore, the idea of open science relies on sharing research data and materials for third-parties to reuse and reproduce experiments, assuring the trustworthiness of the research process [2]. Some companies also need to have their products openly visible on Internet for their business to keep going, e.g., e-commerce platforms. However, privacy issues may keep some kinds of data from being released in order to avoid the exposure of confidential information. As an example, user-generated data, such as user behavior, search interests, and search profiles, demands anonymization steps prior to its release. Otherwise, it should remain unrevealed. Other privacy issue regards the level playing field for search services since those which monitor the intents of users can eventually obtain benefits out of it.

Privacy is a concept related to limiting the extend of information an individual is willing to share [3]. Many countries consider privacy as a right protected by law. For instance, the General Data Protection Regulation (GDPR) [4], which entered into force in May 2018 in the European Union (EU), draws guidelines for the collection, transfer, storage, management, and deletion of personal data within the member states of the economic bloc. GDPR grants the residents of the EU the control over their personal data. Therefore, any processing activities over personal data must comply with the regulation and provide protection measures. In case of data breaches or noncompliance with the legal principles, penalties and fines are applicable. Since 2018, data protection bills have also been passed in several countries worldwide.

In the recent past years, the preservation of privacy has been gaining attention in the field of information retrieval (IR) [5–8]. Several privacy-preserving mechanisms have been developed to safeguard personal data from threats, as attacks, disclosures, and unintended usages. Some of these mechanisms, such as encryption [9, 10], differential privacy (DP) [11], multi-party computation (MPC) [12], and federated learning (FL) [13, 14], are implemented alongside models for enabling applications to safeguard data privacy. There are also some attacking methods that aim at retrieving data samples used to train models, mostly neural networks, which can be seen as a threat or a safety checker, depending on the attacker's intention. Furthermore, some privacy-preserving tools may present computational overheads or performance reductions, which call attention to privacy-utility trade-offs.

Searches on personal data often present privacy risks from both data provider and model sides. There is a need for open data, due to scientific, governmental, and press reasons [1,2], however data generators and IR systems users cannot have their identities and personal data exposed. This work aims, therefore, at pointing out open challenges with regards to privacy in IR tasks, as well as reviewing appropriate privacy-preserving methods to safeguard personal data or model. We focus primarily on tasks featuring text data since the private content of data in written format may be presented explicitly, as a person's name or an ID number, or implicitly like a the gender of a person which can be inferred from the text, based on gendered words like profession terms.

* ssousa@know-center.at

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5887680



Our contributions comprise a summary of research directions on privacy for IR tasks alongside adequate privacy-preserving methods for the privacy risks these tasks may present. Problems that put privacy at risk are also discussed. Moreover, we discuss how privacy-preserving methods can influence the results of IR tasks. Finally, we provide the readers with essential understanding of privacy-related issues and privacy-preserving methods for IR.

This paper is organized as follows. Firstly, we review recent works in the literature of privacy in section *Related Work*. Secondly, we describe open challenges and solutions for privacy preservation in open search tasks in section *Privacy Challenges and Solutions*. Further, we discuss our results in the section *Discussion*. Finally, our contributions and upcoming works are brought into context in the section *Conclusion and Future Works*.

## RELATED WORK

Data privacy is a large concern within IR [6–8] since privacy threats and risks often arise from searches on private data [6] and tasks which involve the search behavior or intent of users [7, 8, 15]. Examples of such issues include private data publishing [6], string searches [16], user's context [8], information sharing of web search logs [17], interactive search [15], and the information selection behavior of users [15]. Additional privacy threats include revealing private data from IR systems that compute distributed information, performing face recognition without the consent of the people whose faces are captured, and so on. Therefore, a large number of real-world IR systems can be prone to breaches of personal data, unintended disclosures of private information, and penalties established by data protection regulations.

Nowadays, many organizations collect and process personal data, such as governments, companies, and search service providers. As a consequence, the volume of collected data has increased alongside the risks related to breaches of sensitive information from these datasets. Zhu et al. [6] survey DP applications for data publishing and data analysis. The authors define data publishing as publicly sharing data itself or the result of queries, whereas data analysis consists on releasing data models to the public. In both scenarios, DP can offer privacy guarantees resistant against attacks and mathematically provable. Riazi et al. [16] come up with a mathematically provable mechanism for privacy preservation in IR tasks. The authors implement a methodology for two-party string search based on the Yao's Garbled Circuit protocol. For instance, two users can hold queries and data for string search simultaneously, whereas they both aim at keeping these search components private without relying on a third party like a trusted server. Therefore, the proposed protocol converts the search algorithm into a Boolean circuit that evaluates the private queries and texts. Tamine and Daoud [8] survey methodologies and metrics for context IR evaluation, specifying the impact of data privacy towards the evaluation design.

User search behavior can also be seen as private information since queries and search result's selection can disclose demographic attributes, preferences, political views, etc. Orso et al. [15] investigate the role of user search behavior and information selection in order to understand which layers of social information, namely personal preferences, tags combined to personal preferences, or tags and social ratings combined with personal preferences, can enhance search efficiency. The authors found empirical evidence that personalized preferences and social ratings make it easier for users to select information without external sources. In this study, the publicly available Yelp dataset[1] was used. Therefore, privacy aspects that would arise from this use case in a real-world scenario, as the compliance with data protection regulations, were not approached. Sharing search log records is a process ruled by data protection laws, which demand anonymization techniques to be applied before the data leaves the servers it is stored on. Mivule et al. [17] introduce an heuristic for privacy preservation of individual web search log records based on swapping. Firstly, an individual has a set of logs $\mathcal{A}$. Secondly, the records in $\mathcal{A}$ are switched within this set. Finally, the swapped files from $\mathcal{A}$ are switched again using records of a set $\mathcal{B}$. This heuristic is efficient when it comes to preventing an attacker from tracing the issuer of a search query. Additional privacy-preserving IR solutions include homomorphic encryption (HE) [18], DP [19], and hashing [7].

Our work focuses on identifying privacy risks across IR tasks, alongside suggesting privacy-preserving mechanisms that have the potential to suit the needs for privacy protection. We briefly introduce the task description followed by the privacy risks associated with both data and IR model.

## PRIVACY CHALLENGES AND SOLUTIONS

This section firstly introduces open challenges with regards to privacy in IR tasks. We survey recently published works for ten IR tasks, highlighting their privacy issues. Afterwards, we overview privacy-preserving methods that are suitable options to address these problems.

### *Privacy Challenges*

#### Search tasks

*Ad-hoc search.* Modern search engines often rely on bag-of-words models to represent documents and search queries. Consequently, accurate quantification of context-specific term importance in documents is a tricky problem since term's context is often not captured by these models. When it comes to data privacy, bag-of-words models pose risks related to data re-identification. For instance, some terms in the vocabulary of the model may refer to personal identities or private attributes, such as age, gender, location, and demographic information, which allow a de-identified document to be re-identified. Dai and Callan [20] propose a novel

[1] `https://www.yelp.com/dataset`.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5887680

document term weighting framework which preserves word context, using BERT [21] embeddings to extract document representations. Although keeping document's contextual information, BERT embeddings can also encode private information and suffer attacks as model inversion, reverse engineering, and membership inference.

*Query expansion.* Query expansion (QE) helps users of IR applications to find more relevant information by expanding the search queries, hence increasing recall. For instance, synonyms and hypernyms of terms in the questions for question answering (QA) are used to rise the likelihood of matching sentences stating the most appropriate candidate answer [22,23]. However, in some QA scenarios the number of retrieved sentences may be small, and then mismatch the intents of the user who queries the QA system [23]. Common privacy threats in applications which use QE regard disclosing query content or private information in the documents in which the sentences are extracted, such as names and locations of people.

*Feature extraction for ranking.* In the context of ad-hoc search, document ranking consists on returning a ranked list of documents from a large collection based on the assumed information need expressed by a search query, maximizing some ranking metric like average precision [24]. Ranking is a cornerstone for IR systems and search engines, which can also be performed by machine learning (ML) classification models [24, 25]. Therefore, in order to reduce computation time, boost learning results, and prevent overfitting for those models, feature extraction techniques can be implemented to better represent documents. Pandey et al. [25] come up with a method for representing documents as matrices with reduced dimensions when compared to the original document representations. Such representations are useful for improving the results of ranking algorithms. However, feature extraction models can have its original training samples disclosed by attacks of model inversion, membership inference, or reverse engineering.

*Online learning for ranking.* A critical drawback of ranking models regards the hardness to obtain labelled data for model training. Thus, Zhuang and Zuccon [26] propose a counterfactual learning to rank method based on logs of user interaction collected from the ranking model in production. In a real-world setting, users would be able to confirm the effectiveness of the ranking model by clicking on the results. In the experimental evaluation, the authors use publicly available datasets and generate user clicks automatically with a cascade click model [27]. This scenario can pose privacy threats to personal data collection if the user clicks are collected from actual system users. Therefore, regulations for personal data collection would be applicable alongside the need for privacy-preserving methods.

*Query composition.* ML algorithms can be used to predict query properties like answer size, run-time, and error class [28]. These algorithms can therefore be prone to unintended memorization of query content alongside the attacks which aim at disclosing training document samples. Another privacy threat regards the location of the ML model, e.g., on a cloud server, since computation parties sometimes may not be trusted, or the communication channels for transferring data or model updates may allow eavesdropping attacks to take place.

### Healthcare tasks

*Healthcare data tasks.* Electronic medical records or electronic health records are digital documents, in which medical staff inputs patient data, including personal information, health condition, disease diagnosis, medication, etc. [29]. This type of document has the advantages of easy storage, transfer, sharing, and deletion. However, healthcare data is inherently private due to the sensitivity of its content. Therefore data protection regulations, as the EU's GDPR [4], prevents publicly releasing such data for research activities and public searches, without the explicit consent of data owners and the application of data anonymization methods. In the context of IR, medical applications have to safeguard medical data from queries by malicious users or computation parties, e.g., corrupted servers.

### Social media tasks

*Opinion mining.* Many online data sources contain opinions, which can be classified with regards to their polarity. For instance, Nguyen and Nguyen [30] predict sentiment polarity on YouTube[2] comments in English and Italian languages, using a DL model based on a Bidirectional Long Short-Term Memory architecture. Tasks performed over user-generated data pose privacy risks of unintended data memorization by the neural network, as well as sensitivity to attacks that aim at retrieving the model training samples. Therefore, there is a room for privacy-preserving mechanisms that prevent such privacy issues.

*Advisor for hashtag sharing.* On social media platforms, users often put their privacy at risk unconsciously by releasing details of their personal lives publicly, revealing their exact location, and posting political or societal views. Furthermore, some features of such platforms like hashtags may induce privacy threats, mainly related to location, since attacking models can easily predict precise user location from the hashtags they post online [5]. This situation can be prevented by privacy-preserving mechanisms of data obfuscation or de-identification, which get rid of privacy-sensitive information before any sharing step by the user. Moreover, social media data is frequently protected by data protection regulations, then, ensuring data privacy to be safeguarded.

[2] `http://youtube.com`.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5887680



*Social media profile linking.* User identity linking is the task of connecting accounts owned by the same user across different social media platforms [31]. For instance, a user can have profile simultaneously on TikTok[3], Twitter[4], Instagram[5], and so on. However, more relevant advertisement can be suggested to this user by linking these different profiles. This scenario, therefore, presents privacy threats related to monitoring user behavior online besides linking profiles where the user uses a pseudonym to remain anonymous.

**Recommendation tasks**

*Recommender systems.* Recommender systems aim at predicting the preferences of users based on their interest, making more effective use of information [32]. However, user interest and their search history should be preserved from leakages, hence these are pieces of private information. A recent application of recommender systems is news recommendation. Qi et al. [33] propose a framework to recommend news to users in distributed computation scenarios, e.g., smartphones running a mobile application. This frameworks computes updates locally on each distributed device separately and, then, sends these local updates to a centralized server which updates the global model by aggregating the local updates. Finally, the updated global model is distributed over the distributed devices to news recommendation and successive updates. This is an outstanding application of FL, which prevents threats of data leakage from the distributed devices. However, this model still can suffer from unintended memorization of user's behavior, preferences, and search history.

## *Solutions*

Many privacy-preserving methods have been proposed for the sake of preventing attacks and unintended data breaches. In this subsection, we review computational techniques for privacy preservation, which can be conveniently integrated into IR tasks. We group these techniques according to their technical aspects into cryptographic approaches and ML-based applications.

**Cryptographic approaches** Cryptographic protocols have been extensively used to protect private data when sharing activities are not advised. In a nutshell, encryption consists on the application of a function over data in raw format, which is referred as '*plaintext*'. This function results on an output called '*cyphertext*', which inhibits the identification of its original format or content.

*Encryption.* Homomorphic encryption (HE) is a form of encryption that allows arithmetic operations to be computed over cyphertexts without the need for decryption [9]. For instance, ML models can be used for inference on encrypted data, whereas the results are still consistent. Encryption schemes that implement HE are particularly useful for scenarios in which data transfers and centralized storage occur, e.g., cloud servers, alongside the lack of trust. However, it is worth to mention the computational overheads that FHE often leads to, hence its use becomes prohibitive for devices with small memory capacity.

IR models can also be privacy-preserving by the use of searchable encryption (SE). This technique encrypts document collections enabling the data owner to delegate search capabilities, whereas the server or a service provider, like a search engine, does not demand decryption [10]. Therefore, the so-called 'honest-but-curious' server can provide searches, whereas the content of the stored data and the input queries is preserved [34]. On the other hand, semantic relations between words and documents may be lost in the encrypted forms, so that decaying search results [35]. SE usually encompasses algorithms for the steps of key generation, encryption, token generation, and search [34]. Finally, the main threat this encryption scheme faces is related to keyword inference attacks that aim at recovering the content of encrypted keywords.

*Multi-party computation.* Document collections may be stored in distributed search system hosted by other companies or even distributed across members of a broad community. Therefore, when the exchange of documents among the members of these computation scenario is not an advisable option, multi-party computation (MPC) can be successfully used. MPC is a cryptographic primitive that computes aggregated functions over multiple sources of data, which cannot be revealed [12]. Formally, MPC assumes a set of inputs $\{x_1, x_2, \ldots, x_n\}$ so that each party $P_i$ will store $x_i$ and agree to compute $y = \hat{f}(x_1, x_2, \ldots, x_n)$, in which $y$ is the output information to be released, and $\hat{f}$ is the agreed function on the entire input set [36]. The input set may be composed of keywords, documents, medical records, etc.

**Differential privacy** DP can be understood as a randomized function $\hat{k}$ which is applied to document collections or query results prior to their public release [11]. Therefore, for all subsets $S$ in the range of $\hat{k}$, and document collections $D$ and $D'$ differing on at most one element, $\hat{k}$ provides $\epsilon$-differential privacy if:

$$Pr[\hat{k}(D) \in S] \leq exp(\epsilon) Pr[\hat{k}(D') \in S]. \quad (1)$$

Approaches for DP exploit mathematical formalism to neutralize de-anonymization attacks and keep a lookout for membership inference attacks that may disclose the original documents in the collection. The function $\hat{k}$ adds random noise to any input query and, consequently, yields the response [11]. Every mechanism that satisfies $\epsilon$-DP will mitigate risks of leakages of private information from any individual element since its inclusion or removal form the collection would not turn the output significantly more or less likely [11]. DP provides privacy guarantees usually at the cost of performance and computational overhead. However, one of the

[3] https://www.tiktok.com/.
[4] https://twitter.com/.
[5] https://www.instagram.com/.

https://doi.org/10.5281/zenodo.5887680

main advantages of this method regards managing the trade-off between privacy and utility, finding the ideal value of $\epsilon$ which preserves data privacy and affects the model results at a controlled extent.

### ML-based approaches

*Federated learning.* FL is a methodology for training ML models in distributed computation scenarios proposed by Google [37]. This methodology prevents data from leaving its owner's device during the computations and addresses privacy threats related to training over private data [14]. The federated training consists on, firstly, distributing copies of a global model with pre-defined parameters, computing local updates on the distributed devices, sending these updates to the server for aggregation, updating the global model parameters, and sending the updated parameters of the global model to each distributed device [13]. In formal terms, FL assumes a model $\hat{m}$ with parameters $\Theta_{\hat{m}}$ which are stored in a matrix $M$. The model $\hat{m}$ is thus shared with a subset $T$ of $\eta$ clients, which will update $\hat{m}$ with their locally stored data at each training step $t \geq 0$ [13, 37]. Every client $i$ will send its update $H_t^i := M_t^i - M_t$ to the central server, which is responsible for aggregating client-side updates for updating global model [13]. FL has advantages for enabling distributed computations over devices with restrained memory, bandwidth, and computation power.

## DISCUSSION

A large number of IR systems are sensitive to privacy threats in scenarios which include personal information, search history, personal preferences, private documents, to name a few. Additionally, regulations as the EU's GDPR establish the guidelines for protecting user generated data from breaches and non-consented usages. As a consequence, many algorithmic methods for protecting data privacy have been proposed and integrated into IR systems. However, protecting privacy using such methods can mean coping with utility performance decays and computational overheads. Therefore, the choice for a convenient privacy-preserving method for an IR scenario has to take into account the target of privacy protection and the computational resources available.

In Table 1, we bring a summary of IR applications for the privacy-preserving methods surveyed in the section *Solutions*. Cryptographic approaches, such as HE, SE, and MPC, are suitable to scenarios in which the original content of documents or datasets should not be revealed to unauthorized parties or some of these parties are not trusted, e.g., malicious servers. DP provides formal privacy guarantees which can be employed on a myriad of applications, such as data anonymization and protecting results of queries and ML models against inversion or reverse engineering. Finally, FL can be successfully implemented for scenarios with distributed devices and limited resources for data sharing or prohibitions of data exchange.

Table 1: Applications of IR alongside suitable privacy-preserving methods.

| IR application | PP methods |
|---|---|
| Ad-hoc search | SE, MPC |
| Query expansion | HE, SE |
| Feature extraction for ranking | DP |
| Online learning for ranking | DP, HE |
| Query composition | SE |
| Healthcare data tasks | DP, HE, SE |
| Opinion mining | DP |
| Advisor for hashtag sharing | SE |
| Social media profile linking | HE, SE |
| Recommender systems | FE, MPC |

## CONCLUSION AND FUTURE WORKS

Privacy is a critical point for the development of IR systems which deal with personal, user generated, or sensitive data. In this work we overview recent developments in the IR literature, pointing out privacy issues and suggesting suitable privacy-preserving methods. Data types, IR tasks, and privacy-preserving method drawbacks are taken into account to provide the reader with essential understating of this research field. As future works, we aim to address the aforementioned challenges for Open Search use cases, as well as studying and discussing compliance with legal requirements, such as those of the EU's GDPR.

## REFERENCES

[1] D. Brickley, M. Burgess, and N. Noy, "Google dataset search: Building a search engine for datasets in an open web ecosystem," in *The World Wide Web Conference*, 2019, pp. 1365–1375.

[2] D.-L. Magazine, "The landscape of research data repositories in 2015: A re3data analysis," *D-Lib Magazine*, vol. 23, no. 3/4, 2017.

[3] A. F. Westin, "Privacy and freedom," *Washington and Lee Law Review*, vol. 25, no. 1, p. 166, 1968.

[4] E. Commission, "2018 reform of eu data protection rules," https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf, 2018, date: 2018-05-25, URL Date: 2019-06-17.

[5] Y. Zhang, M. Humbert, T. Rahman, C.-T. Li, J. Pang, and M. Backes, "Tagvisor: A privacy advisor for sharing hashtags," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 287–296.

[6] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, "Differentially private data publishing and analysis: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, 2017.

[7] L. Weng, L. Amsaleg, A. Morton, and S. Marchand-Maillet, "A privacy-preserving framework for large-scale content-based information retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 152–167, 2014.

https://doi.org/10.5281/zenodo.5887680




[8] L. Tamine and M. Daoud, "Evaluation in contextual information retrieval: foundations and recent advances within the challenges of context dynamicity and data privacy," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–36, 2018.

[9] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 169–178.

[10] D. Cash, P. Grubbs, J. Perry, and T. Ristenpart, "Leakage-abuse attacks against searchable encryption," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 668–679.

[11] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[12] Q. Feng, D. He, Z. Liu, H. Wang, and K.-K. R. Choo, "Securenlp: A system for multi-party privacy-preserving natural language processing," *IEEE Transactions on Information Forensics and Security*, 2020.

[13] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Work- shop on Private Multi-Party Machine Learning*, 2016.

[14] M. Chen, A. T. Suresh, R. Mathews, A. Wong, C. Allauzen, F. Beaufays, and M. Riley, "Federated learning of n-gram language models," pp. 121–130, 2019.

[15] V. Orso, T. Ruotsalo, J. Leino, L. Gamberini, and G. Jacucci, "Overlaying social information: The effects on users' search and information-selection behavior," *Information Processing & Management*, vol. 53, no. 6, pp. 1269–1286, 2017.

[16] M. S. Riazi, E. M. Songhori, and F. Koushanfar, "Prisearch: Efficient search on private data," in *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2017, pp. 1–6.

[17] K. Mivule, "Data swapping for private information sharing of web search logs," *Procedia computer science*, vol. 114, pp. 149–158, 2017.

[18] A. El-Ansari, A. Beni-Hssane, M. Saadi, and M. El Fissaoui, "Papir: privacy-aware personalized information retrieval," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2021.

[19] G. H. Yang and S. Zhang, "Differential privacy for information retrieval," in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 2017, pp. 325–326.

[20] Z. Dai and J. Callan, "Context-aware document term weighting for ad-hoc search," in *Proceedings of The Web Conference 2020*, 2020, pp. 1897–1907.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[22] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, "Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering," *Information Sciences*, vol. 514, pp. 88–105, 2020.

[23] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *Acm Computing Surveys (CSUR)*, vol. 44, no. 1, pp. 1–50, 2012.

[24] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 708–718.

[25] G. Pandey, Z. Ren, S. Wang, J. Veijalainen, and M. de Rijke, "Linear feature extraction for ranking," *Information Retrieval Journal*, vol. 21, no. 6, pp. 481–506, 2018.

[26] S. Zhuang and G. Zuccon, "Counterfactual online learning to rank," in *European Conference on Information Retrieval*. Springer, 2020, pp. 415–430.

[27] A. Chuklin, I. Markov, and M. d. Rijke, "Click models for web search," *Synthesis lectures on information concepts, retrieval, and services*, vol. 7, no. 3, pp. 1–115, 2015.

[28] Z. Zolaktaf, M. Milani, and R. Pottinger, "Facilitating sql query composition and analysis," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 209–224.

[29] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: a review," *Journal of healthcare engineering*, vol. 2018, 2018.

[30] H. T. Nguyen and M. Le Nguyen, "Multilingual opinion mining on youtube–a convolutional n-gram bilstm word embedding," *Information Processing & Management*, vol. 54, no. 3, pp. 451–462, 2018.

[31] A. T. Hadgu and J. K. R. Gundam, "Learn2link: Linking the social and academic profiles of researchers," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 240–249.

[32] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1235–1244.

[33] T. Qi, F. Wu, C. Wu, Y. Huang, and X. Xie, "Privacy-preserving news recommendation model learning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1423–1432. [Online]. Available: https://www.aclweb.org/anthology/2020.findings-emnlp.128

[34] H. Liu and B. Wang, "Mitigating file-injection attacks with natural language processing," in *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, 2020, pp. 3–13.

[35] X. Dai, H. Dai, G. Yang, X. Yi, and H. Huang, "An efficient and dynamic semantic-aware multikeyword ranked search scheme over encrypted cloud data," *IEEE Access*, vol. 7, pp. 142 855–142 865, 2019.

[36] R. Cramer, I. B. Damgård, and J. B. Nielsen, *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, 2015.

[37] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6170083

# SEARCHING ON HETEROGENEOUS AND DECENTRALIZED DATA: A SHORT REVIEW

Julius Möller*, Carl-von-Ossietzky University of Oldenburg, Germany
Dennis Jankowski, OFFIS Institute for Computer Science, Oldenburg, Germany
Axel Hahn, Carl-von-Ossietzky University of Oldenburg, Germany

*Abstract*

Within the last few years, the amount of recorded data has increased significantly. Information is collected from a wide variety of areas, such as healthcare, autonomous driving, or e-commerce. In addition, the recorded data is usually not stored centrally, but is rather distributed across various decentralized infrastructures. This complicates searching for the desired data. Several papers have already been published that discuss approaches for finding the requested information within heterogeneous and decentralized data architectures. To highlight the similarities and differences between the various approaches, this paper conducts an integrative literature review on search architectures that deal with heterogeneous and decentralized data. This is done by decomposing existing architectures in different layers: It was found that the identified architectures first abstract from the original technologies of the heterogeneous data sources, and then use different indexing strategies in combination with a search algorithm to find and present the queried information.

## INTRODUCTION

In an increasingly networked world, the value of data as a common good has significantly increased. New technologies and achievements in science and industry rely on data-driven workflows more than ever. However, as the amount of publicly available data is continuously growing, finding accurate pieces of information in large, distributed, or decentralized infrastructures has become a problem. This problem can not only be observed in the human-readable part of the internet, but is also present for publicly available databases, multi-media content or file repositories [1]. With the increased availability of Big Data in several areas in research and economy, being able to find the right sets of data has become more and more relevant for success. In this context, initiatives like the Open Search Foundation aim at providing open and independent approaches for fulfilling these tasks [2]. These approaches could play an important role for future developments and should not only deal with classical web data, but also with data from a variety of heterogeneous data sources. Nevertheless, web search and search on heterogeneous data sources is not contradictory: Common approaches combine both types of data, to provide intelligent query functionalities that rely on knowledge representations and intelligent indexes generated from heterogeneous data (cf. [3], [4]).

More openly developed search approaches rely on decentralizing the utilized infrastructure and incoming processing tasks to strengthen overall trust and to equally distribute processing power. Analogously, data from heterogeneous data sources can also often be found in decentralized structures. This means, that the data is not stored in a central location like, e.g., a data lake, but is distributed to different sources, typically controlled by different providers, and based on different technologies. This introduces new challenges to the design of search systems.

In this paper, we present the challenges of searching on heterogeneous and decentral data sources in comparison to conventional web search. Furthermore, a short integrative literature is conducted and common architectures for solving the identified challenges are examined on four different architectural levels.

## SEARCH ON HETEROGENEOUS DATA

Conventional web search architectures typically start with a crawler, that crawls web pages and saves them in a local copy with a pre-defined format (crawl dump) that is then processed by an indexer. The indexer analyses the saved data and maps possible search terms to data points in the crawl dump: this creates the index. Finally, a search algorithm is used, that takes input from a user of the search engine and uses the index to find search results (cf. Figure 1). These elements typically differ in their specific realization but can most often be found in conventional search engines. [5]

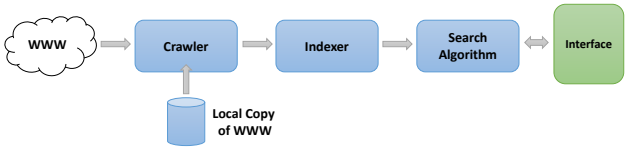



Figure 1: Classical Web Search Engine Architecture (cf. [5])

On the first sight, this process seems easily adaptable to use-cases with other types of data. However, some preconditions are required here, so that the architecture cannot be transferred to heterogeneous data sources without further effort: To start with, the web crawler makes use of the standardization of web content accessibility: The Hyper-Text Transfer Protocol (HTTP) is supported by any website, so that the crawler can access web-content in a standardized way [6]. When considering heterogeneous data sources, this is one of the first obstacles: A common interface to access arbitrary data sources cannot be assumed. Rather, the data sources come in a variety of different technologies, and different types of access-policies. Furthermore, it may not always make sense to create a complete local copy of available data sources, as there is no homogeneous data structure that could manage these and make them easily indexable as it can be done with

* julius.moeller@uni-oldenburg.de

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6170083



HTML documents. Like web-crawlers, which sometimes only store parts of the web data (e.g., filtering out image or video data) or meta-data, storing only processed information about the heterogeneous data sources can be a solution to this problem [7]. After that, web search engines profit from the fact, that most of their indexed content is given in standardized formats (html, xml, etc.) that are known before runtime [6]. These formats are then parsed by the indexer, which usually extracts the human readable text and generates keywords and metadata that map to the desired sections of the crawl dump [8]. As the formats of the data from heterogeneous data sources are not known before runtime, it is much harder to find suitable indexing strategies. Additionally, as heterogeneous data may also include non-human-readable information, the question arises if a simple keyword search is effective in such a case [9]. Another important difference is that it cannot be assumed, that heterogeneous data is linked in any way: Web crawlers can find web pages by following URLs and continuously extend their knowledge about areas of the public internet, which is typically not possible in multiple heterogeneous data sources. This can also become a problem, when utilizing metrics like “number of web pages referring to a specific web page” for search ranking algorithms, as it is often done in web search engines [10].

This problem of searching in heterogeneous and decentrally organised data has already been discussed in several scientific publications. However, to our best knowledge, there has been almost no effort to review the literature regarding the presented search architectures. The only work in this area, that could be identified was a paper by Wang et al. [11], who discuss the problem specifically from the perspective of search queries. Nevertheless, their work is restricted to the application in data spaces and does not discuss the architectural principles behind the different types of queries.

## REVIEW OF SEARCH ARCHITECTURES

To analyse common search architectures that deal with heterogeneous and decentral data, we conduct an integrative literature review: We manually search Elsevier Scopus and Google Scholar (cf. [12]) for the following search term, based on the discussion in the previous section:

*(indexing* OR *search)* AND *(heterogeneous data* OR *data architecture* OR *unstructured data* OR *decentral)*

To limit the large amount of search results, we only consider the 160 search results ranked most relevant for both search engines, leading to a total of 320 search results that are considered within this review. For our review, only publications were considered, that (1) are related to search in data, (2) either deal with heterogeneous data or decentral organization of data and (3) present an architecture. To identify publications that meet these criteria, out of the 320 search results, 113 were selected for further consideration based on their title. Then, after scanning their abstracts, 63 papers were identified, which were examined completely. Finally, and without duplicates, 33 papers were found, that fulfil the defined criteria.

In summary, most of the analysed architectures are based on the presence of heterogeneous data sources, that can contain structured, semi-structured or unstructured data. Secondly, some type of technology abstraction layer for the actual data sources is implemented to enable a unified access to them. As already stated, crawling components usually are not part of search architectures for heterogeneous data sources. Also, the understanding of an indexing strategy is extended to more generic representations than keyword indices. Even though several architectures provide keyword indexing, others completely rely on creating ontologies or other mappings. Several architectures can also deal with basic decentralization of data sources implicitly by their design, as they connect each data source individually to the rest of the system. The following section provides a detailed analysis of the different components in search architectures that deal with heterogeneous or decentralized data. An abstract model of

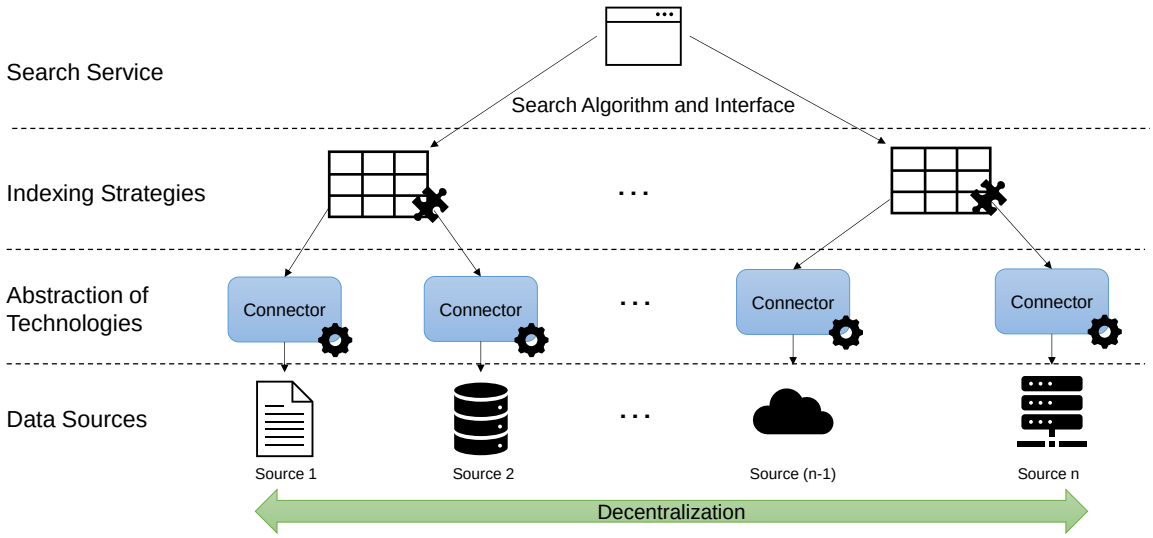



Figure 2: Architectural Framework for Search in Heterogeneous Data Sources.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6170083



these architectures is depicted in Figure 2: Starting with data sources that should be indexed, a technology abstraction layer is introduced to provide unified access. This layer is then used to provide the required data for an indexing strategy. Finally, results of the indexing strategy can be accessed by the search algorithm, which in combination with an interface represents the search service to a user. The following sub-sections present more details of the single components.

### *(Heterogeneous) Data Sources*

The data sources represent the basis of most architecture. Heterogeneous data source can contain structured data in databases (e.g., PostgreSQL, MySQL, MongoDB), semi-structured data (e.g., json, XML) or unstructured data (e.g., text files, images, video data). For many use-cases, this data is often not organized in a centrally controlled structure, but is often distributed to different single data providers, that also use different data models and access policies. Many of the identified search architectures can deal with arbitrary types of data sources. However, this requires manual integration procedures most of the time. On the other hand, some of the identified architectures are focused on specific types of heterogeneous data and therefore are often able to provide a more precise search functionality for the respective type of data. For example in [13] and [14], the authors specifically focus on multimedia data. Also, there are many approaches that rely on textual representation of information (e.g., [15] or [16]) as this can be very helpful for keyword-based search approaches. Lastly, other approaches, such as [17], are closer to common web-search approaches and base their architecture on data sources that are available on the web.

### *Abstraction of Technologies*

When creating an index on heterogeneous data, the challenge lies in making the individual data source accessible for the application of an indexing strategy. In contrast to typical web indexing, the structure in which the data is stored is not uniform. However, indexing strategies need to know in which representation they will find the individual data sources that have to be indexed. Therefore, it is not directly possible to execute a single indexing strategy on arbitrary data sources. For this purpose, an abstraction from the respective structure of the heterogeneous data sources into a uniform structure must be carried out first. We model this abstraction in our generalized architecture as “connectors” (cf. Figure 2). The identified architectures differ in the realization of connectors and can be divided into two classes: Either the data source technology is abstracted to translate search requests to the native technology of the data source and execute them on the data source (as e.g. in [18], [19] or [20]) or to provide access to the data itself in such a way that an index or similar (external) data structures can be created, which is later used for searching (as e.g. in [21], [22] or [23]). In the second case, there are also some approaches like [24] or [25] in which the complete set of data is analysed and transformed to a unified representation, which is then used exclusively for search. It also must be noted that the connectors are not always a standalone component, but often part of indexing strategies or query translation components.

### *Indexing Strategies*

With technological abstraction, the problem of syntactic heterogeneity can be solved, at least for the functionalities that are offered by the connectors. However, the problem of semantic heterogeneity of the data sources and differences in the characteristics of their data remains. The main differences in the identified search architecture could be found in the indexing strategies. Specific implementations of these also depend on the use-case of a search algorithm: Simple keyword-based search algorithms require different indexing strategies than more complex search algorithms, that may accept a variety of additional parameters or even structured queries. As more precise indexing also requires higher levels of data source integration and therefore higher effort and the requirements can be very different, we refer to the term *indexing strategy* as the process of generating a knowledge representation for searching. The following indexing strategies could be identified in the reviewed literature:

**Graph-based indexing strategies.** These strategies mainly rely on the representation of knowledge about the heterogeneous data sources in graphs. Most commonly, the nodes in these graphs represent entities or attributes in the data, while edges are data-specific relationships [24], [25], [25]. Often, the Resource Description Framework (RDF) is used to encode this information (e.g. in [20] or [25]). Search algorithms can then exploit relations in the graph for search terms that can be associated with specific nodes in the graph or answer structured queries in graph querying languages. Another approach is presented by [21], who use graphs to model the similarity of keywords in the data to find data entities, that match to given search keywords. A similar utilization of graphs was found in [27], who use a graph data structure that represents semantic correlations of data objects as a base for their search algorithm.

**Vector space indexing strategies.** Vector space indexing strategies are based on an algebraic model to assess the relevance of possible search results in relation to the search query and is realized by indexing documents as vectors in a vector space [15]. In this way, each dataset can be represented as a vector in the index, in which every entry represents the importance of a specific keyword in the associated dataset [16]. Finally, the query is also represented in the vector space and a similarity measure is used to find the closest dataset representations in the vector space [28]. Apart from the specific vector representation schema and the similarity measure, the identified approaches did not differ significantly in their realization of this indexing strategy.

**Inverted list indices.** One of the most common indexing strategies that was found, was using inverted list indices to index heterogeneous data sources. These indices

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6170083



typically consist of a mapping of possible search keywords to lists of datasets that are related to these keywords [29]. Keywords for the inverted list can, for example, be extracted from text-based data, or available metadata. Realizations of this strategy differ in how the indices deal with the presence of multiple data sources. For instance, [22] implement a concept with local indices for each data source, which are connected by a global index, that maps search keywords to the local indices. However, most of the time, the specific implementation of inverted indices is not discussed in detail in the reviewed literature. Often, this task is outsourced to external libraries, such as Apache Lucene [30], [23], [31].

**Ontology-based strategies.** Ontologies are a common concept to represent semantic knowledge in a formal way, e.g., by defining concepts, relations, instances, types or other entities and interconnecting them. In searching, ontologies are often used to map general user queries to specific queries, that can be answered by specific data sources, by utilizing semantic knowledge about the similarity of concepts in the user query and the heterogeneous data sources [32], [33]. Ontologies can even be used to provide structured query interfaces (such as SQL) for multiple heterogeneous data sources [34]. Apart from that, ontologies in the search context may also be used to enhance user queries with additional information such as synonyms [35], [36], to transform user queries of geographical terms to machine-readable geographic representations (for instance, to translate place names to coordinates)[37] or simply to provide a common data model, which has to be adopted by the data sources, before they can be searched [38].

**Other strategies.** Finally, some other strategies could be identified that did not match the previously introduced categories. [18] and [19] both present a custom translation model, that transforms user queries to native query interfaces of the data sources without ontologies. [39] also implement such a strategy and refer to this type of query translation as “wrapper” or “mediator” architecture. [40] use Distributed Hash Tables (DHT) to index data files. This works by calculating hashes for possible queries to any data set and then lookup the hashes of user queries in the index. To improve the chance to find matching datasets with a single query, similar queries are derived from the original query and also executed. Similarly, [41] also use DHTs for indexing. In [14], machine-learning is utilized to learn hash-functions that project image or text into a common representation, which can later be used to find search results that are related with an image or text query. Finally, [42] propose a framework, which is based on ElasticSearch*, a search and analytics engine which can flexibly be configured to index different types of data.

## *Search Service*

How a search query is executed clearly depends on the underlying indexing strategy and is often evident. For instance, inverted lists directly provide a mapping from a query to a possible search result. Graph-based indexing strategies use graph querying languages or graph traversal, vector space indices use a similarity measure and ontologies are used to enhance the query itself or map the query to a native query interface. However, users of a search engine usually do not want to see all available search results, but the top-k results that are most relevant [25]. Typical search ranking algorithms such as the famous PageRank algorithm, utilize an inverted text index for keyword search to obtain possible search results and then order them by utilizing a ranking algorithm (cf. [10]). In heterogeneous data sources, ranking search results is also of high interest, but not as easy as in the web search scenario due to their inhomogeneities [43]. There are several different criteria, that can be utilized to rank these search results. They include semantic distance between datasets and search query, number of references to the data set (or similar datasets) [38], similarity ([15], [28], [36]), popularity or user preferences [42], structural properties in graph-indices ([24], [26]) or geographical distances for spatial queries [37]. However, methods for ranking search results from heterogeneous data sources were only discussed by a smaller amount of the identified papers.

Finally, depending on the implemented algorithm, the presentation of the search results also has to be considered in the process of developing search architectures. Especially when visualizing the heterogeneous results for human users, other methods than just displaying a list of search results including their descriptions may be more appropriate. Nevertheless, only a fraction of the identified papers discussed this area at all. Some approaches present their search results as graphs or diagrams [22],[24],[25], others only provide a textual list of results [34], [13].

## *Decentralization*

As heterogeneous data sources are often found in the context of decentralization, this is an important part of search architectures, especially in the context of open search approaches. In the worst-case, each source of data is managed separately with its own connector. On the other hand, many connectors can be an advantage when it comes to data sovereignty: The provider of a data source can have full control over the respective connector and does not have to allow complete and uncontrolled access to the data. This also moves the responsibility and obligations to a wider set of institutions and avoids centrally controlled infrastructures, which can be a problem for independent search approaches (cf. [44]). The same principle can be applied on the index level (see also Figure 2): There may not be a single index for all available data sources: Both duplicating indices and partitioning indices based on specific parameters are options. However, if complete coverage for searching is required, the search algorithm needs access to all available data in indices. In general, the problem of decentrally managed data sources and the organisational aspects were not directly covered by any of the identified approaches. Some of the proposed architectures may still be applicable in these cases, but

* https://www.elastic.co/de/elasticsearch/

https://doi.org/10.5281/zenodo.6170083



especially those, who outsource query execution to the data sources or aggregate large amounts of data from the sources in a central infrastructure may lead to organizational problems.

## CONCLUSION

To assess the current state-of-the-art in searching heterogeneous and decentralized data sources, we conducted an integrative literature review in this paper. It could be shown that searching on decentralized and heterogeneous data sources brings some difficulties in contrast to traditional web search approaches. Distributed data sources differ not only in terms of their content and structures but also within their used technologies and organisational aspects. Furthermore, the results of our review lead to an overview of the important architectural layers and their possible realizations. With the approach of Big Data in nearly any sector of today's economy and research activities, it seems very important, that besides web data, other data sources can also be included in search architectures. Therefore, we expect research that deals with decentralization aspects, both on the technical and organisational level, to be growing in relevance, especially in the context of open search approaches.

## ACKNOWLEDGEMENTS

This work has been funded by the project OpenSearch@DLR. The responsibility for the content remains with the authors.

## REFERENCES

[1] E. Birialtsev, N. Bukharaev, and A. Gusenkov, 'Intelligent search in big data', in *Journal of Physics: Conference Series*, 2017, vol. 913, no. 1, p. 012010.

[2] 'About Open Search Foundation – Open Search Foundation'. https://opensearchfoundation.org/en/about-open-search-foundation/ (accessed May 17, 2021).

[3] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, 'Linked data on the web (LDOW2008)', in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1265–1266.

[4] J. Umbrich, M. Karnstedt, J. X. Parreira, A. Polleres, and M. Hauswirth, 'Linked data and live querying for enabling support platforms for web dataspaces', in *2012 IEEE 28th International Conference on Data Engineering Workshops*, 2012, pp. 23–28.

[5] K. M. Risvik, Y. Aasheim, and M. Lidal, 'Multi-Tier Architecture for Web Search Engines.', in *LA-WEB*, 2003, vol. 3, p. 132.

[6] M. A. Kausar, V. S. Dhaka, and S. K. Singh, 'Web crawler: a review', *International Journal of Computer Applications*, vol. 63, no. 2, 2013.

[7] X. Dong and A. Halevy, 'Indexing dataspaces', in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*, Beijing, China, 2007, p. 43. doi: 10.1145/1247480.1247487.

[8] K. A. F. Mohamed, 'The impact of metadata in web resources discovering', *Online Information Review*, vol. 30, no. 2, pp. 155–167, Jan. 2006, doi: 10.1108/14684520610659184.

[9] F. Adamu, A. M. M. Habbal, S. Hassan, R. L. Cottrell, B. White, and I. Abdullahi, 'A survey on big data indexing strategies', 2015.

[10] N. Duhan, A. K. Sharma, and K. K. Bhatia, 'Page ranking algorithms: a survey', in *2009 IEEE International Advance Computing Conference*, 2009, pp. 1530–1537.

[11] Y. Wang, S. Song, and L. Chen, 'A Survey on Accessing Dataspaces', *SIGMOD Rec.*, vol. 45, no. 2, pp. 33–44, Sep. 2016, doi: 10.1145/3003665.3003672.

[12] M. Gusenbauer, 'Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases', *Scientometrics*, vol. 118, pp. 177–214, Jan. 2019.

[13] B. Delezoide *et al.*, 'MM: modular architecture for multimedia information retrieval', in *2010 International Workshop on Content Based Multimedia Indexing (CBMI)*, 2010, pp. 1–6.

[14] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, 'Inter-media hashing for large-scale retrieval from heterogeneous data sources', in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 785–796.

[15] B. Pathak and N. Lal, 'Information retrieval from heterogeneous data sets using moderated IDF-cosine similarity in vector space model', in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017, pp. 3793–3799.

[16] D. Sun, G. Zhang, and S. Gao, 'Data Management across Geographically-Distributed Autonomous Systems: Architecture, Implementation, and Performance Evaluation', *IEEE Transactions on Industrial Informatics*, 2019.

[17] R. Delbru, S. Campinas, and G. Tummarello, 'Searching web data: An entity retrieval and high-performance indexing model', *Journal of Web Semantics*, vol. 10, pp. 33–58, 2012.

[18] Y. Wang and X. Zhang, 'The research of multi-source heterogeneous data integration based on LINQ', in *2012 International Conference on Computer Science and Electronics Engineering*, 2012, vol. 1, pp. 147–150.

[19] J. Luong, D. Habich, and W. Lehner, 'A Technical Perspective of DataCalc—Ad-hoc Analyses on Heterogeneous Data Sources', in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3864–3873.

[20] R. De Virgilio, A. Maccioni, and R. Torlone, 'A Unified Framework for Flexible Query answering over Heterogeneous Data Sources', in *Flexible Query Answering Systems 2015*, Springer, 2016, pp. 283–294.

[21] M. Ceglowski, A. Coburn, and J. Cuadrado, 'Semantic search of unstructured data using contextual network graphs', *National Institute for Technology and Liberal Education*, vol. 10, 2003.

[22] C. Lin, J. Wang, and C. Rong, 'Towards heterogeneous keyword search', in *Proceedings of the ACM Turing 50th Celebration Conference - China*, New York, NY, USA, May 2017, pp. 1–6. doi: 10.1145/3063955.3064802.

[23] K. Aye and N. Thein, 'Efficient Indexing and Searching Framework for Unstructured Data', p. 122, Dec. 2011, doi: 10.1117/12.921130.

[24] H. Hwang, A. Balmin, H. Pirahesh, and B. Reinwald, 'Information discovery in loosely integrated data', in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 2007, pp. 1147–1149.

[25] X.-S. Vu, A. Ait-Mlouk, E. Elmroth, and L. Jiang, 'Graph-based interactive data federation system for heterogeneous

https://doi.org/10.5281/zenodo.6170083



data retrieval and analytics', in *The World Wide Web Conference*, 2019, pp. 3595–3599.

[26] G. Li, J. Feng, B. C. Ooi, J. Wang, and L. Zhou, 'An effective 3-in-1 keyword search method over heterogeneous data sources', *Information Systems*, vol. 36, no. 2, pp. 248–266, 2011.

[27] B. Lu, G. Wang, and Y. Yuan, 'Towards large scale cross-media retrieval via modeling heterogeneous information and exploring an efficient indexing scheme', in *International Conference on Computational Visual Media*, 2012, pp. 202–209.

[28] N. Lal, M. Singh, S. Pandey, and A. Solanki, 'A Proposed Ranked Clustering Approach for Unstructured Data from Dataspace using VSM', in *2020 20th International Conference on Computational Science and Its Applications (ICCSA)*, Jul. 2020, pp. 80–86. doi: 10.1109/ICCSA50381.2020.00024.

[29] V. Sheokand and V. Singh, 'Best effort query answering in dataspaces on unstructured data', in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Apr. 2016, pp. 155–159. doi: 10.1109/CCAA.2016.7813709.

[30] G. Totaro, M. Bernaschi, G. Carbone, M. Cianfriglia, and A. Di Marco, 'ISODAC: A high performance solution for indexing and searching heterogeneous data', *Journal of Systems and Software*, vol. 118, pp. 115–133, 2016.

[31] A. I. Orhean, I. Ijagbone, I. Raicu, K. Chard, and D. Zhao, 'Toward scalable indexing and search on distributed and unstructured data', in *2017 IEEE International Congress on Big Data (BigData Congress)*, 2017, pp. 31–38.

[32] R. Hai, S. Geisler, and C. Quix, 'Constance: An Intelligent Data Lake System', in *Proceedings of the 2016 International Conference on Management of Data*, New York, NY, USA, 2016, pp. 2097–2100. doi: 10.1145/2882903.2899389.

[33] M. Zhen-Zhong, 'Research of Information Retrieval in the Cloud Computing Environment', in *2014 7th International Conference on Intelligent Computation Technology and Automation*, 2014, pp. 476–479.

[34] F. Klan, E. Faessler, A. Algergawy, B. König-Ries, and U. Hahn, 'Integrated Semantic Search on Structured and Unstructured Data in the ADOnIS System.', 2017.

[35] A. C. Filgueiras, J. C. da Silva, and A. M. R. Vincenzi, 'A PROTOTYPE FOR QUERYING HETEROGENEOUS DATA SOURCES ON THE WEB'.

[36] L. Kerschberg *et al.*, 'Knowledge Sifter:Ontology-Driven Search over Heterogeneous Databases (PDF)', *Scientific and Statistical Database Management, International Conference on*, vol. 0, p. 431, Jul. 2004, doi: 10.1109/SSDM.2004.1311245.

[37] C. B. Jones, A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid, 'The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing', in *International Conference on Geographic Information Science*, 2004, pp. 125–139.

[38] D. Ziébelin, P. Genoud, M.-J. Natete, D. Cassard, and F. Tertre, 'A Web of Data Platform for Mineral Intelligence Capacity Analysis (MICA)', in *International Symposium on Web and Wireless Geographical Information Systems*, 2018, pp. 155–171.

[39] C. Bondiombouy, B. Kolev, O. Levchenko, and P. Valduriez, 'Integrating big data and relational data with a functional sql-like query language', in *Database and expert systems applications*, 2015, pp. 170–185.

[40] P. Felber, E. W. Biersack, L. Garces-Erice, K. W. Ross, and G. Urvoy-Keller, 'Data indexing and querying in DHT peer-to-peer networks', 2004.

[41] A. Asiki, K. Doka, I. Konstantinou, A. Zissimos, and N. Koziris, 'A distributed architecture for multi-dimensional indexing and data retrieval in grid environments', *Proc. of Cracow*, vol. 7, 2007.

[42] N. D. Vo and J. J. Jung, 'Towards Scalable Recommendation Framework with Heterogeneous Data Sources: Preliminary Results', in *2018 14th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, Nov. 2018, pp. 632–636. doi: 10.1109/SITIS.2018.00102.

[43] N. Lal and S. Qamar, 'Comparison of ranking algorithms with dataspace', in *2015 International Conference on Advances in Computer Engineering and Applications*, 2015, pp. 565–572.

[44] N. Kulathuramaiyer and W.-T. Balke, 'Restricting the View and Connecting the Dots-Dangers of a Web Search Engine Monopoly.', *J. UCS*, vol. 12, no. 12, pp. 1731–1740, 2006.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5970346

# A PROPOSAL FOR CLIENT BASED USER PROFILES FOR OPEN SEARCH IN LARGE AND HIGHLY CONNECTED ORGANIZATIONS

Igor Jakovljevic, ISDS, Graz University of Technology, Graz, Austria
also at CERN, Geneva, Switzerland
Stefan Russmann, ISDS, Graz University of Technology, Graz, Austria
Andreas Wagner, CERN, Geneva, Switzerland
Christian Gütl, ISDS, Graz University of Technology, Graz, Austria

*Abstract*

Retrieving the right information by navigating through different information sources on the web and in large organizations has proven to be a prominent issue for individuals. This problem has been tackled by search engine providers, where users construct queries to retrieve the relevant information from a collection of indexed web pages or internal organizational information sources. While being able to retrieve the information on the web or from organizational data sources, not introducing a level of personalization to search engines limit the ability to adapt to users' short-term and long-term interests, making them hard to use over time. On the other side, personalization can invade the privacy of users, by collecting and storing personal and sensitive information. In the context of large organizations, it is extremely important to be explicit with user data collection and usage. Based on the analysis of different search engine systems, user profiling methods, and literature survey on information sharing in large organizations, a conceptual extension that integrates aspects of data privacy and protection into an open search architecture for large organizations has been introduced. The extension aims to improve the process of information retrieval and information discovery in large organizations by introducing a level of personalizations to the system.

*Keywords: Open Search, User Profiles, Data Privacy, Large Organisations*

## INTRODUCTION

Since the creation of the internet, the amount of data generated by humans has been increasing year by year because the world has become more data-driven. Every day people send emails, take photos, make videos, create documents, and use diverse techniques of data and information generation [1]. This behavior of rapid information generation also translates into the work-space, especially within large and highly connected organizations [2]. As the amount of data produced by humans on the Web and within large organizations also rapidly increases, new challenges for navigation through information are formed. Nowadays, navigating the web and information within an organization normally requires using a search engine as the main entry point. This search engine can be one of the major search engines used to navigate the Web or/and an internal organization search engine service.

Individuals interact mainly with search engines by submitting queries that contain certain keywords related to the topics they are searching for [3]. On the other side, search engine service providers use databases of pointers to web pages to react to these queries. These database pointers, also called indexes, are generally built on keywords that relate to information in specific web pages. The search engine compiles user queries and estimates relevance statistics based on words in the query and indexed web pages [4]. When the same query is submitted by different users, a standard search engine returns the same result in ranked order, regardless of who submitted the query. In most cases, this result might satisfy some users, but it does not provide adequate results for all users. Taking the query "virus" as an example, a group of users might be interested in documents dealing with "virus" as a infectious agent that attacks the immune system of living beings, while other users may want documents related to computer viruses [5]. For search engines to adjust the results of queries, the search engines must know which user sent the query, the personal information of the user that might benefit the result, and understand the context of the query. This information about the user can be collected explicitly by asking the user to provide the information or implicitly by collecting user behavior data. Collecting user information explicitly requires more engagement from the user and can be error-prone. This is one of the main reasons why search engines prefer implicit user data collection. When a user sends a query to a popular search engine, such as Google, Bing, or Baidu, while the search engine retrieves the requested information among its indexed web pages, it also stores the submitted query and additional metadata into documents called query logs [6].

Retaining user search query logs, search engine service providers can provide additional services like enhancing ranking algorithms, query fine-tuning, improving personalized query results, combating fraud and abuse, enabling shared data for research, and enabling shared data for marketing and other commercial purposes [7]. The downsides of preserving metadata and user data can lead to serious privacy problems as well. The keywords of each query and the related metadata may disclose sensitive user information such as behaviors, habits, interests, religious views, sexual orientation, etc. [3]. Some query contents may even contain identifiers that allow linking a certain query with an individual. An example of these queries are vanity searches in which an individual looks for their own name on the inter-

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5970346



net [8]. Search providers also rely on the use of user browser profiles for extracting user information, one of these examples is Google Chrome, where you can register with your Google account into a browser and your personal information is stored and used within the browser. This information together with the device used, IP information, and browser information is attached to user queries, which can be used to extract more information about the user [9]. Even if the user's private data and query logs are anonymized, it is possible to uncover the identity of users based on their query preferences. In the year 2006, the internet company AOL released a large amount of user search requests to the public for research purposes. The information was anonymized and did not contain any user information, but personally identifiable information was present in many of the queries. This enabled users to be identified by their search histories [10].

According to a Eurobarometer survey on Data Protection and Electronic Identity in the European Union, less than 40 percent of Europeans were comfortable with the idea of search engine providers accessing their online activity to improve advertising or content. Data also showed that search engine providers are among the least trusted companies to collect and store personal information [11].

At the 2nd International Symposium on Open Search[1] an open information-based search system was presented. The purpose of this system was to offer large organizations the ability to share information transparently, enabling information retrieval to organizational users and also external users, while offering a high degree of data protection and privacy to users. The system offered a high level of privacy and data protection by not tracking user behavior and not invading user-sensitive information, which resulted in the system lacking a level of user personalization. The main difficulty with search engines comes from the fact that to provide a competitive and usable service, it is necessary to use personal information. This personalization information is often stored on the servers of the search engine providers. This allows these providers to exploit user information for monetary gain, opportunities to steal personal information, and general misuse of information [10, 12, 13].

This paper focuses on the exploration of the idea, benefits, and drawbacks of a client-based user profile for search engines in large and highly connected organizations by extending the open search concept proposed at the 2nd International Symposium on Open Search. The idea that the user data is not stored centrally on servers but is stored and managed locally on the respective end device of the user is considered to be very promising. This would mean that the safeguarding of privacy is not left to the organisation of the user or an outside company that can misuse the data or generate profit from it [14]. The first part of this research focuses on the analysis of search engines and search engine types and their relation to privacy and user profiling. Based on previous research, an approach for client-based user profiling for a conceptual open search system in organisations is proposed together with the benefits and drawbacks of such a system.

[1] https://opensearchfoundation.org/en/international-symposium-on-open-search-2019/

## BACKGROUND AND RELATED WORK

Big data organizations produce more than 500 TB of data per week. In the case of CERN, just the Large Hadron Collider generates 10 GB per second. Data produced by the collider is used for research, reports, visualization, communication, and more [15]. The produced information is used in communication, processing, and analysis which produces even more information. Apart from the mentioned information, users of large organizations like CERN use different and/or multiple hardware devices [16]. Even though a lot of information is generated by organizations, it only becomes useful to the users when it is stored and organized in a way that it can be easily navigated to and accessed by users when they desire to find the right information [17].

### *Web Search Engines*

Web Search Engines can be defined as software systems that enable users to find information on the internet or within a organizations intranet with the use of user-specific queries [18]. The main building blocks of search engines are web crawler, indexer, search index, query engine, and search engine interface [19]. Figure 1 describes the connection between these main components. Search engines that are structured this way are also known as index-based search engines.

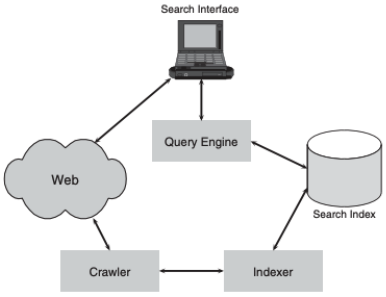



Figure 1: Simplified search engine architecture [19]

Web crawlers have the task to navigate to web pages, download the content of web pages, extract hyperlinks and follow these links for further traversing and download [19]. Content is extracted from the pages, this content can be keywords, topics, meta information, and more. The extracted content is known as a web page index, and the process of index extraction is known as indexing. A collection of web indexes form a data repository, named search index, containing all the information the search engine needs to match and retrieve a web page [20]. For search engines, the query engine is the most important element that enables the user the ability to retrieve and interact with the collected data and the Web. Query engines interpret natural language requests from the user and retrieve ranked information from the search engines about crawled Web pages [19].

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5970346



### *Desktop Search*

As personal computers are becoming more powerful each year, gaining the ability to store and process larger amounts of data. With the increase of available structured and unstructured information, the amount of data formats has also been steadily increasing. The downside of this increase is that the retrieval of information on personal computers is becoming noticeably difficult [21]. Individuals try to organize the data on their personal computers to reduce the time and effort needed to find information. This approach becomes less effective and unusable with the increase of available information. Desktop search engines enable the user easy access to needed local information and are defined as localized search engines. They retrieve references to files on the computer's hard drives based on keywords, file types, or other search criteria [22]. One of the main advantages of desktop search engines is that they offer a high level of privacy to the user, since the user data is not distributed to third-party services, but consumed and stored locally by the local search engine [21].

### *Personalised Web Search Engines*

Providing personalized service to users of Web search significantly satisfies their everyday information needs. As mentioned above, a characteristic of traditional Web search engines is that if different users submit the same query, the system would produce the same list of results, regardless of the user. Personalized search engines, on the other hand, include user information in the search process and retrieve different results for different users [23]. Personalization can be applied to the dimension of the user's knowledge, user's interests, or user's context to produce a user profile for personalized web search [24].

### *User Profiling*

The growth in the volume of available information causes information overloading, defined as the difficulty in understanding an issue and effectively making decisions when provided with too much information about a certain issue. This has led to an increase in demand for personalized approaches for web search engines and information navigation. These systems need to gather personal information about the user, filter out unnecessary information, and recognize supplementary information of possible interest for the user, to create user profiles that can be used for personalization [25]. Common contents of user profiles are: user interests; user knowledge, background and skills; user goals; user behavior; user interaction preferences; and the user context [26]. This information can be divided into structure and unstructured. Structured user information are name, date of birth, gender, etc. and unstructured user information are topic keywords, machine learning models, semantic networks, etc [27]. For large organisations this structural information also includes department, section, skills, work title and position, and more.

Information about a user can be collected explicitly, by providing possibilities for the user to share his information and interests. This is also often called explicit user feedback, which relies on personal information input by the users. Structured information as demographic information ( birthday, marriage status, occupation, or other personal information) represents the type of information collected by explicit user feedback. The drawback of explicit user information collection on the web is that it costs the user time, requires the user to engage and participate in the information collection which can result in inconsistent or incorrect information [25]. Explicit user information collection within organizations has a higher level of validity, making it more usable for personalization. The second method to collect user information is implicit, by gathering and aggregating open user information like navigation, click behavior, submitted queries, and clicked documents [24]. Many live systems on the Web, such as Google, Yahoo, Bing, Facebook, maintain and process the history of users' interactions [28]. The main disadvantage of this method is that it raises privacy concerns for the user, which motivates the user to avoid explicit information collection. In practice, implicit information is more likely to be used because it does not require explicit user engagement to collect information [25]. The collected information aims to define users' short-term and long-term interests to improve the search experience. Short-term interests are temporary interests that are usually satisfied in a relatively shorter period, they are used to personalize search within a current search session. While long-term interests are persistent user interests observed over a longer time frame and can be used to enhance future searches [29, 30].

### *Personalized Web Search Engines and Privacy*

Preserving privacy in a personalized system depends on ensuring that the user feels in control of their information and guaranteeing the integrity of that information [24].

Depending on the security configuration of the device using the search engine, it is possible to extract personal information about the user from a single user query request, without the user knowing or allowing it. The following is a list of some of the most revealing details that the hosts can find out about the computers visiting them: IP address, approximate location, date and time of the visit, browser type, operating system, user language, processor type, display resolution, browser active plugins, and more [13].

Popular search engines also log user queries, these query logs contain the search and navigation history of individual users. They are combined them with the previously mentioned information from user requests to form user profiles [6]. User data gathered by large organizations (e.g. Google, Microsoft, etc.) are kept in centralized servers, which are easy to set up and maintain. Nevertheless, these servers are a potential target for hacking and identity theft. There are additional risks, user data can get destroyed or it can be sold to other organizations whenever the organization gets bankrupt [31]. Besides implicitly collecting user personal information, without the user exactly knowing what is collected while they use web search engines, structured user information (date of birth, legal name, gender, etc.) is also

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5970346



collected [13]. This has left the average user of web search engines not in control of their information and has compromised the integrity of user data stored for personalization.

User information can also be collected at the client's side, specifically on the device/s of the user. An advantage of such a system is that it offers the user a higher degree of privacy preservation and control of information.

### *Discussion*

As previously mentioned, there are many types of search engines, but the system structure of search engines has not changed greatly over the years, and the amount of user information collected by popular search engines has been increasing steadily. Surveys, like the Eurobarometer survey, show that users are reluctant to share their data with search engine providers, but because these providers are the most helpful tools for navigation across the internet, they are forced to use them [11]. Current research shows that by exploiting anonymized information about search engine users, it is possible to retrieve user identifiers like age, gender, and zip code [32]. Looking at modern search engines like google and Bing, we can see that the transparency in what is done with user data is non-existent, based on the fact that the last paper published on the inner workings of such a system is more than 20 years old [33]. Large organisations within the European Union have also complained that search engines such as Google and Bing and their respective algorithms are too vague, which has led the European Union to introduce a set of guidelines that require technology giants such as Microsoft and Google to be more transparent about their inner workings [34]. The identified privacy concerns in common web search engines have produced a need to create new concepts and methods for search engines, and even new types of privacy-aware search engines.

## TOWARDS A CLIENT BASED SEARCH

The open search structure proposed at the 2nd International Symposium on Open Search integrated concepts of information retrieval in large organizations and information sharing between external and internal organizational users. The proposed system lacked the ability to provide personalized information to users, which means that external and internal users would receive the identical result when sending a request to the internal organization search engine. We propose an extension of the open search concept for large and interconnected organizations [35]. This extension needs to provide transparently to the user a way to retrieve personalized information without exposing users' private information (IP, location, browser type, device type, etc.) while enabling the creation and maintenance of user profiles. Based on the research from previous chapters, we have determined that users have a certain level of distrust of modern web search engines. The distrust is based on the fact that the structure, algorithms, and user information usage of modern web search engines are not disclosed transparently to the user. The proposed extension of the open search system aims to tackle privacy issues and empower the user to select the information shared with search engines by storing sensitive information on the client machine ( user's computer, laptop, etc.) and not on the infrastructure of the search engine.

### *System Proposal Overview*

Compared to traditional web search engines, with a client-based search engine, the role of the user changes. The evaluation methods that are used to determine user interests can be adapted to fit a more client-centric or user-centric system. Where the user can agree or disagree on the information usage, information source, and the purpose of the information. This new paradigm would require more research in the area of the user-centric client-based searching process.

Figure 2 describes the conceptual architecture of such a user privacy-oriented system. The local user profiling element of such a system has the task to convert user queries to personalized queries and maintain a local version of a user profile which is updated each time a user executes a query. Personalized queries are based on previous user search information, user explicit information, and more which is stored in local user storage. This enables the user to keep track and secure personal and/or sensitive information without sharing it with the search engines while maintaining the ability to send personalized search queries. Similar to traditional search engines, this component should adapt, learn from the user's interaction and user search queries by improving the user profile. It should also enable the user secure storage, preview, and deletion of stored information, preference, personal data, and other user data.

The search engine proxy has the goal to remove identifiable information from user requests while keeping a level of personalization that would enable the user to retrieve necessary information. This anonymous user query is then forwarded to the search engine. An important aspect of the search engine proxy is that it does not store queries or user information.

### *Creation and Maintenance of User Profiles*

Unlike traditional web search engine user profiles, which are stored on a central server on the web or within an organization, our proposal is to store user profile information within the user client, being that within the browser database or with the use of a plugin to store it on the file system. User information can be generated from many sources: devices (mobile phones, personal computers, wearables); social media, messaging platforms, governmental systems, and much more. It is necessary to keep track of all these information sources and aggregate them into one coherent user profile, while securely sharing it among multiple user devices and services. To store user profiles on local devices and to securely share them across multiple devices it is necessary to find efficient representations without taking up too much space on the device, but still preserving important user details. These user profile representations need to store approximations of users' long-term and short-term interests, which is why we propose that the local user profile are represented twofold.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5970346

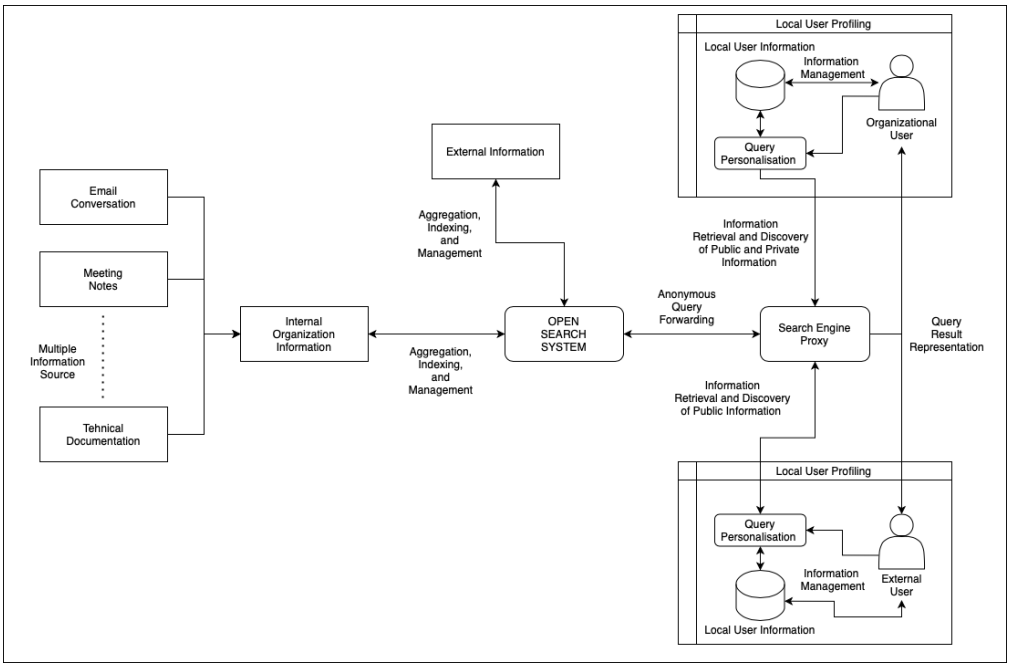



Figure 2: Updated Conceptual Integration Diagram of the Open Search System for Large and Highly Connected Organizations

The first representation is the representation of structural data ( e.g. age, profession, location, etc.), while the second user profile representation contains extracted information from user queries with a time component ( keywords and topics) [24, 30]. Structural data about the user would be obtained by providing an interface for the user to manage her/his information. As mentioned in previous research, this might increase the risk of obtaining false information about the user, since it requires a high level of user engagement. While the implicit user profile data would be collected from user queries by extracting keywords and processing them to determine the topical interests of the user. To create a representative user profile while keeping the information collected about the user to a minimum it is necessary to find a method of continuous processing of user queries while taking into consideration past information.

### *Query Personalisation*

Based on the previously mentioned representation of user information, the query personalization component is used to enhance a user query into a personalized user query. For example, when a user submits a query "virus", the personalization would enhance the query with information from local user profiles. In the case that the user is interested in computer science, the query might be transformed to "computer virus". In the case that the implicit user profile contains extracted information that indicates that the user is interested in a medical topic, the query can be enhanced to "medical virus" or specifically "brain virus". It is easy to see that it is necessary to find a method to rank different user information based on context, time constraints, and changes in user interest before personalizing the query.

### *Data Protection and Search*

As previously mentioned, since users use different devices for everyday activities it is necessary to find a way to securely share user profiles with these devices. Our proposal is the user profiles are securely shared between multiple devices using a local blockchain, where each user would create a local blockchain network, which would be updated as soon as one device updates a user profile [31]. Even with this mechanism of storing the user profile securely, it is possible to extract user information from personalized queries. To increase the level of privacy of user data, we propose the integration of a search engine proxy that would act as a middle man for queries. The main benefit of this would be that search engines would not be able to trace user queries, since different personalized user queries would come from the proxy service.

## SUMMARY AND FUTURE WORK

Many challenges need to be addressed to make the client-based open search in large organizations a more complete and useful tool for information discovery and information retrieval. The technical challenges vary from efficiently storing, generating, maintaining, and sharing user profiles, using those profiles for secure and effective personalization and information retrieval. Other than technical challenges, issues with user privacy and user rights need to be analyzed in-depth to find new ways to educate the user about the data that is being collected and processed.

## CONCLUSION

The level of user information stored by search services like Google and Microsoft has been increasing significantly

https://doi.org/10.5281/zenodo.5970346



over the years, which has lead to users expressing dissatisfaction with the way these services use their data. In this paper, we have addressed the problem of extending web search engines in large organizations with user personalization, while maintaining a high degree of data integrity and respecting user privacy. The proposed concept system uses the idea of securely storing user data on user devices, intending to give control of the way the data is used by the user, while enabling personalized information retrieval.

## REFERENCES


[1] B. Marr, "How much data do we create every day? the mind-blowing stats everyone should read," *Forbes*, p. 1–5, 2018.

[2] I. R. M. A. (IRMA), *Information Resources Management Association, Information Diffusion Management and Knowledge Sharing: Breakthroughs in Research and Practice (2 Volumes)*, vol. 2. 2019.

[3] D. Pàmies-Estrems, J. Castellà-Roca, and A. Viejo, "Working at the web search engine side to generate privacy-preserving user profiles," *Expert Systems with Applications*, vol. 64, 08 2016.

[4] K. Kumar, "Privacy protection in personalized web search using obfuscation," *International Journal of Emerging Trends in Engineering Research*, vol. 8, pp. 1410–1416, 04 2020.

[5] F. Liu, C. Yu, and W. Meng, "Personalized web search for improving retrieval effectiveness," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, pp. 28– 40, 02 2004.

[6] M. Chau, X. Fang, and O. R. L. Sheng, "Analysis of the query logs of a web site search engine," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, p. 1363–1376, Nov. 2005.

[7] A. Cooper, "A survey of query log privacy-enhancing techniques from a policy perspective," *ACM Trans. Web*, vol. 2, Oct. 2008.

[8] C. Soghoian, "The problem of anonymous vanity searches," *SSRN Electronic Journal*, 01 2007.

[9] G. Sudeepthi, G. Anuradha, M. Surendra, and P. Babu, "A survey on semantic web search engine," *International Journal of Computer Science Issues*, vol. 9, 03 2012.

[10] M. Barbaro and T. Z. Jr., "A face is exposed for aol searcher no. 4417749," 2006.

[11] T. O. . Social, "Attitudes on data protection and electronic identity in the european union," 2011.

[12] O. Tene, "What google knows: Privacy and internet search engines," *SSRN Electronic Journal*, 10 2007.

[13] H. Aljifri and D. Navarro, "Search engines and privacy," *Computers Security*, vol. 23, pp. 379–388, 07 2004.

[14] E. Toch, Y. Wang, and L. F. Cranor, "Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 203–220, 2012.

[15] T. Smith and A. Wagner, "Open data open science open search," 2020.

[16] P. Jones, "It department user survey report," 2020.

[17] M.-C. Lee, "Knowledge management and innovation management: Best practices in knowledge sharing and knowledge value chain," *International Journal of Innovation and Learning*, vol. 19, p. 206, 01 2016.

[18] D. T. Seymour, D. Frantsvog, and S. Kumar, "History of search engines," *International Journal of Management Information Systems (IJMIS)*, vol. 15, 09 2011.

[19] M. Levene, "An introduction to search engines and web navigation (2. ed.)," 2005.

[20] M. Thelwall, "The responsiveness of search engine indexes," *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics, ISSN 1137-5019, Nº. 5, 2001*, vol. 5, 01 2001.

[21] B. Markscheffel, D. Büttner, and D. Fischer, "Desktop search engines – a state of the art comparison.," 12 2011.

[22] C.-T. Lu, M. Shukla, S. Subramanya, and Y. Wu, "Performance evaluation of desktop search engines," pp. 110 – 115, 09 2007.

[23] P. Brusilovsky and C. Tasso, "Preface to special issue on user modeling for web information retrieval," *User Model. User-Adapt. Interact.*, vol. 14, pp. 147–157, 06 2004.

[24] M. R. Ghorab, D. Zhou, A. O'Connor, and V. Wade, "Personalised information retrieval: Survey and classification," *User Modeling and User-Adapted Interaction*, vol. 23, 08 2013.

[25] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," vol. 4321 LNCS, 2007.

[26] S. Schiaffino and A. Amandi, "Intelligent user profiling," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5640 LNAI, pp. 193–216, 2009.

[27] C. M. A. Yeung, N. Gibbins, and N. Shadbolt, "A study of user profile generation from folksonomies," in *CEUR Workshop Proceedings*, vol. 356, 2008.

[28] S. Stamou and A. Ntoulas, "Search personalization through query and page topical analysis," *User Model. User Adapt. Interact.*, vol. 19, no. 1-2, pp. 5–33, 2009.

[29] J.-D. Ruvini, "Adapting to the user's internet search strategy," vol. 2702, pp. 145–145, 06 2003.

[30] X. Shen, B. Tan, and C. Zhai, "Implicit user modeling for personalized search," *International Conference on Information and Knowledge Management, Proceedings*, 01 2005.

[31] A. Shrestha, R. Deters, and J. Vassileva, "User-controlled privacy-preserving user profile data sharing based on blockchain," 09 2019.

[32] R. Jones, R. Kumar, B. Pang, and A. Tomkins, ""i know what you did last summer": Query logs and user privacy," CIKM '07, (New York, NY, USA), p. 909–914, Association for Computing Machinery, 2007.

[33] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 30, pp. 107–117, 1998.

[34] REUTERS, "Eu sets out search ranking guidelines for google, microsoft, platforms," 2020.

[35] I. Jakovljevic, C. Guetl, and A. Wagner, "Open search use cases for improving informationdiscovery and information retrieval in large andhighly connected organizations," 2020.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6144571

# TOWARDS OPEN DOMAIN LITERATURE BASED DISCOVERY

O. M. Bensch*, Maastricht University, [6200] Maastricht, The Netherlands
T. Hecking †, German Aerospace Center (DLR), [51147] Cologne, Germany

*Abstract*

Literature based discovery (LBD) is concerned with the extraction of implicit knowledge from large corpora of scientific publications inferring previously unseen links between terms and concepts, which can potentially lead to new hypotheses and findings. Most LBD systems are used in biomedical and related domains, where the existence of well elaborated domain taxonomies and ontologies support the automatic extraction of relevant information from texts. Due to a lack of such resources and different nature of publications LBD has been rarely used in other scientific areas. This work explores and evaluates text and graph methods for open domain concept and relation discovery in scientific literature. First results indicate that several different approaches have to be combined to detect a sufficient amount of concepts and meaningful relationships in an open domain corpus. The work can contribute to broaden the scope of LBD systems and potentially lead to new applications.

## INTRODUCTION

Search applications let users express their information needs as specific search queries that are matched against a search index for documents or translated into database queries. In contrast, discovery systems focus on exploration of information items in a less targeted manner. Here the goal is to exploit rich datasets to discover something new, unexpected, and possibly inspiring. This paradigm has been applied in literature-based discovery (LBD) systems [1] that aim at fostering new scientific developments and hypotheses in an automated manner. The ratio behind this is that large publication databases contain a lot of implicit knowledge that is not manifested in one publication alone but becomes salient when insights from several publications are combined. In this sense, Swanson established the "ABC-Model" for LBD to automatically generate and evaluate new hypotheses [2]. In a first step relations between concepts (or meaningful terms) are extracted from scientific text corpora, e.g. based on co-occurrence in documents, which eventually results in a concept network. In the discovery part, one aims at predicting previously unseen relationships in such networks from transitive relations. A prominent example from Swansons seminal work [3] is: The relation that fish oil (A) lowers the blood viscosity (B) was found in one set of publications. Another set of publications reports that a high blood viscosity (B) causes Raynaud's disease (C). With that explicit knowledge a new hypothesis can be stated that there is an implicit relations between A and C. This hypothesis was later proven correct [3].

* o.bensch@student.maastrichtuniversity.nl
† tobias.hecking@dlr.de

Based on the ABC model two search approaches can be used to generate new hypotheses from literature corpora (illustrated in 1. In open discovery, the search term "A" is given. This concept is used to identify "B" concepts that are related to "A", as well as "C" concepts that are related to the "B" but not "A". In closed discovery two concepts "A" and "C" are given and the aim is to find bridging concepts "B" that connect "A" and "C".

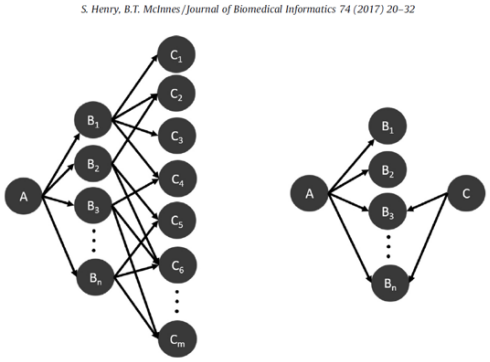



Figure 1: LBD approaches from [4]

The typical workflow comprises of (1) identification of important concepts in documents, (2) relation extraction and creation of a concept network, and (3) ranking of unconnected concept pairs. While the first literature-based discoveries mainly came from manual inspection of documents, with the advancing methods in natural language processing and graph mining the process becomes increasingly automatised [4].

However, the vast majority of LBD systems focus on the biomedical domain, which might be attributed to highly specific and descriptive content in research papers in this domain, as well as the existence of well elaborated taxonomies such as the Unified Medical Language System (UMLS), which alleviates the extraction of meaningful information [5].

As a step towards broader adaptation in scientific discovery and monitoring systems, the main goal of our research is to explore techniques for LBD in open domain text corpora. This paper reports on our first results as well as technical issues along examples from a literature corpus of 25.161 English abstracts retrieved from the publication server elib [6] of the German Aerospace Center (DLR). We, furthermore, reflect on possible future directions for open scientific discovery systems including semantic augmentation and the usage of existing scientific knowledge graphs.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6144571



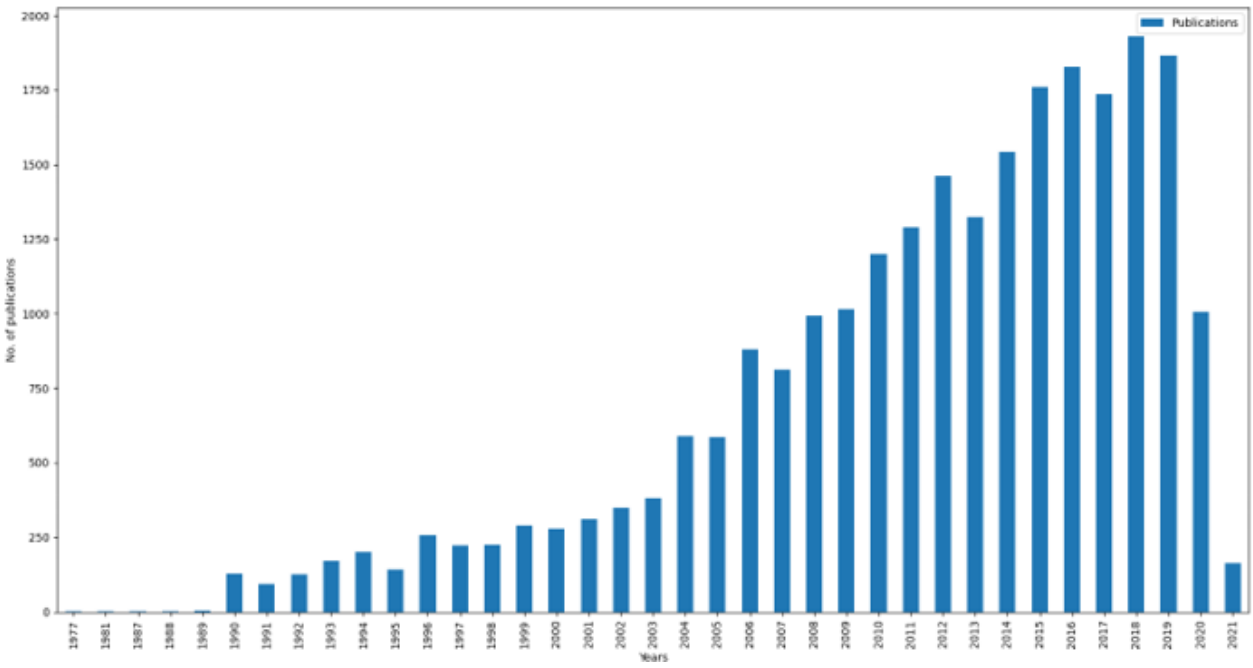



Figure 2: Abstracts per year.

## RELATED WORK

Although literature-based discovery has a long research history originating in the 80's [1], it is still mainly focused on variations of the classical ABC model with applications in biomedicine and chemistry.

In these contexts, LBD systems often use the Unified Medical Language System (UMLS) [7] to identify important concepts mentioned in medical and biological literature. It contains more than 2 million terms, 900,000 concepts, and over 12 million relations. These are used in conjunction with natural language processing methods, like named entity recognition for information extraction from text corpora [8]. One state of the art example for an existing LBD system in the medical domain is LION LBD [9] focusing on the molecular biology of cancer. It uses the PubTator [10] to annotate important medical concepts in publications. Based on co-occurrences of concepts LION LBD uses machine learning to infer promising new relationships for open as well as closed discovery. Due to the lack of comparable ontologies in other domains, semantic augmentation for LBD is rarely used in other domains than biology, chemistry, medicine, and biomedicine [5].

However, most recently it has been shown that automatic discovery of implicit knowledge in publication databases is also possible in other domains. For example, in the field of material science Tshitoyan et al. [26] could predict scientific discoveries years before corresponding experiments were actually conducted by applying machine learning techniques on word embeddings created from older literature.

## METHODS AND EXAMPLES

We created a test data set of 25.161 English abstracts retrieved from the publication server of the German Aerospace Center (DLR) elib [6], to evaluate open domain LBD approaches. The abstracts in this dataset were composed on average of 188,74 words in 8,3 sentences. The distribution of abstracts per year of this dataset can be seen in figure **??**. It can be seen that the amount of papers published per year increases from year to year. Abstracts until the end of may 2021 were included in this dataset.

In the following different approaches for the three main steps of the LBD process (concept detection, relationship identification, and concept pair ranking) are described pointing out their strength and weaknesses for the task at hand.

### *Concept detection*

There are several ways to detect terms in sentences that refer to concepts of interest. Several approaches that solely work on the syntactic level can only detect single words, which will miss out multi-word expressions such as 'machine learning', while other techniques involving grammatical analysis can also detect coherent terms as one concept. Another important aspect of concept detection is matching expressions to ontologies as external knowledge bases, which are, however, often not available or not sufficiently elaborated.

**Named Entity Recognition** One example of a technique that can detect multi-word expressions as one concept, as well as a corresponding ontology is named entity recognition (NER).

The most common open domain model for NER is the Stanford NER [18] which was trained on the OntoNotes 5.0

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6144571

dataset. This NER model was built to detect 18 different types of entities such as persons, organizations, products, dates, or monetary terms. Consequently, this model works well for information extraction in economic domains. However, our experiments have shown that the Standford NER could not be used to extract meaning full concepts from scientific literature since this model was not trained to detect terms such as technical components, methods, etc., which are important for LBD. Overall this model detected 2-3 terms per abstract, which was not sufficient to extract meaning full concept relations.

An example of this approach can be seen in figure 3.

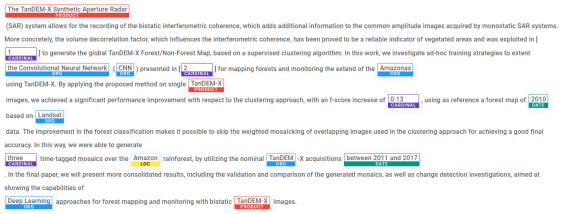


Figure 3: Stanza example.

**Wikification** Another recent approach to detect open domain terms is wikification [15]. Wikification makes use of Wikipedia as external knowledge base to identify meaningful concepts in texts by matching terms to titles of Wikipedia articles (see [15] for an example). Apart from a good precision also for multi-word expression, another advantage is that one can make use of additional information associated with an article especially the category of the page for concept classification or redirects to other articles for synonym resolution. To evaluate the wikification approach a list of all occurring words in the test dataset was created. In the next step, we used the following SPARQL query to query DBPedia (an ontology based on Wikipedia) to detect English terms with matching articles.

```
SELECT DISTINCT * WHERE {
 ?url rdfs:label "' + searchText + '"@en .
}
```

The "searchText" variable was replaced by a single word for each query. In this example, the query was only performed for single terms. However, it can easily be modified to also match multi-word expressions by checking expressions with variable length.

This approach performs better for abstracts of older publication, as Wikipedia entries are created over time. Very novel concepts, for which no wiki article exists will be missed which is definitely a disadvantage for LBD. On average 6.3 concepts could be identified in the abstracts in our dataset, and most of them appear to be useful for scientific information extraction.

An example of this approach can be seen in Figure 4.

**TF-IDF** Another approach is to detect terms with TF-IDF. In contrast to the previous approaches, TF-IDF solely

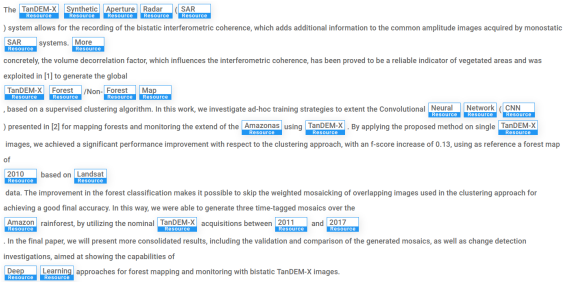


Figure 4: Wikification example.

relies on term occurrence statistics and no ontologies or grammatical analysis are needed. On the downside, this approach can only detect single terms as a concept resulting. Furthermore, one has to specify a threshold for TF-IDF scores of words to be included as a concept. Depending on this threshold common words like "within" can also be classified as a concept. This can be circumvented by using proper stopword lists to remove such common language words.

With well adjusted threshold (in our example case 0.7) and in combination with stopword filtering and lemmatization for pre-processing and possibly further post processing steps, this approach can detect single-word concepts quite reliably (see Figure 5). While TF-IDF detects more concepts than wikification, there are also terms that are not useful e.g. 'single' or 'exploited'. This indicates that TF-IDF may have a higher recall in concept detection but lower precision compared to wikification.

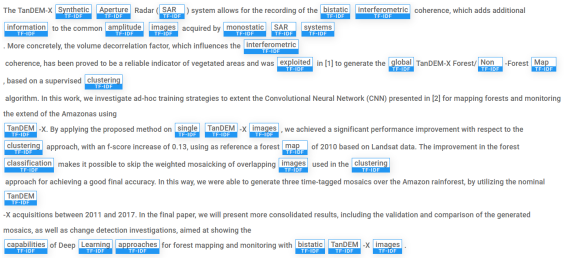


Figure 5: TF-IDF example.

## *Relation identification*

Once important concepts are identified, the next step is to discover their connections. This is often done based on co-occurrences. The most simple approach would be to connect all concepts from the same abstract pairwise. However, co-occurrence can also be defined on sentence or paragraph level. Although co-occurence based concept linking discards semantic information, i.e. the nature of their relationship, it is known that co-occurrence models often inherently captures the semantic structure of a text to a sufficient extent [19]. The significance of a relation between two concepts can be determined from the number of such co-occurrences, which can be used to filter sporadically and noisy relations.

Based on the selected technique for concept detection, this method has to tackle several issues. For methods that only

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6144571



detect single words, like TF-IDF, a co-occurrence-based concept relation detection might also detect coherent concepts instead of relations between independent concepts.

Our experiments have shown that relation detection on sentence-based co-occurrence with concepts extracted with TF-IDF and a threshold of 0.7 can not only be used to detect concept relations but also coherent concepts. In this case, the highest concept relation rank was achieved by the two tokens "open" and "source" as these often co-occur in sentences. As a result, a concept relation with a high co-occurrence rank and a low word distance between those concepts could indicate a coherent concept. Based on a threshold this approach could be used to detect coherent concepts with methods that would normally only detect single word concepts. Depending on the selected threshold words like "within" might be considered as a concept by TF-IDF. Stopword lists could be used as a filter for those words.

In our experiment we used a threshold of 0.1 for TF-IDF. The top 50 co-occurring words can be seen in figure 6.

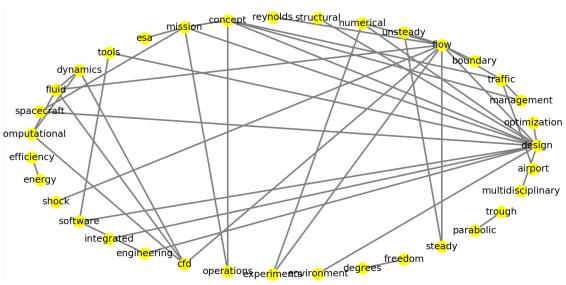



Figure 6: TF-IDF top 50 co-occurring concepts and relations.

By increasing the amount of concept relations detected by this method more and more sub-graphs were combined to larger graphs. However, less important concept relations were included this way.

In a zoomed version of this figure example relations between the extracted concepts can be seen. This zoomed version can be seen in figure 7.

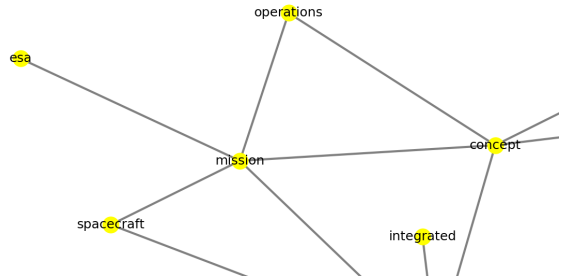



Figure 7: TF-IDF top 50 co-occurring concepts and relations, zoomed.

### *Concept pair ranking*

The result from concept and relation identification can be represented as a network. For discovering new previously not observed relationships that can potentially lead to new hypotheses and ideas, one can borrow from link prediction methods established in social network analysis [20]. The most simple approach is to rank unconnected node pairs based on the number of their common neighbours, which corresponds to the classical ABC model of LBD [1] described in the beginning of this paper. The idea is that for two concepts that share a large number of common neighbours but are not connected by an edge themselves, there could exist an implicit connection. A drawback, however, is that two concepts that are well connected in the network have a higher chance to share neigbours than concepts sparsely connected to the rest of the network. To account for this, one can use the Jaccard similarity of the neighbourhoods of two concepts instead of the total number of shared neighbours. While the aforementioned methods only take two-step connections into account, the Katz coupling measure [20] counts all paths between two concept nodes and includes a weighting parameter that downweights longer paths. Thus, this measure can be considered as a generalisation of the common neighbours method. However, it suffers from the same problem that it can be a biased towards well connected nodes.

To illustrate the differences between different link prediction methods for concept pair ranking, we build a graph of concepts identified by the TF-IDF method with more than 10 co-occurrences on the sentence level. The result is a network of 6109 nodes with 275513 edges. Excerpts of the top ranked concept pairs are given in table 1

It can be seen that the Katz measure and the common neighbours measure share most of the new detected relations, as well as the weights between these. The term 'sar' (search and rescue) is dominant. As mentioned above, one reason can be that terms that have many more connections than the average (also called hubs), which is a well-known property of many networks [21], are connected to many others via short paths. In contrast, the relations discovered with the Jaccard measure differs completely from the other two and appears to be less biased toward such hub concepts.

## CONCLUSION

We have explored different approaches for extracting meaningful concepts and their relationships from publication abstract for open-domain literature based discovery. Initial findings suggest that several approaches should be combined along with pre- and post-processing to improve the detection rate. Apart from that, open domain LBD remains challenging since every discipline differs in the nature of findings and how to communicate them. Consequently, there will be no one-fits-all solution so that open domain LBD systems should rather provide a set of configurable tools from which specific applications can be built.

We believe that the potential of using computational tools to link existing pieces of information to support scientific discovery has not yet fully been exploited, especially in the light of emerging open web indices and scientific knowledge

https://doi.org/10.5281/zenodo.6144571

Table 1: Top concept pairs extracted by different link prediction methods

| Common Neighbours | Katz | Jaccard |
|---|---|---|
| sar – aircraft | sar – aircraft | files – streams |
| images – design | images – design | exchangers – pumps |
| sar – temperature | sar – temperature | housekeeping – streams |
| presented – presents | sar – preasure | linearity – uniformity |

graphs, e.g. Open Academic Graph (OAG) [22]. Advances in this direction can be a key element to make better use of the rapidly growing amount of scientific information available not only in traditional publications but also on the web.

## FUTURE WORK

On the technical level, further approaches for entity extraction could be evaluated. Similar to the wikification approach, a dictionary model like wordnet [23] or, when available, curated domain taxonomy could be used. This would also alleviate the problem of synonyms and multi-word expressions.

Further processing steps like dependency parsing or machine learning models could be used to detect coherent concepts and relations between them. These could also be combined with pattern-based approaches similar to Hearst patterns for discovery of hyponyms [**?**], for example, based on keywords like "influences" that indicate specific connections.

To combine the mentioned approaches, we plan to use weak supervision frameworks such as Snorkel AI [24]. These, take the results of various heuristics expressed as (imperfect) data labelling functions and combines them in a probabilistic framework to create consistently labelled training data for building machine learning models for information extraction.

Apart from technical advancements of components in LBD pipelines, in the future one can also go beyond publication data for knowledge discovery. Since scientific output is increasingly available on the web and manifested also in form of software publications in public repositories or open datasets, one can also include these diverse sources into the discovery process (c.f. [25]).

## REFERENCES

[1] M. Thilakaratne, K. Falkner, T. Atapattu "A Systematic Review on Literature-based Discovery: General Overview, Methodology, Statistical Analysis" in Association for Computing Machinery New York, USA, December 2019, Article No.: 129.

[2] D.R. Swanson "Migraine and magnesium: Eleven neglected connections." in Perspectives in Biology and Medicine 31 Summer 1988, pp. 526–557.

[3] R. A. DiGiacomo, J. M. Kremer, D. M. Shah "Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study" in The American Journal of Medicine Volume 86 Issue 2, New York, 1989, pp. 158–164.

[4] S. Henry, McInnes, T. Bridget "Literature based discovery: models, methods, and trends" in Journal of biomedical informatics 74, 2017 pp. 20–32.

[5] M. Thilakaratne, K.Falkner, T. Atapattu "A Systematic Review on Literature-Based Discovery: General Overview, Methodology, amp; Statistical Analysis" in Association for Computing Machinery 52, 2020.

[6] elib `https://elib.dlr.de`

[7] O. Bodenreider "The Unified Medical Language System (UMLS): integrating biomedical terminology." in Nucleic Acids Research, Volume 32, January 2004, pp. D267—D270.

[8] D. Hristovski, C. Friedman, T. Rindflesch, B. Peterlin "Exploiting Semantic Relations for Literature-Based Discovery" in AMIA Annual Symposium proceedings. AMIA Symposium, 2006, pp. 349–353.

[9] S. Pyysalo et al. "LION LBD: a literature-based discovery system for cancer biology", Bioinformatics, Volume 35, Issue 9, 1 May 2019, pp. 1553—1561.

[10] C. Wei, A. Allot, R. Leaman, Z. Lu "PubTator central: automated concept annotation for biomedical full text articles", Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, pp. W587—W593.

[11] T. C. Rindflesch, M. Fiszman "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text" in Journal of Biomedical Informatics 36 (6), pp. 462–477.

[12] C. Chantrapornchai, A. Tunsakul "Information Extraction on Tourism Domain using SpaCy and BERT" in ECTI Transactions on Computer and Information Technology (ECTI-CIT). 15, pp. 108-122.

[13] E. Sang, F. De Meulder "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition" in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142—147.

[14] R. Weischedel et al. "OntoNotes Release 5.0" in Linguistic Data Consortium Philadelphia, October 2013 `https://catalog.ldc.upenn.edu/LDC2013T19`

[15] Y. Taskin, T. Hecking, H. Hopppe "ESA-T2N: A Novel Approach to Network-Text Analysis" in Complex Networks and Their Applications VIII, pp.129–139.

[16] SPARQL `https://www.w3.org/TR/rdf-sparql-query`

[17] S. Auer et al. "DBpedia: A Nucleus for a Web of Open Data" in 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference Busan, Korea, 2007, pp. 722—735.

https://doi.org/10.5281/zenodo.6144571



[18] J. Finkel, T. Grenager, C. Manning "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." in Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370.

[19] , Carley, Kathleen, Palmquist, Michael (1992). Extracting, representing, and analyzing mental models. Social forces. 70(3), 601–636.

[20] Liben-Nowell, D., Kleinberg, J. (2007). The link-prediction problem for social networks. Journal of the American society for information science and technology, 58(7), 1019-1031.

[21] Barabási, A. L., Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.

[22] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[23] Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

[24] A. Ratner et al."Snorkel: rapid training data creation with weak supervision." in Proc. VLDB Endow. 11, 3 November 2017, pp. 269-–282.

[25] R. el Baff, S. Santhanam, T. Hecking "Quantifying Synergy between Software Projects using README Files Only" in Proceedings of the International Conference on Software Engineering Knowledge Engineering, 2021, Pittsburg, USA, pp. 265–270.

[26] Tshitoyan, V., Dagdelen, J., Weston, L. et al. (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98 .

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772576



# MODULES FOR OPEN SEARCH IN MATHEMATICS TEACHING

Melanie Platz, Lea Marie Müller, Saarland University, 66123 Saarbrücken, Germany
Engelbert Niehaus, Svenja Müller, University of Koblenz-Landau, 76829 Landau, Germany

*Abstract*

To create awareness of the interpretation of search results from some frequently used search engines, learning modules on Open Search are developed and possibilities for a sustainable integration into education are examined. In this paper, a learning module addressing the black box behaviour of some search engines is addressed and links to the German curriculum of mathematics in primary and secondary education are elaborated to enable an implementation into regular teaching. The modules are based on fundamental ideas that can be addressed from kindergarten to upper secondary schools within a spiral curriculum.

## HOW TO REACH PEOPLE?

Even children can have a great influence on problem awareness and decision making in society. Especially the movement "Friday for Future" shows this high impact. Lessons learnt in school could have an influence on problem awareness in the family and triggered activities that arise from discussions between kids to their parents similar to environmental risk literacy (e. g. waste sorting and use of plastic, see BMU, 2018, [1]). The Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) in Germany is convinced that children can argue across from their parents and impart action strategies (BMU, 2018, [1], p. 3). This indicates that sustainability education can be very important if the goal is to change the thinking and decision making of generations. Consequently, this strategy could be transferred to other topics like digital media and search literacy.

The JIM study (Feierabend, Rathgeb & Reutter, 2018, [2], p. 35) showed that children and young people use search engines like Google. The search engine Google ranks 6th out of 13 in the survey of children's and young people's favourite internet offers. Search engines like Google are used by 85% daily or several times a week. The trend is slightly upwards with a 2% increase from the previous year's survey (Feierabend et al., 2018, [2], p. 52). Another German study by the "Deutsches Jugendinsitut" (Feil, Gieger & Grobbing, 2013, [3]) showed similar results with more than 1200 children and adolescents between 6 and 12 years of age. Children and adolescents primarily used the search engine Google to search the internet, which is also used by their parents to search for information (Feil et al., 2013, [3], p. 7). Only one-third of parents know alternative search engines for children. Furthermore, the study showed that searching with search engines was accompanied by problems: especially searching with keywords and dealing with the search results posed challenges for the children. The children usually only use the first results when displaying hits. In addition, they mostly did not go to the website displayed, but presumably took the information primarily from the hit display. The study showed another interesting result regarding the periods of use of search engines: the children and young people used the search engines for search queries much more often during school time instead of during leisure time. This again clearly shows that this topic should be addressed in school.

This creates another problem because the curricula are already filled with important content that must not be omitted and in primary school no new school subject is supposed to be created in Germany: the development and acquisition of the necessary competencies for living in a digital world "[...] go far beyond the necessary basic knowledge of informatics and concern all subjects. Therefore, they cannot be assigned to an isolated learning area." (KMK, 2016, [4], p. 12, translated by the authors). Dealing with digitalisation in the school sector should focus on the "primacy of the pedagogical" (KMK, 2016, [4], p. 51) and the "primacy of subject didactics" (GDM, 2017, [5], p. 41) and must be integrated into pedagogical and subject didactic concepts in which learning is at the forefront (KMK, 2016, [4]; Platz, 2019, [6]).

To implement the topic of Open Search into the school curriculum, fundamental ideas have to be identified and linked directly to the existing curriculum. As the authors work in the field of mathematics education in Germany, this is exemplarily done in this field in this paper.

The fundamental ideas should already be laid down in a child-friendly way in the initial lessons and be taken up again at the further stages of the learning process, i.e., in later grades, and be structurally enriched in the process. In doing so, the fundamental ideas are taken up again and again at different stages, on the one hand at a higher level and on the other hand in a structurally enriched form (Krauthausen, 2018, [7]). According to the spiral principle, these basic concepts and relationships (fundamental ideas) should be dealt with in mathematics lessons in several cycles, each at a different level, using means of representation, language and didactic models appropriate to the developmental stage of the pupils. The knowledge on a learning topic is developed step by step (in a spiral) (Käpnick & Benölken, 2014, [8], p. 54).

In this project, the learning modules are developed using Design Science Research (DSR; e. g. Peffers et al., 2006, [9]; Prediger et al, 2012, [10]). DSR aims to develop and evaluate solutions to problems to codify knowledge about design sciences as design theories. DSR usually involves the creation of an artefact (in our case: learning modules on Open Search in mathematics teaching) and/or a design theory (in our case: design principles for learning modules linking computer science content to mathematics education) to extend and optimise the current state of

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772576



practice as well as existing research knowledge (Vaishnavi, 2019, [11]). With this goal in mind, the Dortmund model for subject didactic development research on diagnosis-guided teaching-learning processes for researching and further developing teaching (e.g. Prediger et al, 2012, [10]) is combined with the DSR Methodology Process for researching and further developing information systems research (e.g. Peffers et al., 2006, [9]) to be able to productively use synergy effects of both approaches for developing learning environments on the topic of open search in mathematics teaching. In the following one exemplary learning module in mathematics education is suggested.

## PRIMARY EDUCATION

For search engines to be used purposefully, users need to understand how search engines work and are structured. This is also important for a critical examination of the field. In this learning module (called "Black Box"), the non-transparency of some powerful and frequently used search engines is addressed.

### *Transparent sorting*

In the context of search engines, sorting algorithms play an important role (Halavais, 2017, [12]). Starting already in kindergarten, the functionality of such sorting algorithms can be discovered and actively experienced by the children. Bubble Sort, which is because of its slowness rather of didactic importance (Fischer & Hofer, 2011 [13], p. 839), can already be investigated in kindergarten. Sönnerås (2019) [14] (p. 49f) provided kindergarten children (n=6) with the information, that they learn how a search engine works and that they enter data to be processed on the one "end" and receive sorted data at the other "end". Rods of different lengths are prepared and the children are asked to pick one and to choose a position on a line (drawn on the floor), where they want to stand first (drawn rectangles can be used for marking the different positions on the line, the number of rectangles corresponds to the number of children). Then the functionality of the flow chart is described: Starting on the left side, the two children who are standing next to each other are supposed to compare the length of their bars. The child with the longer bar should step right, the child with the shorter bar should step left. (Older children might need to help the younger children to get the sorting right). When the child on the right side is reached, the sorting starts again on the left side. After several steps, the sorting is finished and the children put the bars on the floor in front of them to see if it worked. If the order is not correct, a concrete error search can be done with the children. Sönnerås (2019) [14] reports, that the children had doubts about whether the correct sorting was just a coincidence and whether it would work on the next run. They repeated the experiment and started in other rectangles and with other rods, but the result was the same. The children wondered if this would also work with other things instead of the bars. They sorted themselves by size, among other things, and saw that it also worked. This inductive testing convinced them, that the algorithm would always work. The basis of this activity is measuring, in more detail: direct comparison. This activity promotes competencies in the field of geometry and measuring and sizes (sorting geometric figures according to properties through comparing the lengths of the bars, e. g. KMK, 2004, [15] p. 10 & 11). The direct comparison picks up on previous experiences of the children of ordering and comparing and stimulates a conscious examination of the concepts of relation ("… is as long as ..."; "… is longer as …", etc.) through actions. Without these basic experiences in each size range, children cannot build an understanding of the equivalence and order relation in that range (Franke & Ruwisch, 2010, [16], p. 185f). Related to sorting algorithms, the direct comparison is connected with sorting and ordering from the smallest to the largest. In this process, the transitivity of the order relation becomes clear: if bar A is longer than bar B and bar B is longer than bar C, then bar A is longer than bar C. Even though transitivity related to lengths of bars is mastered by almost all children at the beginning of the first school year in primary school, difficulties can arise with weak children in early teaching, and that this requires special attention. The transitivity underlying the order relation, which manifests itself on the children's level of activity in the fact that they not only compare pairs but can also carry out a ranking (seriation) of several objects, represents an essential aspect of the size concept and measurement concept. However, there is no simple correspondence between action, concept and verbal description for the children (Franke & Ruwisch, 2010, [16]). Through such sorting tasks, process-related competencies can be promoted like communicating (especially working on tasks together, making and keeping agreements) and arguing (especially when checking the sorting process on correctness and finding and correcting errors) (e. g. KMK, 2004, [15] p. 8). Besides Bubble Sort, Insertion Sort and Sorting Networks can be investigated (Quiring-Tegeder, 2016, [17]).

Following this "computer science unplugged" learning environments, "searching as a game" (Menzel, 1978, [18] p. 151ff) can be implemented in primary school using the Scratch programming environment. Schwätzer (2018) [19] describes the approach of constructive programming from the point of view of mathematics didactics and presents some tried and tested examples of lessons in primary school using the Scratch programming environment, including arithmetic and geometry. The subject matter is addressed on three levels:

- (1) exploring with paper and pencil,
- (2) creating a programme flowchart
- (3) independently dealing with code structures.
- In addition, Schwätzer (2018) [19] emphasises the important step of returning to the mathematical content (e. g. by working on mathematical questions about the programme or questioning the sense of using the created programme).

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772576

In the learning environment "guessing numbers" Schwätzer (2018, [19], p. 50ff) two problem variations can be investigated and programmed with Scratch:

- (1) the computer "thinks up" a number and gives hints and the child in front of the computer guesses;
- (2) the child in front of the computer thinks up a number, remembers this number and gives feedback to the computer, if the guessed number is too big or too small or if it is the right number. The computer uses the "trick" of always reducing the search spaces by "guessing" exactly the middle number of the search space by forming the new search space from the old one by cutting off the part that is no longer relevant (obtained from the information "too big" or "too small", i.e. whether it is above or below the middle number) (see also https://pikas-mi.dzlm.de/pikasmifiles/uploads/images/Spielanleitung_Mister%20X.pdf ).

With this game among others understanding of number relations and the sorting of numbers by size (KMK, 2004, [15], p. 9) can be promoted as well as systematic sampling as a problem-solving method (KMK, 2004, [15], p. 7).

### *Intransparent sorting*

In the described learning environments, the supposed sorting order was given (lengths of bars with the relation "… is longer as …", cardinal numbers with the relation "… is bigger as …") and each involved learner agreed to these orders naturally (because of the order relations on this sets). Especially in the initially described computer science unplugged learning environments, the children might start to ask themselves, which other properties instead of length can be used to sort (e. g. weight, etc.) which could also lead to discussions when the decision for the "right" order is not clear (already in kindergarten: Sönnerås, 2019, [14], p. 50).

In the next step, objects or even the children themselves could be sorted by the teacher or another child in a certain (maybe even random) order and the children make assumptions about the sorting criteria and find arguments (KMK, 2004, [15], p. 8) for their assumption. As a basis for such sorting profiles can be used. Such profiles could be generated for the whole school class or several school classes and in this way, the pupils collect and afterwards structure and present data in tables, charts and diagrams (competence data, frequency and probability, KMK, 2004, [15], p. 11). If e. g. the question "Which drinking chocolate do you like most?" is asked (e.g. to collect information for the school kiosk) and a ranking of drinking chocolates starting with the most popular (e.g. drinking chocolate A) is done which could represent the search results for the investigated group of children (that could be given to the school kiosk to provide the children with the drinking chocolate that is statistically liked by most of the children in the investigated group). Then the question could be asked: "What would happen if the school kiosk would have a contract with a drinking chocolate provider, whose drinking chocolate (drinking chocolate B) is not liked by most of the children?" as the starting point for a discussion about issues (e. g. economic interests) that could influence a ranking. Furthermore, some children (especially those whose favourite drinking chocolate is not the one liked by most of the children) could perceive the procedure as unfair. In the next task, the option-selection is not binary (like – not like), but fuzzy: The children are supposed to place themselves on a line on the floor where the starting point is "I do absolutely not like" and the endpoint is "I like very much". With a certain question like "Do you like drinking chocolate C"? Then the children are asked to place themselves on the line. Then a first ranked search result can be assigned based on their position on the line, e. g. as visualized in figure 1:

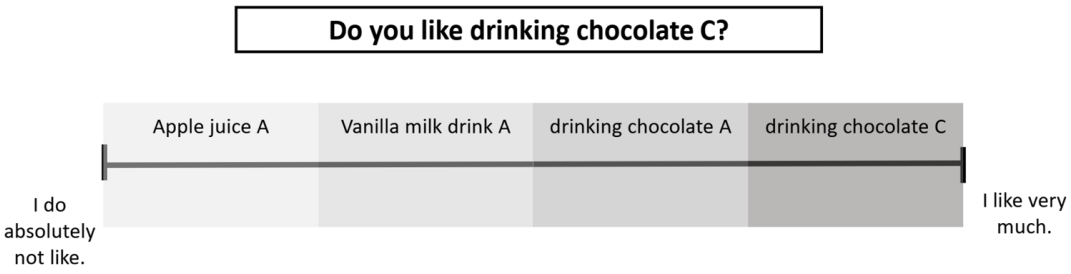



Figure 1: Example for an assignment of the first search result

This way the assignment of search results based on user profiling can be made experienceable by the children. The children can discuss now, how the assignment was done, what other options there could have been and that the search results individually differ if a search engine uses such procedures.

## SECONDARY EDUCATION

Adapting the example to secondary education in schools probability theory and stochastics is a relevant topic in the curriculum. For search engines the ranking is a key element to present the results of the search in a specific order. Learners need to understand how the ranking of search results have an impact on the visibility of results to society and therefore which educational content has an impact on societal changes, risk literacy and awareness. The link

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5772576



between ranking in search engines and probability theory applied to the visibility of search results is a key to understand the basic principles of search engines. This is also important for a learning module, that addresses ranking in a non-transparent way and how the weights on the search results are assigned, which induces up-ranking and down-ranking of specific results and in turn the visibility of those results in society.

## *Manual ranking*

In the context of search engines, we start with a specific content (e.g., plastic waste and plastic bottles used for soft drinks) with four different internet resources:

- One states that plastic bottles have less weight, need less energy for production and during the transport process than glass bottles and therefore plastic bottles are more environmentally friendly than glass bottles (Stefanini, Borghesi, Ronzano & Vignali, 2021, [20]).
- The next resource is referring to the recycling of the plastic waste in concrete and (Hassani, 2005, [21]) focuses on recycling in comparison to avoiding plastic waste.
- The third source focuses on plastic mitigation and therefore suggests eliminating plastic bottles (Jia, Evans & Linden, 2019, [22]) and switching to alternatives, such as glass bottles.
- The fourth statement addresses microplastics in food chains (Zarfl, 2011, [22]) and how microplastics may have an impact on food safety (World Health Organization (WHO), 2019, [23]).

## *Curriculum Integration*

Integration into the curriculum beyond mathematics education is possible by addressing the same issue in social subjects in schools, by transferring the issue of plastic waste to elections, and by making the specific arguments of political parties visible to the public. A sound problem solving takes multiple views, scientific results from different disciplines, minorities of society and humanitarian side effects into account when decision making is performed. If search results are ranked according to personal preferences of search engines that reflect personal preferences in the ranking violates that principles of decision making based on multiple criteria and neutral point of view (NPOV). A neutral point of view is difficult to achieve even if the NPOV principle is explicitly addressed e.g. Wikipedia's NPOV rules (Matei, 2011, [24]). The main topic may be regarded also as generic for education in school to train critical thinking and encourage the learner to perform evidence-based decision making and being risk literate about the consequence of living in a search bubble in which only the preferences of the own attitudes are communicated to the learner. So going back to the plastic waste example the task for upper secondary students is to search for peer-reviewed articles on the topic and create a manual ranking according to the results and define the criteria in which the articles are ranked. The students should assign weights to all the search results. Finally, they should discuss which criteria are reflecting personal interest in specific views of the topic and which criteria could be used to perform a more preference independent ranking of the search results.

## *Ranking and visibility*

Now we accept that the assignment of weights to the search result is not transparent to the public. The learner should address the link between ranking and visibility in the context of stochastic and probability theory especially addressed in lower secondary grades as a topic of the curriculum (Schupp, 1982, [25]) starting with a random ranking of search results over three or more visible pages on a computer defined as static HTML pages. These static pages of search results that contain the real links of the internet sources can be ordered according to a specific content by the students. The results in a specific order are presented to other learners who are requested to create a summary of a specific topic. All results will have the same header but have different links of real search results mapped to the header. The intended result would be that learners who receive a particular designed order of search results will produce a different summary of the topic than another group of learners who receive a different order. The learners know that the results are real results of a search engine, but the header has the same topic for all search results e.g. "microplastics, plastic bottle, food chain".

The learning objective is that ordering and ranking of results have an impact on the public opinion derived from the search results. Students should be able to explain what a bias in content delivery is and how it can be generated by an intransparent ranking of search results. The second step is to quantify that approach within the stochastic. The learners should identify a non-Laplace probability distribution on the search results. Assume we have $n$ search results and the click statistics of learners on which of the search results from the list was clicked. Because the learners are exposed to random orders of search results the click statistic of the learner show how ranking creates a preference to use first results at first. If they find suitable answers for their search query they may stop without looking at later ones presented e.g. at the very end of the results. Applied to the context of plastic waste the students will quantitatively analyse the search results they have looked at in more detail and calculate simple probability distributions. Discrete stochastic analysis of non-Laplace probability distributions is required for page rank. Going back to the curriculum consideration in mathematics education, it is possible to link the topic of search engines to probability theory and stochastic without introducing a new topic into the curriculum of lower and upper secondary schools.

https://doi.org/10.5281/zenodo.5772576

## CONCLUSION

This paper addresses the possibility to assign a learning module about Open Search addressing the Black Box behaviour of some frequently used search engines which can be linked to the standard curriculum in primary schools and secondary schools in mathematics education in Germany. Stochastic and combinatoric content was identified as a link to the standard curriculum. In lower secondary schools non-Laplace distributions that refer to page ranks and access statistics are identified as existing content. Furthermore, the content addressed in the stochastic analysis offers different interdisciplinary links that are shown with the plastic waste example (e.g. organic chemistry in upper secondary schools, biology, ethics and philosophy with a neutral point of view example, ....) In the next step, the learning module will be optimized by teacher training students in a university course using DSR and it will be tested with children. The learning modules will be published as OER via Wikiversity.

## REFERENCES


[1] Bildungsministerium für Umwelt, Naturschutz und nukleare Sicherheit (BMU) (2018). *Abfall - Bildungsmaterial für die Grundschule. Informationen für Lehrkräfte.* https://www.bmu.de/fileadmin/Daten_BMU/Pools/Bildungsmaterialien/gs_abfall_handreichung_lehrer.pdf (retrieved 19.06.2021).

[2] Feierabend, S., Rathgeb, T., & Reutter, T. (2018). *JIM-Studie 2018. Jugend, Information, Medien. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger. Suttgart: Medienpädagogischer Forschungsverbund Südwest, c/o Landesanstalt für Kommunikation.* https://www.schau-hin.info/fileadmin/content/Downloads/Sonstiges/JIM_2018_Gesamt.pdf (retrieved 19.06.2021).

[3] Feil, Ch., Gieger, Ch., & Grobbing, A. (2013). *Projekt: Informationsverhalten von Kindern im Internet – eine empirische Studie zur Nutzung von Suchmaschinen.* München: Deutsches Jugendinstitut. https://www.dji.de/fileadmin/user_upload/www-kinderseiten/898/1-BMBF-Fkz%2001PF08017.pdf (retrieved 19.06.2021).

[4] Kultusministerkonferenz (KMK) (2016). *Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt“* [Electronic version]. (Beschluss der Kultusministerkonferenz vom 08.12.2016) https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2016/Bildung_digitale_Welt_Webversion.pdf (retrieved 19.06.2021).

[5] Gesellschaft für Didaktik der Mathematik (GDM) (2017). *Die Bildungsoffensive für die digitale Wissensgesellschaft: Eine Chance für den fachdidaktisch reflektierten Einsatz digitaler Werkzeuge im Mathematikunterricht* [Electronic version]. Mitteilungen der Gesellschaft für Didaktik der Mathematik, [S.l.], n. 103, (pp. 39–41), july 2017. https://ojs.didaktik-der-mathematik.de/index.php/mgdm/article/view/59 (retrieved 19.06.2021).

[6] Platz, M. (2019). Vorstellung eines Entscheidungsunterstützungssystems zur Auswahl passender Apps und Applets für den Mathematikunterricht der Grundschule. In R. Rink, & D. Walter (Ed.), *Beiträge zum 5. Band der Reihe „Lernen, Lehren und Forschen mit digitalen Medien“. Digitale Medien in der Lehreraus- und -fortbildung von Mathematiklehrkräften - Konzeptionelles und Beispiele* (pp. 167–182). Münster: WTM-Verlag.

[7] Krauthausen, G. (2018). *Einführung in die Mathematikdidaktik-Grundschule.* Heidelberg: Springer Spektrum.

[8] Käpnick, F., & Benölken, R. (2014). *Mathematiklernen in der Grundschule.* Berlin, Heidelberg: Springer Berlin Heidelberg.

[9] Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The design science research process: a model for producing and presenting information systems research. In *Proceedings of the first international conference on design science research in information systems and technology* (DESRIST 2006), 83–106.

[10] Prediger, S., Link, M., Hinz, R.; Hußmann, S., Thiele, J., & Ralle, B. (2012). Lehr-Lernprozesse initiieren und erforschen – Fachdidaktische Entwicklungsforschung im Dortmunder Modell. *Der mathematische und naturwissenschaftliche Unterricht, 65*(8), 452–457.

[11] Vaishnavi, V., Kuechler, W., & Petter, S. (2019). *Design Science Research in Information Systems.* January 20, 2004 (created in 2004 and updated until 2019 by Vaishnavi, V. and Kuechler, W.). http://www.desrist.org/design-research-in-information-systems/ (retrieved 19.06.2021).

[12] Halavais, A. (2017). *Search engine society.* Hoboken, New Jersey: John Wiley & Sons.

[13] Fischer, P., & Hofer, P. (2011). *Lexikon der Informatik*. Berlin, Heidelberg: Springer.

[14] Sönnerås, K. (2019). *Programmieren im Kindergarten.* Berlin: Bananenblau – Der Praxisverlag für Pädagogen.

[15] Kultusministerkonferenz (KMK) (2004). *Bildungsstandards im Fach Mathematik für den Primarbereich.* (Beschluss der Kultusministerkonferenz vom 15.10.2004) München, Neuwied: Wolters-Kluwer, Luchterhand Verlag.

[16] Franke, M., & Ruwisch, S. (2010). *Didaktik des Sachrechnens in der Grundschule.* Heidelberg: Spektrum.

[17] Quiring-Tegeder, A. (2016). *Sortieralgorithmen auf dem Schulhof. Kinder geben Kommandos*. https://kinder-geben-kommandos.de/2016/04/11/sortieralgorithmen-auf-dem-schulhof/ (retrieves 19.06.2021).

[18] Menzel, K. (1978). *Elemente der Informatik: Algorithmen in der Sekundarstufe I.* Berlin, Heidelberg: Springer.

[19] Schwätzer, U. (2018). *Programmieren in der Grundschule.* CreateSpace Independent Publishing Platform.

[20] Stefanini, R., Borghesi, G., Ronzano, A., & Vignali, G. (2021). Plastic or glass: a new environmental assessment with a marine litter indicator for the comparison of pasteurized milk bottles. *The International Journal of Life Cycle Assessment*, *26*(4), 767–784. https://doi.org/10.1007/s11367-020-01804-x.

[21] Hassani, A., Ganjidoust, H., & Maghanaki, A. A. (2005). Use of plastic waste (poly-ethylene terephthalate) in asphalt concrete mixture as aggregate replacement. *Waste Management & Research*, *23*(4), 322–327.

[22] Jia, L., Evans, S., & Linden, S. van der. (2019). Motivating actions to mitigate plastic pollution. *Nature Communications*, *10*(4582). https://doi.org/10.1038/s41467-019-12666-9.

https://doi.org/10.5281/zenodo.5772576



[22] Zarfl, C., Fleet, D., Fries, E., Galgani, F., Gerdts, G., Hanke, G., & Matthies, M. (2011). Microplastics in oceans. *Marine Pollution Bulletin*, *62*, 1589–1591.

[23] World Health Organization (WHO) (2019). *Microplastics in drinking-water*. Genf: World Health Organization. http://edepot.wur.nl/498693 (retrieved 19.06.2021).

[24] Matei, S. A., & Dobrescu, C. (2011). Wikipedia's “neutral point of view”: Settling conflict through ambiguity. *The Information Society*, *27*(1), 40–51.

[25] Schupp, H. (1982). Zum Verhältnis statistischer und wahrscheinlichkeitstheoretischer Komponenten im Stochastik-Unterricht der Sekundarstufe I. *Journal für Mathematik-Didaktik*, *3*(3-4), 207–226.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6045722

# CREATING A DATASET FOR KEYPHRASE EXTRACTION IN PHYSICS PUBLICATIONS AND PATENTS

André Rattinger *, ISDS, Graz University of Technology, Graz, Austria
Christian Gütl, ISDS, Graz University of Technology, Graz, Austria

*Abstract*

Extracting keyphrases and entities can be an important first step in many Natural Language Processing (NLP) and Information Retrieval (IR) Tasks. There are many datasets to train models for standard entities, but it is hard to find data that can be used for more domain specific applications. The types of keyphrases someone wants to extract vary enormously between different fields, which makes otherwise successful algorithms perform poorly on them. One of the fields where this is the case is Physics, specifically to process physics publications and patents. In comparison to news articles or social media, the typical entities like Organization, Location or Person are not helpful when extracting important information from publications or patents. There are few dataset annotations for specific domains, and even when they exist they are not easily transferable. This work contributes an annotated dataset for the facilitation of information retrieval and extraction in Physics. The dataset spans Physics Patents as well as Publications. It covers both of these document types to enable future work between them. This can facilitate future work such as tracking inventions from the first emergence in a publication to the adaption in a patent.

*Keywords:* Keyphrase Extraction, Named Entity Recognition, Dataset, Semantic Search, Physics, Patents, Publications

## INTRODUCTION

Extracting keyphrases and named entities is one of the fundamental tasks of semantic search based on knowledge graphs. Those tasks can be crucial for retrieving the most relevant body of work, but it can also help in analysing the developments in research over time, recommend citations or search for a combination of keyphrases.

As more papers and patents get published every year, it takes a lot of effort to keep up with the current state-of-the-art in a lot of fields. Retrieval systems only address this to an extent as it is hard to search for certain methods or even combinations of entities, in particular for a specified domain such as physics. Especially patents are hard to retrieve because of the fact that certain descriptions are very close to legal language [1]. For some types of data and specific types of entities this task is practically solved, but the extraction of keyphrases is a harder task than just identifying entities. Extracting and classifying keyphrases and entities also helps with the creation of knowledge graphs and the population of existing ones. Considering the relationships between the different parts of a query helps to improve the retrieval accuracy.

Most of the research on Keyphrase Extraction and Named Entity Recognition (NER) takes place on a few datasets, and application to more general data can be disappointing. NER datasets are usually restricted to a few key entities such as Person, Organization and Location. In addition, there is Fine-Grained Named Entity Recognition [2] that differentiates between types of those three entities and adds a few more categories. The more general categories help with the retrieval when it comes to every day topics, but fail when it comes to something really domain specific. There have been shared tasks sometimes to improve the state of NER and keyphrase extraction, that try to improve the state of the problem, but they usually limited in scope and what they cover. Two applications of this that we are particular interested in are building a knowledge graph for semantic search and searching for entity combinations, which can be build on top of the open search index or linked to other search engines (Example: Materials in semiconductors that can change over time as more research and development takes place).

Even if a scientific publication is in the same domain as the patent, the language can be very different, which makes it difficult to apply existing algorithms to a patent corpus for retrieval [1]. This work is probably closest to the shared task of SemEval 2017 [3] that proposed a task with a dataset that covered the keyphrases of process, material and task for the fields of Computer Science, Material Science and Physics. We use the types Process, Material and Task that are used in this dataset but also add some more entities that fit the domains of Physics in Patents and Publications well. With this work we want to build a bridge between Publications and Patents and we envision work that can use this to its advantage. All of the annotated documents that are used to create this dataset are freely available, either from the patent offices or from arxiv. Those resources are linked together with the dataset.

Our main contribution is a new annotated dataset. In addition, we provide reasonable baselines and models for all of the tasks to enable future comparisons and improvements on those baselines. Those first results are provided by a neural network based model we trained on the data, and therefore provide a good starting point for future research with room for improvements. The task is generally harder than NER which is reflected in those baselines. Similar observations have been made in the past for similar tasks. Keyphrase extraction is generally a harder problem than NER because they can vary significantly between domains,

* ...@protonmail.com

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6045722



and the extracted keyphrases can be longer than typical NER tags [3]. The paper is structured in the following way: The next section describes the related work and the background to the task including the most important datasets that have been created in the field. Section Collecting the Corpus describes how the patents and publication for the annotation were selected. Section Annotation Process describes how the dataset was created and annotations were collected. Section Dataset Overview gives an overview over the dataset and Baseline Results creates simple baselines for the dataset future work can be compared against. The paper concludes with Section Conclusion and describes potential future work in Section .

## BACKGROUND AND RELATED WORK

Notably, there have been a few common datasets that are generally used as a test for new named entity recognition. There are also a few domain specific datasets, but there is very little in the domain of physics Some of them are either not openly available, use only a limited to the three most common entities.

### *Background*

NER and keyphrase extractions are fairly well explored field when it comes to news articles and content of every day life (such as social media), but there are a lot of domain specific keyphrases where almost no datasets exist. News articles are a fairly broad as they can talk about a lot of different domains, but the topics are more general than a very specific scientific domain such as physics. In those cases, the state-of-the-art approaches for those fields don't fall short, but they encounter tasks that they were not trained for.

The second most common annotations after news articles can be found in the domain of medicine. A lot of annotations for different molecules, diseases or pharmaceuticals exist. Other domains do have very little annotation in comparison.

An advantage of the news based annotations with the entity types of Person, Location, Corporation is that other extracted keyphrases from the text that don't fall into those categories can be linked to existing knowledge bases such as dbpedia, a knowledge base based on wikipedia [4]. This knowledge base contains fairly general knowledge, making it a good target to link to from news articles. An ontology extracted keyphrases and entities of this dataset could be linked against is the ScienceWise Ontology [5]. Some other promising Ontology projects exist in scientific domains, such as Biology and even Math.

### *Related Work*

One of the most popular tasks or dataset is the CoNLL-2003 Shared Task [6]. This task has a huge body of resource associated with it and the state-of-the-art has been beaten many times. The main types in this dataset are Location, Organization and Person. This makes this task very interesting for specific applications such as tagging news articles, but is not helpful as a source for training in many other domains such as our target domain of physics.

Another dataset that is frequently used as a testbed is the english part of Ontonotes5 [7]. Beside other annotation, Ontonotes provides more entity types such as Person, Organization, Location, Work Of Art, GPE and others. Along with those eleven entity types, it also provides annotations for different measurable types of quantities. As some of the entity types in this task are less specific, the task is closer to the one in this work, but still doesn't include most of the keyphrases that would be useful to build a physics based knowledge graph.

The WNUT2017 shared task [8] is yet another task that comes with a dataset of annotated entity classes. Entity classes it provides are Person, Location, Corporation, Consumer Good, Creative Work and Group. The goal of this task was to solve emergent entities in the domains of social media or news. Emergent entities are newly emerged entities previously not seen that might cause problems to NER models. We envision the usage of the dataset in a similar way, where it would enable models to be trained that can identify newly emergent technologies in the physics domain.

The task that includes the dataset that is the closest to this work is the SemEval 2017 task [3]. The SemEval task was constructed out of three different tasks, where one dealt with the extraction and classification of keyphrases and entities.

In addition to this, there are several dataset for different scientific domains, such as the NCBI Disease Dataset [9], the i2b2/UTHealth shared task [10] or astronomy [11].

## COLLECTING THE CORPUS

The corpus for annotation was collected from arxiv[1] and freely available patents from the USPTO[2]. We extracted 32,832 documents in a pre-selection stage.

### *Patents*

A seed of patents was selected from the computing related patent class in physics (G06). We extended to the most semantically similar patent classes. This selection was based on previous analysis on which patent classes were connected with each other [1]. A selection this way has the advantage that we do not just use a single class, but a bigger subset of the big field of physics. We sourced patents that were created from 2010 to 2018 and created document embeddings for the claims sections with the gensim toolkit [12]. The following hyper-parameters were used in creating the embeddings:

- algorithm: skip-gram
- iterations: 50
- window size: 7

[1] `https://arxiv.org/help/bulk_data`
[2] `https://developer.uspto.gov/data`

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6045722



- dimensions: 300
- min count: 10
- sub-sampling threshold: $10^{-5}$
- negative sample: 5

We have chosen the claims section as it generally is one of the best sections for information retrieval and the most important section when it comes to prior art search [13]. We sourced patents that were not too similar to each other because we wanted to avoid overlaps between patents that are very similar to each other to have a broader selection of tokens. The similarity was measured with euclidean distance and we maximized the distance between the pre-selection documents. Unlike this approach, it would also be reasonable to select patents that are similar to each other as it could highlight the evolution in certain topics.

### *Publications*

Next we collected physics publications from arxiv between the years 2010 and 2018. Similar to the physics patents, we also created document embeddings for the publications with the same parameters as we used for the patents. For this selection, we identified pairs of patent and publication that were close to each other. The selection was made this way to enable future work to draw parallels between the two sources that use very different language.

A total of 1000 initial documents were sourced this way for annotation, although not all of them were used in the end as the annotators didn't get to them. We included author and organization descriptions whenever possible form different sections of the documents, because they are rarer in full texts of publications and patents compared to other sources.

## ANNOTATION PROCESS

The dataset is designed for the tasks of mention level keyphrase indentification and classification. We use the following keyphrase types:

- ORGANIZATION (Any organizational form, can be institution, agency, department)
- PROCESS (Methods used in creating the product, e.g. execution of instructions; oxidization)
- MATERIAL (Physical materials or any tangible elements used. Example: hydrogen, water, catalyst)
- TASK (The problem or task at hand to solve: Named Entity Recognition, Information Extraction)
- PART (Parts used in creating the final product or parts of a whole. Example: touch sensor, graphical user interface)

The annotation process was done in two steps: Pre-annotation and manual annotation. We decided to split the annotation in those two steps to conserve the time of our volunteers, which led to us being able to collect more annotations.

### *Automatic Pre-annotation*

We selected the keyphrase types because it is a good mixture between classical named entity recognition and more challenging keyphrase like types. We automatically pre-annotated the data wherever possible, because solely manual annotation of keyphrases would be very time consuming. Domain specific annotations for our selected types are more complicated to spot and identify for manual annotators than Organization or Person. To support this, we used a pre-trained bert based NER model [14] that was fine-tuned on the CoNLL dataset [6] to do a first pre-annotation. Afterwards we transfer the model and fine-tune it to the SemEval dataset. The main one that was missing was the PART tag. The main disadvantage of this approach was that longer keyphrases were not identified in a lot of cases. Nevertheless the approach facilitated the annotation process as many of the simpler keyphrases and entities were correctly tagged.

### *Manual Annotation*

For the manual annotation, we recruited five volunteers. All of the volunteers are experts in the field of physics and are employed in different capacities at CERN. We prepared sample annotations from the pre-annotated corpus that fulfilled our quality assessments and showed those example annotations to the annotators. We instructed the annotators to also correct pre-annotated keyphrases when they thought the models from the pre-annotation step made an error. In addition, we provided the annotators with guidelines to help them getting introduced to the task. As the annotators are all experts in their field, we opted to double annotate the documents. When there was a disagreement between the annotations, a manual check was performed by a third expert. The agreement between the annotators was measured with the average cohen's kappa and fleiss coefficients, which are respectively 0.81 and 0.82.

Documents were annotated with the doccano open source annotation tool [15], which allows each user to have their own user space. The tool was publicly hosted and all annotators had access to it over two months.

## OVERVIEW OVER THE ANNOTATIONS

This section gives a brief description of the annotated dataset and its characteristics. The annotated dataset is split into a training and validation set. The training data set contains a total of 300 annotated documents, with even splits between Publications and Patents. The test dataset contains 100 annotated datasets with an equal split as well. While

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6045722



Table 1: Overview over the dataset.

| | |
|---|---|
| Domains | Physics Pubs and Patents |
| Classes | Org, Process, Mat, Task, Part |
| Training Documents | 300 |
| Training Publications | 150 |
| Training Patents | 150 |
| Validation Documents | 100 |
| Validation Publications | 50 |
| Validation Patents | 50 |
| Number of Keyphrases | 4,642 |

splitting between the training and testset, there was no consideration for splitting them according to their semantic similarity as we did when selecting the documents from the two fields. This could have been done as well, but we decided against this approach and randomly sampled the table. This still give us a good mix of already seen and unseen keyphrases. Table 1 shows an overview over the dataset statistics. The data for the Publications and Patents texts come in different forms. Publication abstracts are available in textual forms and don't need any further processing to be used for the annotation process. There are the few data types for patents, but the most useful one for this analysis is the xml based. We extract the claims section from the patents. For this dataset we only use patents that where granted by the patent offices.

## BASELINE MODEL AND RESULTS

We present a baseline model and results in this section as a way to compare future results against the annotated dataset and to provide an easy way to get started with keyphrase extraction.

### *Baseline model*

We trained a bidirectional LSTM (long short term memory) model with a CRF (Conditional random field) layer on the data. Models similar to this have shown to provide good results on NER tasks in the past. One of the other advantages of this approach is that compared to other models, we do not have to do careful feature engineering. The bidirectional LSTM layer does a forward and backwards pass over the data to utilize contextual information before and after the tag which can be especially useful for the more complex longer keyphrases. Each word is represented as a vector of character and word embeddings. We used the adam optimizer [16] and trained the model for 20 epochs. Figure 1 shows the general layout of the bidirectional LSTM with all the network layers.

The baseline shows that this task is harder than pure NER tasks. NER task results on CoNLL for example tend to reach F1 scores higher than 90. The baselines is similar

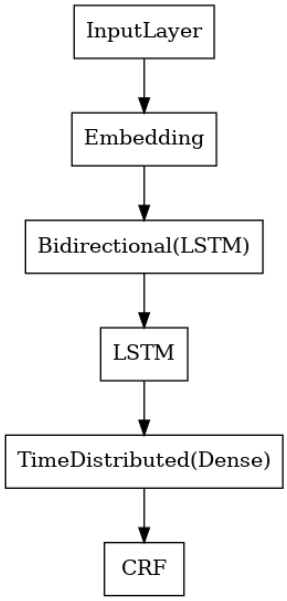



Figure 1: DNN model used to create the baseline. The model and the model layer visualization were created in keras.

Table 2: Baseline evaluation results for the dataset.

| Metric | Result |
|---|---|
| F1 | 0.33 |
| Precision | 0.27 |
| Recall | 0.42 |

to the the results some of the participants in the SemEval task reached, which is plausible as the task also focused on the more complicated field of keyphrase extraction and identification.

### *Results*

For the baselines, standard evaluations metrics are calculated, namely F1-score, precision and recall. Table 2 shows the baseline results for the datasets which can be improved upon by more intricate methods. We provide a baseline for that annotated dataset to have a starting point for future comparisons.

## CONCLUSION

Keyphrase Extraction or the extraction of named entities is an important first step in many applications in NER or Information Retrieval. With this work, we introduced a new dataset for keyphrase extraction in the physics domain for publications and patents. In addition, we trained a neural network model for this task. This model provides a baseline future work can be evaluated against. The bi-LSTM model with CRF layer has been chosen because this approach was applied successfully to similar tasks in the past. The sim-

https://doi.org/10.5281/zenodo.6045722

ple baselines are as expected lower than other baseline for NER tasks considering that keyphrase extraction is a harder problem than NER. Similar previous tasks that dealt with keyphrase extraction showed similar results in terms of F1 score, recall and precision. The annotations span publications and patents on purpose to enable work that can track progress between the fields, highlight the differences between the language or make it easier to train models that are successful on patent data.

## FUTURE WORK

We plan to improve on the baselines of the dataset and use the result of the extractions to link to different ontologies and knowledge bases. For new emergent nodes that don't exist in the knowledge bases, the knowledge base can be populated with the newly extracted nodes. In addition, we plan to expand the dataset with more annotations in the future. Training the models on more data would further improve the models trained on the dataset. Another interesting direction we would like to explore is to annotate the dataset for semantic relationships extraction. As the concepts are already annotated, this would be less effort than the intial creation of the dataset.

## REFERENCES


[1] A. Rattinger, J.-M. Le Goff, R. Meersman, and C. Guetl, "Semantic and topological patent graphs: Analysis of retrieval and community structure," in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2018, pp. 51–58.

[2] X. Ling and D. Weld, "Fine-grained entity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012.

[3] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum, "Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications," *arXiv preprint arXiv:1704.02853*, 2017.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[5] K. Aberer, A. Boyarsky, P. Cudré-Mauroux, G. Demartini, and O. Ruchayskiy, "Sciencewise: A web-based interactive semantic platform for scientific collaboration," in *10th International Semantic Web Conference (ISWC 2011-Demo), Bonn, Germany*, 2011.

[6] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[7] R. Weischedel et al., "Ontonotes release 5.0 ldc2013t19. web download." *Philadelphia: Linguistic Data Consortium*, 2013.

[8] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 140–147. [Online]. Available: https://www.aclweb.org/anthology/W17-4418

[9] R. I. Doğan, R. Leaman, and Z. Lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.

[10] A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1," *Journal of biomedical informatics*, vol. 58, pp. S11–S19, 2015.

[11] T. Murphy, T. McIntosh, and J. R. Curran, "Named entity recognition for astronomy literature," in *Proceedings of the Australasian Language Technology Workshop 2006*, 2006, pp. 59–66.

[12] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[13] A. Rattinger, J. L. Goff, and C. Guetl, "Local word embeddings for query expansion based on co-authorship and citations," in *Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018.*, 2018, pp. 46–53. [Online]. Available: http://ceur-ws.org/Vol-2080/paper5.pdf

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang, "doccano: Text annotation tool for human," 2018, software available from https://github.com/doccano/doccano. [Online]. Available: https://github.com/doccano/doccano

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044976



# The Impact of Main Content Extraction on Near-Duplicate Detection

Maik Fröbe,* Matthias Hagen,* Martin-Luther-Universität Halle-Wittenberg, Germany
Janek Bevendorff,† Michael Völske,† Benno Stein,†Bauhaus-Universität Weimar, Germany
Christopher Schröder,‡ Robby Wagner,‡ Lukas Gienapp,‡ Martin Potthast,‡
Leipzig University, Germany

*Abstract*

Commercial web search engines employ near-duplicate detection to ensure that users see each relevant result only once, albeit the underlying web crawls typically include (near-)duplicates of many web pages. We revisit the risks and potential of near-duplicates with an information retrieval focus, motivating that current efforts toward an open and independent European web search infrastructure should maintain metadata on duplicate and near-duplicate documents in its index.

Near-duplicate detection implemented in an open web search infrastructure should provide a suitable similarity threshold, a difficult choice since identical pages may substantially differ in parts of a page that are irrelevant to searchers (templates, advertisements, etc.). We study this problem by comparing the similarity of pages for five (main) content extraction methods in two studies on the ClueWeb crawls. We find that the full content of pages serves precision-oriented near-duplicate-detection, while main content extraction is more recall-oriented.

## INTRODUCTION

Typical web crawls contain many pages with identical or very similar content and different URLs [10]. Search engines retrieving pages from such web crawls may encounter those near-duplicates in multiple stages of their pipeline. During indexing, omitting near-duplicates might reduce the index size. During retrieval, near-duplicates might occur in the search engine result pages, reducing the user experience because users gain nothing from viewing the same result twice or more on the search engine result pages [5]. Hence, identifying near-duplicates is a mandatory step in web search, with commercial search engines like Google showing only the "best" version from a set of near-duplicates for a query.*

Widely available web crawls—most notably the ClueWebs† and the Common Crawl‡—contain the (near-)duplicate documents that the crawler encountered during the crawling process. While the inclusion of near-duplicates enables many applications (like research on text reuse [2]), it introduces problems for search engines (that we will discuss later in this paper). The CopyCat resource [12] addresses the problems introduced by near-duplicates in information retrieval experiments by providing a precision-oriented near-duplicate detection. CopyCat comes in two parts: (1) ready-to-use compilations of near-duplicate documents within and between selected web crawls, and (2) a software library to deduplicate arbitrary document sets, e.g., search engine result pages before they are shown to searchers.

* <first-name>.<last-name>@informatik.uni-halle.de
† <first-name>.<last-name>@uni-weimar.de
‡ martin.potthast@uni-leipzig.de
* developers.google.com/search/docs/advanced/guidelines/duplicate-content
† lemurproject.org/clueweb09.php/ and lemurproject.org/clueweb12/
‡ commoncrawl.org/

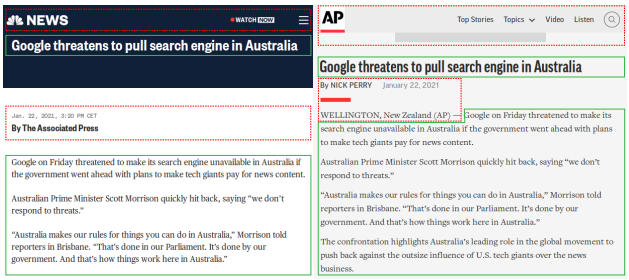



Figure 1: Pages with different URLs and the same article. Both pages have identical content (indicated by green boxes) but vary in parts irrelevant to searchers (indicated by red dashed boxes).

Removing near-duplicates from the search engine result pages with a framework like CopyCat comes with the inherent difficulty of balancing precision and recall. A low precision might reduce the effectiveness of a search engine because relevant and novel documents might be omitted. In contrast, low recall reduces the user experience because users see more near-duplicates. In the context of web search, tuning a similarity measure faces additionally the problem that some parts of the pages that are irrelevant for searchers can increase or decrease the similarity of pages. Figure 1 shows an example of two identical articles located at different URLs where the "noise," i.e., the navigation bar, reduces the similarity, eventually having a negative impact on the recall.

Using only the "retrieval-relevant" part of documents for the near-duplicate detection might be a promising direction to improve the recall in cases as exemplified in Figure 1. However, the impact of main content extraction on near-duplicate detection is not studied so far. While a "perfect" main content extraction should positively affect precision and recall, existing implementations might even harm precision and recall. E.g., invisible changes in the HTML structure of a page might cause tag-based main content extraction approaches to extract different main contents from pages that have identical content from the users' perspective.

We conduct experiments on the ClueWeb crawls to investigate to what extent main content extraction may improve near-duplicate detection in information retrieval. Therefore, we extract document pairs from the ClueWebs containing redundant content (according to their canonical URL) to draw a sharp line between "roughly similar" documents and near-duplicates. We calculate the syntactic similarities for all document pairs after extracting the text from the raw HTML with four main content extraction methods and one full content extraction method that does not discard the "noise" parts from the document. We evaluate the precision and recall of all five content extraction methods based on manual near-duplicate judgments in two case studies. First, we review document pairs uniformly sampled from the full similarity range for all content

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044976

extraction methods confirming that main content extraction increases the recall as exemplified in Figure 1. Secondly, we review 100 cases per main content extraction method in which main content extraction changes the similarity drastically, e.g., with identical documents having disjoint main content, or dissimilar pages having duplicate main content, finding inaccurate main content extraction in most of those cases.

## RELATED WORK

This section reviews definitions for near-duplicates, their prevalence on the web, and approaches to near-duplicate detection implemented in the CopyCat framework.

### *Defining Near-Duplicates*

The fact that there is no universal near-duplicate definition renders their detection and comparable analysis difficult. Restrictive near-duplicate definitions [15, 17] consider documents as near-duplicates if they differ only by their session or message IDs, timestamps, visitor counts, server names, invisible differences, URL parts, or if they are entry pages to the same site. Note that documents, even with minimal content changes, are often not considered near-duplicates under such a restrictive definition. Bernstein and Zobel [6] relax the near-duplicate definition by applying an information retrieval focus, allowing minimal changes in the content. They consider a document pair as near-duplicate if users get the same information from both documents for all "reasonable queries." We adopt the near-duplicate definition of Bernstein and Zobel since it considers a pair of pages as duplicates if they are equivalent in terms of information provided, i.e., ignoring parts of pages irrelevant to searchers (templates, advertisements, etc.).

### *Studies on Near-Duplicates on the Web*

According to previous studies by Fetterly et al. [10, 11], 30% of the pages on the web are near-duplicates. While web pages change regularly, consecutive versions of the same web page are usually highly similar [9]. Subsequent investigations [1, 10, 11, 18, 19] confirm this observation by tracking web pages between 5 weeks and one year. For example, Adar et al. [1] repeatedly crawl 55 000 URLs over 5 weeks finding two-thirds of the pages changed their content, observing that most of these changes were minimal. Ntoulas et al. [18] tracked 150 pages over one year, finding that 40% of them were still accessible after one year, noticing only insignificant changes on most pages.

### *Near-Duplicate Detection*

There are syntactic, URL-based, and semantic algorithms for detecting near-duplicates [3], from which the detection of syntactic near-duplicates received the most attention, resulting in many effective algorithms based on fingerprinting techniques [7, 8, 15, 17]. The CopyCat framework implements syntactic near-duplicate detection in large web crawls with the SimHash algorithm using a fingerprint size of 64 bit and a Hamming-threshold of 3 bits as suggested by Manku et al. [17], while reducing the number of calculated pairwise similarities with the partitioning scheme proposed by Henzinger et al. [15]. Complementary to estimating the similarity of documents with SimHash, CopyCat can calculate the lossless $S_3$ fingerprint similarity [5] for near-duplicate detection in small sets of documents, such as run and qrel files frequently used in information retrieval experiments.

## NEAR-DUPLICATES IN WEB CRAWLS: RISKS AND POTENTIALS

We recapitulate two risks and one potential of near-duplicate pages in web crawls that the CopyCat framework addresses. Please note that we here focus on information retrieval and that other risks, e.g., in the training of large language models [21], exist.

### *Risk: Evaluation of Search Engines*

Bernstein and Zobel [6] found that near-duplicates cause problems in information retrieval evaluations because search engine users do not benefit from seeing near-duplicates. Therefore, they introduce the so-called novelty principle, which states that a document, though relevant in isolation, is irrelevant if it is a near-duplicate to a document the user has already seen on the search engine result page. Especially on web crawls with many near-duplicates, the novelty principle has a non-negligible impact on evaluating search engines [14]. E.g., applying the novelty principle on the runs submitted to the Terabyte track 2004 decreases mean average precision scores by 20% on average [6].

The classical evaluation setup of search engines employs the Cranfield paradigm, making it pretty easy to oversee negative impacts caused by near-duplicates. Relevance assessors judge the relevance of documents to a query in isolation, seeing only one document at a time. Hence, situations that would severely reduce the experience for searchers, e.g., when many near-duplicates occur at subsequent positions in the ranking, can be overlooked because assessors do not look at the ranking. Topic 194 of the ClueWeb09 Web Tracks includes a particularly striking example, where among 47 relevant documents, there are 40 near-duplicates of the same Wikipedia article.

### *Risk: Training of Learning to Rank Models*

Near-duplicates form a kind of oversampling because multiple identical or very similar copies of a page are in the dataset. As recently exemplified [22], oversampling data before partitioning it into training and test sets can invalidate evaluations in machine learning because models may see the same object during training and test. This leakage of information is not possible during the training of learning to rank models because the train/test partitioning is done per query. Still, not removing near-duplicates during the training of learning to rank models decreases the effectiveness of models and biases the trained models [13].

A study [13] on the ClueWeb09 with 42 ranking features using popular algorithms finds that near-duplicates in the training data harm the retrieval performance, since the presence of near-duplicates is unaccounted for in the loss-minimization of learning to rank and in subsequent evaluations. Furthermore, by varying the number of Wikipedia near-duplicates in the training set, the study showed that models might be biased towards retrieving near-duplicates at higher positions. Hence, these observations make a strong case that learning to rank pipelines benefit from removing duplicate documents from the data before training the model.

Mitigating the negative effects of near-duplicates during the training of retrieval models is easily possible with the CopyCat framework, which can deduplicate the training and test set.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044976



### *Potential: Transfer of Relevance Labels*

In contrast to the previous two risks to the validity and robustness of search engine evaluation and tuning, near-duplicate detection enables the transfer of relevance judgments between different editions (or updates) of web crawls [12]. Relevance judgments—obtained from click logs or expert assessments—are an important and costly resource for the development of search engines. E.g., the effort for the 73,883 relevance judgments for the TREC Web tracks on the ClueWeb09 crawl can be estimated at a manual labor of about 4–8 full-time person-months (assuming 40-hour weeks with 30–60 seconds per judgment [23]).

To "reduce" the costs of keeping the relevance judgments up-to-date for ever-evolving web-indices, search engines might transfer relevance judgments from the previous version of a crawl to the next version when they find the judged documents (or near-duplicates of them) in the newer version of the crawl. In a showcase [12] using precision-oriented near-duplicate detection with the CopyCat framework, 10% of the ClueWeb09 relevance judgments could be transferred to the ClueWeb12. The number transferred relevance judgments would even increase to 15% when the ClueWeb12 crawling process would have ensured that the URLs judged in the ClueWeb09 are part of the URL seeds for the next crawling round. More frequent updates (compared to the gap of three years in the relevance transfer showcase) would likely further increase the amount of transferrable relevance judgments. Additionally, the reported experiments on the transfer of relevance labels have used only the full content of the pages. Hence, further improvements, e.g., by leveraging main content extraction to increase the recall while maintaining good precision, are possible.

## CONTENT EXTRACTION EXPERIMENTS

To experimentally compare the impact of main content extraction on near-duplicate detection, we construct a dataset of 186 819 ClueWeb document pairs with redundant content as indicated by canonical URLs. For each document pair, we calculate its syntactic similarity with the lossless $S_3$ fingerprinting [5] for four main content extraction algorithms and the full content of pages. We label 900 document pairs as near-duplicates or not sampled with two approaches: (1) with 100 document pairs stratified sampled from the $S_3$ distribution of each of the five content extraction methods, and (2) with 50 document pairs with maximal positive/negative $S_3$ differences between each of the four main content extraction methods and the full content of a page.

### *Dataset Construction*

We aim at constructing a manageable dataset for our experiments that allows us to draw a sharp line between "only similar" documents and near-duplicates. Therefore, we identify document pairs in the ClueWeb09 and ClueWeb12 that should contain redundant content because they share the same canonical URL. We inspect all documents in the ClueWeb, group them by their canonical URL, and select 5000 groups having the same canonical link at random. From each group, we select all possible pairs (with a maximum 50 document pairs per group) giving us 186,819 document pairs.

### *Document Preprocessing*

We preprocess all documents with the CopyCat framework. CopyCat provides five content extraction approaches that transform the raw HTML of a page into text: four main content extraction approaches and one full content extraction. The four provided main content extraction approaches are Boilerpipe [16], Jericho,[§] Justext [20], and Trafilatura [4]. The full content extraction uses JSoup[¶] to extract the plain text—without any main content extraction—from the HTML. After extracting the documents text, we remove stop words using Lucene's default stop word list for English, apply stemming with the Porter Stemmer, and lower case the remaining words.

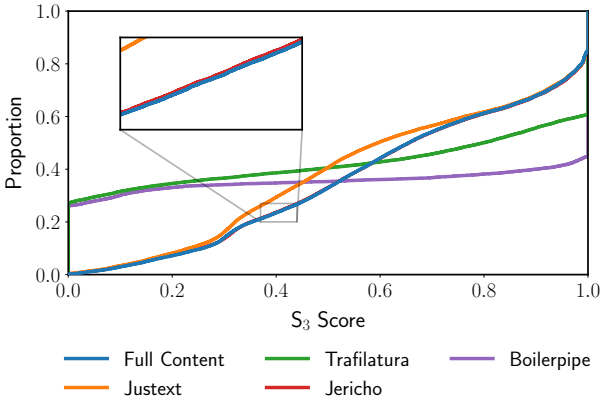



Figure 2: Cumulative distribution plot showing the proportion of document pairs in our dataset below a given similarity measured by their $S_3$ score for all five considered content extraction methods.

### *Similarity in our Dataset*

We use the lossless $S_3$ fingerprint similarity [5] using word-8-grams to calculate the similarities between all document pairs for all five content extraction methods in our dataset. An $S_3$ score of 0 indicates no overlap between documents, and an $S_3$ score of 1 means equality. Figure 2 shows the cumulative distribution plot for all five content extraction methods regarding the portion of document pairs below a given $S_3$ score.

We can identify two groups of content extraction methods that share similar overall behavior. The first group consists of the main content extraction methods Trafilatura and Boilerpipe that show many document pairs with an $S_3$ score of 0 (26% for Boilerpipe and 27% for Trafilatura), indicating that the main content extraction produces disjoint main contents from the considered documents in a pair. This group additionally contains many document pairs with an $S_3$ score of 1 (55% for Boilerpipe and 39% for Trafilatura), indicating that for documents in a pair often the same main content is extracted.

The second group consists of the full content, Justext, and Jericho methods of content extraction. Approaches in this group have very few document pairs with an $S_3$ score of 0 (the Justext approach from this group has the maximum of 0.35% of document pairs with an $S_3$ score of 0), and much fewer document pairs with an $S_3$ score of 1 (all three have around 15.6% of document pairs with an $S_3$ score of 1). We calculated the Pearson correlation between the full content extraction and all other methods. We found a very

[§] http://jericho.htmlparser.net
[¶] https://jsoup.org/

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

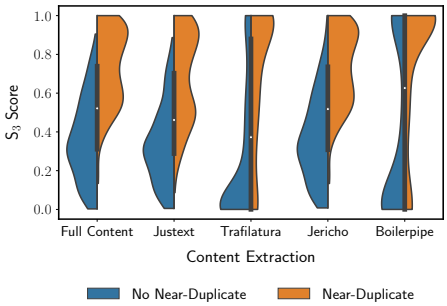
https://doi.org/10.5281/zenodo.6044976

Table 1: Overview of (a) precision and recall for near-duplicates in the uniform sampled document pairs at high syntactic similarity ($S_3 = 1$ and $S_3 \geq 0.9$), and (b) near-duplicates per $S_3$ similarity in the uniform sampled document pairs for all five content extraction methods. Lastly, (c) shows near-duplicates per $S_3$ similarity in document pairs with large $S_3$ differences to the full content extraction.

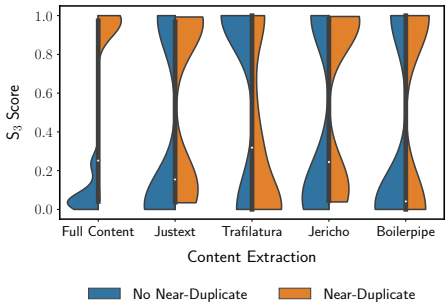
(a)

| | $S_3 = 1$ | | $S_3 \geq 0.9$ | |
|---|---|---|---|---|
| | Pr. | Re. | Pr. | Re. |
| Full Content | **1.00** | 0.06 | 0.98 | 0.26 |
| JusText | **1.00** | 0.07 | **1.00** | 0.24 |
| Trafilatura | 0.76 | 0.33 | 0.77 | 0.41 |
| Jericho | 0.93 | 0.07 | 0.97 | 0.26 |
| Boilerpipe | 0.60 | **0.50** | 0.61 | **0.59** |

(b)




(c)




high correlation to Justext and Jericho (0.97 respectively 0.99) and only a moderate correlation to Boilerpipe and Trafilatura (0.65 respectively 0.73). Overall, Figure 2 shows that the $S_3$ scores in our dataset differ substantially between the two groups, which motivates us to manually verify which of the document pairs are indeed near-duplicates.

## *Labeling Near-Duplicates*

After calculating the $S_3$ scores for all document pairs with all content extraction methods, we sample two sets of document pairs for manual review. First, we sample 100 document pairs uniformly covering $S_3$ scores between 0 and 1 for all five content extraction methods. Second, we sample document pairs with large $S_3$ differences between a main content extraction method and the full content extraction aiming at identifying document pairs where main content extraction yields opposite $S_3$ scores to full content extraction. Therefore, we select the 50 document pairs with the largest positive and largest negative $S_3$ difference for all four main content extraction methods for manual review.

We use the near-duplicate definition and review guidelines of Bernstein and Zobel [6] to label near-duplicates: A document pair is considered as near-duplicate when both documents are content-equivalent, and users would be able to extract the same information from either one for all reasonable queries. Two versions of the same Wikipedia article with only minor non-content changes are an example of near-duplicates under this definition.

We labeled the two document pair samples with two assessors. We applied a $\kappa$-test on the 100 document pairs sampled for $S_3$ similarities between 0 and 1 for the full content method, finding a high Fleiss' $\kappa$ of 0.78, indicating good agreement between both assessors. In a follow-up discussion among the annotators, we discussed all 11 document pairs with different near-duplicate judgments, finally agreeing in all cases. After our $\kappa$ test, each annotator judged the document pairs for the same two main content extraction methods for both user studies.

## *Evaluation*

Table 1a and Table 1b shows the ability of all five content extraction methods to identify near-duplicate documents in our set of 500 manually reviewed document pairs that uniformly cover $S_3$ scores between 0 and 1. In Table 1a, we report precision and recall for $S_3$ thresholds of 1 (for exact duplicates after content extraction) and 0.9 (highly similar extracted content). As in our initial discussion on similarity scores produced by the five content extraction methods, Trafilatura and Boilerpipe (the group with many document pairs at an $S_3$ score of 1 in Figure 2) as well as the full content, Justext, and Jericho (the group with fewer document pairs with an $S_3$ score of 1 in Figure 2) show similar behavior in terms of precision and recall. The full content and Justext approaches show a perfect precision of 1.0 at an $S_3$ threshold of 1, and Justext even has a perfect precision at an $S_3$ threshold of 0.9. On the other side, Trafilatura and Boilerpipe show a very high recall. Even for an $S_3$ score of 1, Boilerpipe achieves a remarkable Recall of 0.5.

Table-1b shows the correctly and wrongly identified near-duplicates per $S_3$ score for all content extraction methods in our set of 500 manually reviewed document pairs that uniformly cover $S_3$ scores between 0 and 1. Again, we can see similar behavior for the full content, Justext, and Jericho methods which make almost no mistakes at high respectively low $S_3$ scores. In the opposite group, with Trafilatura and Boilerpipe, we see quite some mistakes (even at $S_3 = 1$ and $S_3 = 0$).

Table 1c shows the correctly and wrongly identified near-duplicates per $S_3$ score for all content extraction methods in our set of 400 manually reviewed document pairs for which the main content extraction changes the similarity drastically. In almost all cases, barring few exceptions, we find that for such large differences, the $S_3$ score calculated on the full content correctly identifies near-duplicates and non-near-duplicates. This is visible since the full content method assigns, almost perfectly, non-near-duplicates an $S_3$ score near 0, and near-duplicates an $S_3$ score near 1. All other approaches make substantial mistakes in this selection of document pairs, indicated by assigning many non-near-duplicates an $S_3$ score near 1 (for which the full content method assigned scores near 0, since we selected large differences), and many near-duplicates an $S_3$ score near 0 (for which the full content method assigned scores near 1). Especially for cases in which highly similar documents get dissimilar main content extracted, we often found that the main content extraction had problems in identifying the correct main content. Overall, Trafilatura is the most vulnerable in this setting (the most near-duplicates near $S_3$ of 0, and most non-near-duplicates near $S_3$ of 1). Still, even main content extraction approaches with a very high correlation to the full content

https://doi.org/10.5281/zenodo.6044976



extraction method, like Jericho and Justext in our experiments, make substantial mistakes.

## CONCLUSION AND FUTURE WORK

We have recapitulated two risks and one potential application of near-duplicates in web search to motivate the maintenance of metadata on duplicate and near-duplicate documents. Given metadata on near-duplicates, it is easy to remove risks such as overestimated evaluation scores of retrieval systems or overfitting learning to rank models. Additionally, updating relevance judgments to the next version of the underlying web crawl can be done at lower costs because relevance labels might automatically be transferred to near-duplicates in the newer version.

In a first attempt to simplify the difficult decision of choosing an appropriate similarity threshold, we investigated how removing parts of documents that are rather irrelevant for the retrieval impacts the similarity of documents. Therefore, we have compared document similarities after preprocessing documents with five (main) content extraction methods. We found that main content extraction can yield very high recall for near-duplicate detection, even when only documents with identical main content are considered as near-duplicates.

An interesting prospect for future work is to include more main content extraction methods and expand the experiments to more document pairs. Another interesting direction for future work might be a further inspection of our observation that highly similar documents having very dissimilar extracted main contents were in most cases caused by mistakes in the main content extraction. This technique might help bootstrap a distant supervision dataset of documents with main content that is difficult to extract.

## REFERENCES


[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The Web Changes Everything: Understanding the Dynamics of Web Content. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 282–291, 2009.

[2] M. Alshomary, M. Völske, T. Licht, H. Wachsmuth, B. Stein, M. Hagen, and M. Potthast. Wikipedia Text Reuse: Within and Without. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, editors, *Advances in Information Retrieval. 41st European Conference on IR Research (ECIR 2019)*, volume 11437 of *Lecture Notes in Computer Science*, pages 747–754, Berlin Heidelberg New York, Apr. 2019. Springer.

[3] B. Alsulami, M. Abulkhair, and F. Eassa. Near Duplicate Document Detection Survey. *International Journal of Computer Science and Communications Networks*, 2(2):147–151, 2012.

[4] A. Barbaresi. Generic web content extraction with open-source software. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*, 2019.

[5] Y. Bernstein and J. Zobel. A Scalable System for Identifying Co-derivative Documents. In *String Processing and Information Retrieval, 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004, Proceedings*, pages 55–67, 2004.

[6] Y. Bernstein and J. Zobel. Redundant Documents and Search Effectiveness. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 736–743, 2005.

[7] A. Z. Broder. On the Resemblance and Containment of Documents. In *Compression and Complexity of SEQUENCES 1997, Positano, Amalfitan Coast, Salerno, Italy, June 11-13, 1997, Proceedings*, pages 21–29, 1997.

[8] M. Charikar. Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 380–388, 2002.

[9] J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 200–209, 2000.

[10] D. Fetterly, M. S. Manasse, and M. Najork. On the Evolution of Clusters of Near-Duplicate Web Pages. In *1st Latin American Web Congress (LA-WEB 2003), Empowering Our Web, 10-12 November 2003, Sanitago, Chile*, pages 37–45, 2003.

[11] D. Fetterly, M. S. Manasse, M. Najork, and J. L. Wiener. A Large-Scale Study of the Evolution of Web Pages. In *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*, pages 669–678, 2003.

[12] M. Fröbe, J. Bevendorff, L. Gienapp, M. Völske, B. Stein, M. Potthast, and M. Hagen. CopyCat: Near-Duplicates within and between the ClueWeb and the Common Crawl. In *44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021)*. ACM, July 2021.

[13] M. Fröbe, J. Bevendorff, J. H. Reimer, M. Potthast, and M. Hagen. Sampling Bias Due to Near-Duplicates in Learning to Rank. In J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1997–2000. ACM, 2020.

[14] M. Fröbe, J. P. Bittner, M. Potthast, and M. Hagen. The Effect of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, editors, *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 12–19. Springer, 2020.

[15] M. R. Henzinger. Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 284–291, 2006.

https://doi.org/10.5281/zenodo.6044976

[16] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate Detection Using Shallow Text Features. In B. D. Davison, T. Suel, N. Craswell, and B. Liu, editors, *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 441–450. ACM, 2010.

[17] G. S. Manku, A. Jain, and A. D. Sarma. Detecting Near-Duplicates for Web Crawling. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 141–150. ACM, 2007.

[18] A. Ntoulas, J. Cho, and C. Olston. What's new on the Web?: The Evolution of the Web from a Search Engine Perspective. In S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, editors, *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 1–12. ACM, 2004.

[19] C. Olston and S. Pandey. Recrawl Scheduling Based on Information Longevity. In J. Huai, R. Chen, H. Hon, Y. Liu, W. Ma, A. Tomkins, and X. Zhang, editors, *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, pages 437–446. ACM, 2008.

[20] J. Pomikálek. *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masarykova univerzita, Fakulta informatiky, 2011.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[22] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, F. D. Backere, F. D. Turck, K. Roelens, J. Decruyenaere, S. V. Hoecke, and T. Demeester. Overly Optimistic Prediction Results on Imbalanced Data: A Case Study of Flaws and Benefits when Applying Over-Sampling. *Artif. Intell. Medicine*, 111:101987, 2021.

[23] E. M. Voorhees. The Philosophy of Information Retrieval Evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5884312



# IMPROVED DISCOVERY AND ACCESS TO RESEARCH DATA IN ENERGY SYSTEMS ANALYSIS*

C. Hoyer-Klick, German Aerospace Center (DLR), Germany,
J. Frey, InfAI - Leipzig Universtiy, Leipzig, Germany

## INTRODUCTION

Research in the domain of energy systems analysis deals with the evaluation of future sustainable energy systems. This is often done by computational modelling of the energy flows, trading at energy markets and investment decisions of the various actors within the energy system. The analysis is therefore driven by data to very large extend. The modelling uses a lot of different input datasets from large variety of domains as e.g. engineering, economy, climate and societal developments. The models itself produce a lot of data as a result which contains energy balances, time series of energy provision or trading on energy markets, possible future investment needs into energy technology and much more. To be able to model the complex interactions of energy, markets and society, models more often get chained, so the output of one model becomes the input to the next model. Therefore, a lot of data is used and produced within the research domain by very heterogenous actors.

## IMPROVING FINDABILITY

Within the domain there is move towards more open data, but even then, all the datasets are stored on different places, in different formats with better or worse descriptive metadata. Within our project LOD-GEOSS, we try to use some lessons learned in the development of the Global Earth Observation Systems of Systems (GEOSS) and from linked open data (LOD) to improve the findability of all the research data. A central element is the DBpedia Databus which serves as a central communication platform and searchable catalog for the research data in our domain. If research data is made available suitable metadata along with a link to the dataset is registered to the Databus. The registration generates a unique id for the dataset which can be used to identify a specific dataset in the future. The metadata description and link to the data file improves the findability of the produced research data. A user of this data file can access the data though the provided link and reference this dataset by the unique id provided from the Databus. If a user republishes improved or derived data, a link to the originating data source is created. Data processing therefore gets traceable back to the originating source data. This can be complemented by data about the involved agents and activities to form a provenance graph of the data. The Databus and the distributed data storages therefore form a network of federated data bases of research data within our domain and improve the findability and access to the research data which is produced and used in our research area.

## IMPROVING INTERPRETABILITY

As described above, the data comes from a variety of different domains and interpreting the data is often a difficult task. In a parallel stream we contribute to the development of the Open Energy Ontology for the annotation of research data. This will create a common understanding of different data fields within the datasets and will ease data exchange within and across domains. Additionally, if data on the Databus is annotated with the ontology it will enable semantic searches for research data.
With our presentation we want to show the current status of the development of a distributed data architecture to improve findability and access to our research data.
This research is funded by a grant for German Ministry for Economics and Energy by grant number 03EI1005A.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6139647

# REQUIREMENTS FOR AN OPEN SEARCH INFRASTRUCTURE FROM THE PERSPECTIVE OF A VERTICAL PROVIDER

L. Martin*, F. Engl, A. Henrich, University of Bamberg, 96047 Bamberg, Germany

*Abstract*

A major advantage of an Open Search Infrastructure, as propagated by the Open Search Foundation[1], should be that it facilitates the development of special search solutions for special purposes, so-called verticals. In order to design a beneficial infrastructure in this respect, it is crucial to understand the requirements from the perspective of a vertical provider. Therefore, in this extended abstract we describe the requirements that result from our experiences with the IT-Altas Upper Franconia.

## MOTIVATION

Starting in 2014 we created a small vertical search engine for local IT companies[2] in cooperation with an IT business association in Upper Franconia [1]. There are some specific ideas implemented in this search engine: We do not search for documents (web pages) but for companies described by web pages. To derive the company ranking for a query, well fitting web pages vote for their company [2]. Furthermore, topic modelling [3] is used to improve the search results by injecting domain knowledge. The search engine result page (SERP) lists small automatically generated and query dependent profiles of the best ranked companies. In the remainder of this extended abstract we will sketch the requirements which arise if such a "company search" should be implemented on the basis of an Open Search Infrastructure.

## A COMMON INDEX TO SEARCH

To avoid the need for a custom index, the search for company web pages relevant to a query has to be executed on an index provided by the infrastructure. Of course, this initially requires limiting the search to pages of specific web domains. However, there are further wishes: since the pages found vote for the relevant companies [2], it seems reasonable to have not only pages of a domain itself, but also pages that are directly linked from this domain in the results. Furthermore, we mentioned above that we use domain-related resources (e.g. thesauri, topic models, or ontologies) to optimise the ranking. It should therefore be possible for the vertical to inject corresponding resources to the central index such that it takes them into account in the ranking, similar to the relationship of WordPress and Typo3 being WCMS.

## DOCUMENT DATA STORE

For generating snippets in the SERPs—which are usually query dependent—search engines commonly use a document data store [4, p. 16]. To allow for rich company descriptions in the SERP of a vertical it is necessary that the document data store contains not only information on web page level, but also on web site level and, where applicable, even on brand or company level. This could be achieved by interlinking among the different levels and connecting to knowledge graphs like Wikidata[3]. A rather specific aspect is that our current search engine integrates thumbnails of the companies' landing pages into the SERP. It can hardly be required from an infrastructure to provide such specific assets but this example shows that verticals will have the need to build their own add-ons to the document data store. Here, clear interfaces and update mechanisms are needed.

## CRAWL DATA FOR LANGUAGE MODELS

As mentioned above, the current version of the company search uses a topic model [3] to enhance the ranking and the result presentation. One could of course use other techniques as well but the shared requirement is to perform some type of corpus analysis on a well defined subset of the crawled web corpus. Regarding our search engine, this subset consists of IT-related documents (e.g. web pages) in German (we will not discuss the aspects of many web pages containing information in their local language and in English here). One solution for this feature request would be to allow the download of a defined subset of the corpus. A more resource efficient solution for the client (vertical provider) would be an API for executing the analyzer on the server side, similar to concepts like FaaS (e.g. AWS Lambda[4]) or Apache Spark[5].

## REFERENCES

[1] D. Blank, S. Boosz, and A. Henrich, "IT company atlas upper franconia: A practical application of expert search techniques," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, ACM, 2016, pp. 1048–1053. DOI: 10.1145/2851613.2851695.

[2] A. Henrich and M. Wegmann, "Search and evaluation methods for class level information retrieval: Extended use and evaluation of methods applied in expertise retrieval," in *SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021*, ACM, 2021, pp. 681–684. DOI: 10.1145/3412841.3442092.

[3] D. M. Blei, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, pp. 55–65, 2012.

[4] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010, vol. 520.

* leon.martin@uni-bamberg.de

[1] https://opensearchfoundation.org (accessed 15/06/2021)

[2] https://it-atlas-oberfranken.de (accessed 15/06/2021)

[3] https://wikidata.org (accessed 15/06/2021)

[4] https://aws.amazon.com/de/lambda (accessed 15/06/2021)

[5] https://spark.apache.org (accessed 15/06/2021)

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6148347



# CONCEPTUAL CONSIDERATIONS FOR COMPREHENSIVE AND COOPERATIVE CRAWLING AND INDEXING THE WEB

S. Voigt[1], Open Search Foundation, Germany
M. Granitzer, Passau University, Germany

*Abstract*

The web is a very large and dynamic digital database. Currently there are about 341 million domains registered in a bit more than 1500 top level domains (TLDs)[2]. The .com TLD alone accounts for 155 million registered domains (46%). As few as 65 TLDs together account for 95% of all registered domains. Assuming an average of 250 documents/pages per domain, the current Web consists of an estimated 85 billion documents or pages. Beyond this, the Web of course stores petabytes of media data with images, films, voice/music recordings, social media data as well as petabytes of public data bases with scientific data, satellite imagery and administrative data in the 'deeper' part of the web.

As of today, the Web is a very dynamic digital ecosystem, which is influenced by different organisations, governments, individuals and a few monopolistic commercial entities, with different, often contradicting interests. This makes the Web a very innovative, however, partially even toxic environment. Particularly, since a few monopolistic gatekeepers control access to information on the web and shape corresponding access patterns, while the innovation potential and core value is generated by the enormous number of individual or organisational contributors. This structure puts pressure on those small contributors in science, economy, art, culture, media and society, to follow either the rules of the gatekeepers, or get lost in the vastness of the digital space. Consequently, information as public good, with free, open and unbiased access – one of the core principles that the Web was built on – currently is in the hand of a few corporate entities along with the (commercial) interest of their shareholders.

Based on this analysis we argue, that unbiased access to digital data and information requires a collaborative effort of a variety of different organisations to build an open index and metadata pool of the Web. Such an open web index will be the basis for open search engines, a diversity of public and commercial web services, science, training of AI and many future applications we can't even think of today – while at the same time decreasing the dependency on the current monopolistic gatekeepers.

In this talk we discuss concepts and approaches for aligning efforts, currently ongoing in parallel, in an open and democratic manor, to jointly build and maintain a comprehensive and open index for the web. Such an activity will involve a large number of public, scientific and commercial organisations in a large-scale cooperative effort.

Generating a complete index of the Web is typically beyond the capacity of single organisation, not only because of the high technical complexity, but more so, due to the high associated costs. However, we argue that one can generate a comprehensive web index and web data pool by combining ongoing and dedicated efforts of individual organisations and by integrating public computing facilities into the task. To achieve this, we suggest concentrating first on the surface web and aiming at 80% of completeness for a large-scale cooperative web-crawling and indexing campaign. To set this up, two key questions need to be answered: (i) What is the most efficient way to share and distribute such a comprehensive web crawling and indexing effort, while minimizing network load and ensuring energy-efficient hosting, sharing and updating of the distributed web index? (ii) How to develop a value-oriented service-based ecosystem - beyond monopolistic actors - around such an open web index and how to recover the costs for generating and maintaining such a web index?

There are different strategic options for distributing the tasks in a scalable peer to peer framework, allowing for an overall trade-off between individual/voluntary support and achieving the common greater goal. Crawling and indexing goals of contributing computing and science centres may be different, however, as long as the synergy with the larger group is beneficial, there is a win-win situation in contributing to the larger group effort and its goals. It will be important to find a good match between such voluntary participation and financial remuneration to sustain the needed computational capacities.

Technically we suggest a decentralised approach, splitting the overall task vertically, i.e. along the necessary computing tasks, and horizontally, i.e. focusing on different sub-parts of the Web. Vertical splits involve coordinated crawling of sub-parts of the Web, storage (and access) to crawls (i.e. via WARC files), processing and enriching crawls as well as the coordinated generation of indices and sub-indices. In order to generate meaningful, manageable and shareable web crawls, web repositories, web indices and web graphs, the task has to be shared among dozens of computing facilities. It will be important to consistently enrich the web indices with a substantial and extensible number of pre-processed signals and attributes, such as nominal age group, ethical annotations,

[1] sv@opensearchfoundation.org

[2] https://research.domaintools.com/statistics/tld-counts/ (access on 18.6.2021)

comprehensive geo-tagging and many more. An open web index will fuel not only the creation of specific search engines, but also will also reduce efforts in creating AI products, like for example knowledge graphs or neuronal language models. In return, those efforts can contribute to improving the index of the web.

In the talk we discuss different strategies for dividing the generation and maintenance of a global, still distributed, open web index. Candidate factors are: division by language, TLD, geographic region, topic/application field, network topology/latency, hosting facilities etc. Of course, also the type of contributing computing centres and overall funding and governance scheme of the cooperative effort will be of relevance for sharing of tasks: Public science and computing centres may have a different mandate and motivation to contribute to the joint undertaking than e.g. private Internet service providers, industry or fully commercial computing facilities.

Finally, we discuss, how the division of tasks needs to ensure that different individual web repositories and indices have relevance in itself and can be searched, queried and accessed by the cooperating entities in an energy efficient and sustainable way.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5886647



# THE DEVELOPMENT OF A SOCIAL-MEDIA-STRATEGY FOR THE OPEN SEARCH FOUNDATION APPLYING THE SOCIAL-MEDIA-CYCLE

A. J. Decker, Technical University Ingolstadt, Ingolstadt , Germany and
Open Search Foundation, Starnberg, Germany

*Abstract*

On the Second Open Search Symposium in 2020 the foundations for a tailor-made OSF communication approach were presented [1]. As demonstrated, for an unknown Non-profit organization (NPO) like the OSF the necessary focal points in the beginning of the communication approach must be the set-up of awareness and attention for the existing monopoly-problem in the search market as well as the building of a group of supporters. Due to the fact, that the OSF still lacks the necessary funding to run big campaigns, alternative ways must be found. In this context, Social Media plays a big role, because major steps can be taken towards the goals described without big investments in terms of money. A solution can be provided, that can be executed with the help of the existing volunteer members.

Talking about Social Media, one thing can unfortunately still be observed in corporate practice today: there is a lack of competence in developing Social Media strategies. Michelle Charello [2], author of the American textbook "Essentials of Social Media Marketing" summarizes the situation in the market as follows: „Too many businesses struggle with social media because they lack a well defined social media strategy."

As a consequence, particularly an NPO like the OSF should not start the adventure Social Media without such a well defined Social Media strategy. Just like Warren Buffet said: "It takes 20 years to build a reputation and five minutes to ruin it." This is true for all companies, but it is especially true for an NPO that pursues charitable goals.

Hence, the first step in Social Media for the OSF must be the profound development of a Social Media strategy. In order to do so, a project at the Technical University of Ingolstadt from March until June 2021 with 26 master students of the Marketing / Sales / Media program was set up under the coordination of Professor Alexander Decker. As the basic framework to develop such a Social Media strategy the so called Social-Media-Cycle [5] was used. This ten-step approach guides companies systematically through all the necessary strategic and operative steps. Out of the ten steps of the Social-Media-Cycle, six were executed in the project:

- Step 1: Social Media monitoring: all relevant social media platforms were observed with regard to mentions of the OSF and the most important alternative search engines.
- Step 2: Definition of objectives per target groups: Based on the findings of the monitoring and the objectives and target groups of the general OSF communication approach, nine target groups were identified and further described in detail (including the objectives), focussing on both, people we need to address with regard to the awareness problem, and those we want to gain as supporters.
- Step 3: Selection of the focus platforms used by the OSF: The target groups identified with their respecting objectives led to the selection of three platforms: LinkedIn and Twitter to start with. Instagram to follow later on.
- Step 4: Organisational aspects: Due to the lack of resources, organisational aspects had to be taken into consideration as well. Particularly, a social media integration model (following the well-known Altimeter approach) had to be chosen and the roles of the Social Media team had to be defined.
- Step 5: Iteration: All information from step 1 to 4 were put together and iterated in order to come up with a first Social Media strategy approach. This resulted in a Social Media architecture, that encompasses the final seven (out of nine) target groups, the objectives to be achieved, the channels chosen for them and first indications of the content to be produced in step 7.
- Step 7: Content calendar: A content calendar for the three chosen platforms was developed with ideas and first posts as well as a set of five explanation videos.

Due to the fact that the project is still ongoing and the final strategy has yet to be approved by the OSF board, the operative steps 6 and 8 to10 could not be implemented so far.

The presentation will show the major outcomes of the project and the strategy developed in more detail.

## REFERENCES

[1] *Decker, A. / Hiemer, C. (2020):* Beyond Tech: Rising Awareness for the Open Search Foundation through a Tailor-Made Communication Approach. *Proceedings for the 2. International Symposium on Open Search, CERN, Geneva, Switzerland. Electronically published via:* https://doi.org/10.5281/zenodo.4593332.

[2] *Charello M (2020):* Essentials of Social Media Marketing. The Most Up-to-date Social Media Marketing E-textbook. *Stukent, Idaho Falls.*

[3] *Decker, A. (2019)* Der Social-Media-Zyklus. *Springer-Gabler, Wiesbaden.*

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6204092

# THE EFFECT OF SEARCH ENGINE OPTIMIZATION ON SEARCH RESULTS: THE SEO EFFECT PROJECT*

S. Schultheiß†, S. Sünkler, Hamburg University of Applied Sciences, Hamburg, Germany

## INTRODUCTION

On search engine result pages (SERPs), numerous actors can influence the visibility of results, one of them being search engine optimization (SEO). SEO is a multi-billion-dollar industry and defined as "the practice of optimizing web pages in a way that improves their ranking in the organic search results" [1]. Despite this importance, little is known about its impact on result rankings and user perspective on SEO, which is what our project SEO Effect[1] addresses. The project has components that focus on *users* and *SERPs* (Fig. 1), yet they are directly related to each other. In the following, we will outline how we have approached the project goal.

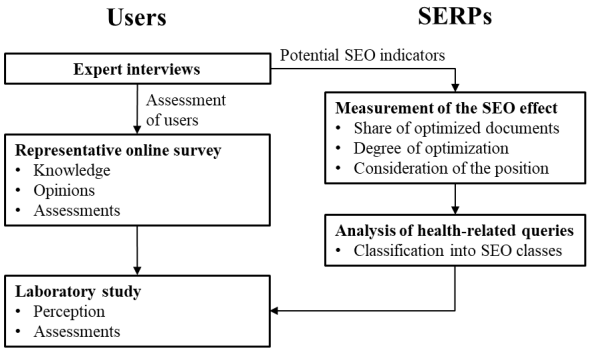



Figure 1: Project parts within the SEO Effect project.

## MAJOR RESULTS

### *Expert interviews*

As a basis for further project parts, we conducted expert interviews with stakeholder groups involved in search engine rankings: Search engine optimizers, content managers, and online journalists [2]. The interviews aimed to gather assessments of user perspectives regarding SEO. The interviewees assumed that SEO is barely known to the users and that the user opinions of SEO strongly depend on their SEO knowledge. From these assumptions, hypotheses were derived for the online survey described later. The interviews also served the SERPs part of the project, providing us with valuable clues on how SEO can be identified on websites.

### *Measurement of the SEO effect*

We developed a multidimensional approach to make the SEO effect measurable and implemented this approach in a software tool that detects on a URL whether SEO measures have been taken [3]. For our approach, we use a model of $n = 48$ indicators based on an extensive literature review and the aforementioned interviews with SEO experts [2]. The model is the basis for a decision tree classifier to determine the probability of SEO.

### *Representative online survey*

We conducted a representative online survey with $n$ = 2,012 German Internet users to get insights into the user perspectives on SEO [4, 5]. The results widely confirm the assumptions from the interviews. Less than half (43%) of Internet users know that ranking improvement is possible outside of paid ads, and only 8% are familiar with the term "SEO." Due to this ignorance, SEO was rarely associated with organic results on SERP screenshots. With increasing knowledge of SEO, a more positive opinion could be observed. In the laboratory study described later, we revisited some elements of the survey, such as querying the SEO knowledge.

### *Analysis of health-related queries*

We applied our approach to determine the probability of SEO on several datasets. One of our evaluation was conducted in preparation for our laboratory study. This involved automated analysis of $n = 318$ health-related search queries with a total of $n$ = 22,426 search results from Google. The analysis showed that 32.6% of the search results were most probably optimized, 46.8% were probably optimized, and 20.7% were (probably) non-optimized. In terms of result positions, the distribution of optimized and non-optimized documents is equally distributed.

### *Laboratory study*

The aim of the laboratory study is to investigate *whether* and *which* quality differences users perceive between optimized and non-optimized websites. Thus, this study brings together results of both project parts and consisted of evaluations of website quality based on various criteria (e.g., trustworthiness, expertise) plus think aloud protocols. The results show that (probably) non-optimized websites are rated as having a higher level of *expertise* than optimized websites. This assessment is independent of the subjects' SEO knowledge and was mainly justified with the competent and reputable appearance of the operator in the case of non-optimized websites (e.g., websites of ministries).

## CONCLUSION

Both subprojects have directly benefited from each other by triangulation of methods. The structure of our project thus proved to be fruitful, as we were able to examine the SEO effect from multiple angles. Further questions and starting points for follow-up projects are derived from the results. These include the influence of SEO on users' knowledge acquisition and the development of a more differentiated SEO classification.

* Funding: This work is funded by the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft), grant number 417552432.

† sebastian.schultheiss@haw-hamburg.de

[1] https://searchstudies.org/research/seo-effekt/

https://doi.org/10.5281/zenodo.6204092



## REFERENCES

[1] K. Li, M. Lin, Z. Lin, and B. Xing, "Running and Chasing -- The Competition between Paid Search Marketing and Search Engine Optimization," in *2014 47th Hawaii Int. Conf. on System Sciences*, Waikoloa, HI, January 2014, pp. 3110–3119. doi: 10.1109/HICSS.2014 . 640.

[2] S. Schultheiß and D. Lewandowski, "'Outside the industry, nobody knows what we do' SEO as seen by search engine optimizers and content providers," *J. Doc.*, vol. 77, no. 2, pp. 542–557, Dec. 2020. doi : 10. 1108/ JD- 07- 2020- 0127.

[3] D. Lewandowski, S. Sünkler, and N. Yagci, "The influence of search engine optimization on Google's results: A multi-dimensional approach for detecting SEO," in *WebSci '21: Proc. 13th ACM Conference on Web Science*, June 2021, to be published. doi : 10. 1145/ 3447535. 3462479.

[4] S. Schultheiß and D. Lewandowski, "Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference," *J. Inf. Sci.*, May 2021. doi : 10. 1177/ 01655515211014157.

[5] S. Schultheiß and D. Lewandowski, "(Un)bekannte Akteure auf der Suchergebnisseite? Ein Vergleich zwischen selbst eingeschätzter und tatsächlich vorhandener Suchmaschinenkompetenz deutscher InternetnutzerInnen," in *Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proc. 16th Int. Symposium of Information Science (ISI 2021),* Regensburg, Germany, March 2021, pp. 218–246. doi : 10. 5283/ epub. 44946.

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5884441

# AVOIDING USELESS CONTENT WHILE CRAWLING THE WEB

O. Behrendt*, A. Hierle, infotiger UG, Munich, Germany

*Abstract*

Practical experience with a web crawler showed that while processing huge amounts of data is not easy, the task of separating waste from precious information can be even more challenging. Here we present the basic design of a real-world web crawler and focus on practical strategies to deal with useless content.

## INTRODUCTION

Crawling the web is a challenging endeavor and as such an hurdle for start-ups of web search engines (SE). An independent, open and freely available web crawler index would not only strengthen the right of free speech, decrease dependency on monopolies but also triggers innovation. Sadly so far no such index exists so that crawling is a precondition to build a SE. In this presentation we focus on *adaptive domain limiting* (ADL), a heuristic developed by the authors and used by the infotiger crawler. We define useless content (web pages) to be link farms, unlimited subdomains or content not in the focus of the SE. Obviously this definition is partially based on subjective choices.

## WEB CRAWLER OVERVIEW

Figure 1 shows core components and data flow in a single web crawler node. The parser removes HTML mark-up, decides which text blocks are boilerplate and creates fingerprints of ”good” text blocks. *Trace vectors* encode parsed documents in a semi-metric topic space which serve as input for similarity search or for categorization (not shown). Link analysis is used by the SE for ranking but also by the crawler in context of ADL.

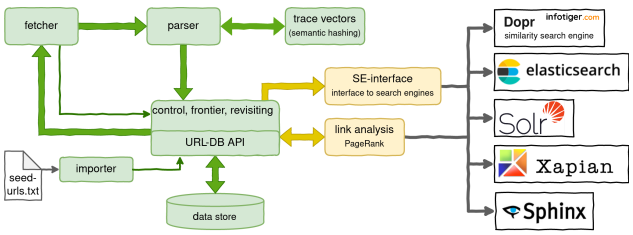



Figure 1: Overview of components and data-flow of a single web crawler node.

## ADAPTIV DOMAIN LIMITS

Experience with simple breadth-first search showed that the web crawler eventually becomes trapped in useless content like link spam (compare with [1]). Since a crawler has to be polite and the number of parallel requests is limited the download speed will decrease. Consequential impacts of crawling useless content include inflating the data store and query index, reducing revisiting frequency and lowering query precision. Domain limits (DL) appear to be a promising approach to drastically reduce useless content. ADL is defined as the maximum allowed number of pages for a single domain. When starting a new crawl all DLs are set to a low number like 2000. When the limits are hit, we want to increase them *only for valuable* domains. Solely manual inspection is not feasible due to the number of hit limits on multiple crawler nodes. Increasing the DLs for domains with a SiteRank[2] higher than a given threshold (e.g. 90% if SiteRank is converted to quantiles) proofed to be a good heuristic to adapt DLs for ”valuable” domains.

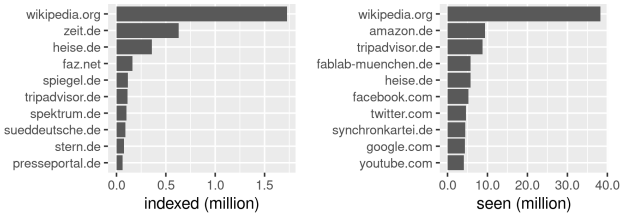



Figure 2: Left side: Top ten domains ordered by number of indexed pages per domain after applying ADL. Right side: Top ten domains ordered by number of seen links per domain, which indicate the situation when ADLs were not applied. Data taken from a single crawler node.

## CONCLUSION

We encountered two main problems with the ADL approach using SiteRank. Firstly it is vulnerable to link spam and secondly SiteRank cannot identify all useless content. The first problem could be addressed by a preprocessing step to detect and remove link spam from the document graph. The latter problem depends on the definition of useless content. For example if a SE does not want to index internet shops, the number of forward links could be a hint for detection, since shops are often bad hubs. For other unwanted content like advertising spam a deeper content-related analysis might be needed. Even with limited resources a web crawler that produces high quality input for a SE can be successfully implemented. Experience showed that one key factor are effective heuristics to reduce useless content.

## REFERENCES

[1] H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov, “Irlbot: Scaling to 6 billion pages and beyond,” *ACM Transactions on the Web (TWEB)*, vol. 3, no. 3, pp. 1–34, 2009.

[2] G. Feng *et al.*, “Aggregaterank: Bringing order to web sites,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 75–82.

* ossym21@infotiger.com

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5886715



# PROCESSING CRAWLED DATA

M.A.C.J. Overmeer MSc (Mark)*, MarkOv Solutions, Arnhem, The Netherlands

*Abstract*

Building an internet-wide search engine starts by collecting and pre-processing the content of webpages. There are dozens of initiatives to crawl for page content to build a text search interface. However, there is a huge difference between performing a text search as such, and presenting quality search results. To be able to put a label "best match" on a fragment of text, a considerable number of additional complex algorithms must be implemented. Those require far more resources than the full text search database.

The "Crawl Pipeline" project combines many kinds of information collection activities: provides those additional algorithms with the facts they need for their work. The Pipeline is also used to collect facts which are not search engine related, like counting web-server types, detecting expired certificates, or discovering phishing sites.

## PIPELINE

Project "Crawl Pipeline"' provides a base infrastructure to extract information from web content. The software may be run as post-processing step on a web-crawler, but in its initial set-up, it runs on dedicated hardware. It is able to process Terabytes per day per pipeline instance, potentially with a few dozen pipelines in parallel.

A crawler instance has a list of URLs (pages to visit) as input. For each URL, the crawl sends a request to a website. The response contains some data we need. Meanwhile, metadata is collected too. The triplet (request, response, metadata) is often written into WARC archives[1], or into a database. At that moment, it becomes a static data-set. The Pipeline transforms those static data-sets back into a stream of collection actions for interested parties to filter and extract into the data they need.

The Pipeline spends time to reconstruct the crawl, just as each user would need when processing the data-set on their own. Probably even a bit more time. But once the crawl is restored, it is reused for many purposes. Those activities are abstracted into Tasks: per end-user a separate Task.

## (EXTERNAL) TASKS

Researchers who need data for their investigations can submit "external Tasks" for the Pipeline. OSF search engine support components will also implement Tasks to enhance its knowledge.

Each Task description contains:

- filter rules to select answers;
- simple data extraction steps; and
- packaging instructions.

The Task will reduce the amount of data for their requesters down to only a few percent of the original data size. Usually from Terabyte down to Gigabyte scale per day.

Many research projects can get their minor subset of crawl-data without the expense of processing Terabytes of data themselves. Besides, they do not need to implement communication with each crawler instance which will emerge in the near future.

## RUNNING THE PIPELINE

The main components of the pipeline instance are

- incoming feed collects the static crawl data-sets published by crawlers;
- batch control manages a set of processing queues;
- each queue reconstructs one crawl result set at a time;
- each result is passed to each of the defined Tasks sequentially;
- when a result is selected for a Task, the related extract is made;
- after a short time, at most a day, the extracts are packaged in the way which is also defined in the Task; and finally
- the packaged results must be retrieved within a day, otherwise they get lost.

## PRESENTING

The presentation will show which kind of Tasks can be run on the "Crawl Pipeline": how can you use the facility to support your own research. It will also show challenges and opportunities for (near) future extensions.

## ACKNOWLEDGEMENTS

This project is made possible by a generous donation from the NLnet Foundation.

## REFERENCES

[1] The WARC File Format – ISO 28500

* mark@overmeer.net

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6044920

# FROM WEB GRAPHS TO PRIORITIZING WEB CRAWLS

S. Nagel, Common Crawl

*Extended Abstract*

This talk describes how web graphs are used at Common Crawl to prioritize which sites and pages are visited by the crawler.

The Common Crawl data sets are sample collections of web pages with no intention to mirror websites completely. In order to achieve a balanced, both diverse and representative sample, the crawler is "steered" to web sites found to be relevant by analyzing the hyperlink structure of preceding crawls. Harmonic centrality and page rank scores are calculated on hyperlink graphs and define the authority of a web site both on the level of hosts and registered domains. The authority is then mapped to a likelihood that a URL from the corresponding site is sampled and to a "crawl budget" which limits the max. number of pages crawled.

The talk will cover the following points:

1. we start with a short introduction of Common Crawl, the goals and give an overview of the crawl technology used between 2008 and 2021.
2. we present the Common Crawl in-house webgraphs and rankings (2017 until now) and how they are constructed, starting from the extraction of hyperlinks, the aggregation on the level of host and domain names and the transformation into a numeric graph representation. We will also look into prior work of building hyperlink graphs from the Common Crawl: page ranks used in the 2012 crawl and earlier, the graphs and rankints from the "Web Data Commons" and "Common Search" projects, and the webgraph framework developped at the University of Milano.
3. the Common Crawl rankings are compared with other open accessible web site rankings – the top-1-million sites published by Alexa, the "Cisco Umbrella Popularity" list, "The Majestic Million" and the Tranco list.
4. we show how the authority of a site in terms of the harmonic centrality rank is used
5. to sample hyperlinks,
6. to define a "crawl budget" per domain (how many web pages or subdomains the crawler is allowed to crawl from this domain)
7. and how domain-level score can be projected to the level of web pages
8. we discuss the challenges associated with the use of link-based centrality measures, namely link spam and other attempts (eg. aggressive SEO) to influence the domains ranks by artificially creating sites, pages and hyperlinks. We look into a few spam clusters, demonstrate how these can be identified and which strategies can be used to prevent the crawler from hitting and getting trapped in spam clusters.
9. finally, we evaluate how the usage of centrality measures impacts the crawled data:

a. we outline the preliminaries and constraints of the Common Crawl data sets: the focus on HTML pages and the 1 MiB content limit, the impact of operating a crawler from a data center in North America, the need for sampling and shuffling and the immediate release of the data.

b. multiple aspects of representativity are discussed in order to define an evaluation baseline: the domain-level coverage compared to different crawling strategies (breadth-first, depth-first), regional coverage (top-level domains, content languages), the amount of duplicates and other.

c. we analyze the crawls 2017 – 2021 and evaluate how the crawled data fits the various aspects of representativity.



ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.5888198



# NEUROPIL – A DISTRIBUTED, PRIVACY-PRESERVING, SEARCH INDEX STRUCTURE

Stephan Schwichtenberg, pi-lar GmbH, Cologne, Germany

## *Abstract*

Neuropil is an open-source de-centralized messaging layer that focuses on security and privacy by design. Persons, machines, and applications first have to identify their respective partners and/or content before real information can be sent. The discovery is handled internally and is based on so called "intent messages" that are secured by cryptographic primitives. Our project aims to create distributed search engine capabilities based on neuropil, that enable the discovery and sharing of information with significantly higher levels of trust and privacy and with more control over the search content for data owners than today's standard.

## *Setting The Scene*

As of now large search engines have implemented "crawlers", that constantly visit webpages and categorize their content. The only way to somehow influence the information that is used by search engines is by using a file called „robots.txt“. Details about algorithms and especially their parameters are only known to the search engine provider. Furthermore there are a couple of misalignments in this model: Content hosted at a central search engine provider is probably outdated. Any search engine provider has to grow with the size of the connected information sources, which triples energy costs: not only the data owner holds information on his servers or electronic devices, the search engine provider needs to crawl the information and needs to store a copy of the material obtained. And although all this happens, nobody can be sure whether his information will appear as a search result: any search engine provider can (and for some parts must) withhold certain information, e.g. due to legal constraints.

## *A Different Approach*

The neuropil messaging layer already uses a highly standardized "intent" format that protects the real content of users, i.e. the public keys contained in the intent messages are used to encrypt before any content is sent. These digitally secured “intent” (loosely modelled after JWT) are the public parts of the messaging layer, and therefore can be shared without any impact on copyrights. Using these “intent” token the above mentioned model is reversed: data owners define the searchable public content, and data users can discover available data sources. Because data owners would like to be found, we ask them to host a proportion of our new, distributed data structure: Based on the available “intent” token we derive cryptographic long term key (based on CLKHash* / PPRL**) and a new index hash (based on mmhash / LSH / LPH), which allows to distribute data in a privacy preserving, secure manner and thus enable each user / data owner to participate and maintain a search index and contents. We believe that it is thus possible to build a distributed search engine database that is able to contain and reveal any kind of information in a distributed, concise and privacy preserving manner, without the need for any central search engine provider.

## *Status Quo*

In 2019 our project has received funding from the NGI ZeroDiscovery project. As part of this project we have not only improved the way how discovery for specific data models work, we also have implemented the required search capabilities in terms of algorithms and data structures into the Neuropil messaging layer. Our current experiments simulate a distributed setup ranging form 256 to 4096 nodes, and works well on smaller datasets. On a standard PC we are able to distribute one million of data records, and achieve a query time of ten milliseconds, but our code is running without any optimizations which leaves room for improvement.

## *The Road Ahead*

In the upcoming months we would like to utilize larger datasets, in terms of “number of records”, “number of index entries” and “data size” to validate our distributed structure. We are also looking out to distribute the created index structure throughout our Neuropil network. Even if our index structures is keeping its promises, we already see a couple of open questions that we have not been evaluated yet: Will energy consumption in our fully distributed setup be really less than crawling sites? Managing trust in such a setting will requires new paradigms beyond PKI / Web Of Trust, but as of now we think that TSA (Time Stamping Authority) could be a way to move forward. How can a synchronized understanding of time between all nodes be established. And how can illegal content be banned from the search index? How will it be possible to establish such a distributed structure, when different organizations and data-owners are co-creating and participating a fully distributed setup? As of now we could just end up with a few more technical capabilities.

## *Our Presentation*

In our OSF presentation we will show our work in terms of algorithms, data structures and communication topologies so far. We hope that we are able to conclude our NGI Zero Discovery project until October, so that participants can experiment and add content with our implementation themselves. We also would like to give an outlook how our approach could work in alignment with other technical initiatives (crawling / pipelining / preprocessing) in the OSF-Technology working group.

* see also https://clkhash.readthedocs.io/en/stable/index.html

** Efficient private record linkage of very large datasets

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6168831

# INDICO & CITADEL SEARCH: A COLLABORATION CASE STUDY

C. Antunes, P. Lourenço, A. Mönnich, A. Wagner, M. Kolodziejski,
P. Panero, P. Ferreira, CERN, 1211 Geneva 23, Switzerland

*Abstract*

Modern web applications, and content management systems, in particular, are expected to provide their users with simple and efficient methods to query data. While most modern relational databases do provide full-text indices, those are often not sufficient for rich querying of application contents, especially when compound queries are required. Moreover, they often provide mostly bare-bones functionality which has to be expanded on by the application developers. It does then pay off to offload full-text search on specialized systems which are capable of performing this task efficiently.

This paper introduces one such system, the *Citadel Search* service, and documents an attempt to use it to provide an efficient, transparent and open-source search solution for *Indico* - an open-source tool for event organisation, archival and collaboration built at CERN and deployed around the world.

For many years, *Indico* provided no out-of-the-box search engine and instances relied on locally-developed custom search strategies or services, often based on paid enterprise solutions, to fulfil this gap. At CERN, a proprietary search implementation based on *Sharepoint Search* [1] was used. Proven useful in the past, as it relieved the service from the complexity of maintaining an external search engine, it had the downside of not being reusable by the community and being based on a commercial product, which limited its affordability.

With more than 10 million documents at CERN, including events, attachments and metadata, *Indico* needed not only to preserve the quality of the existing search functionality, but also provide long-requested improvements on it, such as, better quality of results and a native interface.

*Citadel Search* is an open-source enterprise search solution that makes use of state of the art technologies such as *Elasticsearch* [2], *Tika* [3] and the *Invenio Framework* [4] for large-scale digital repositories. It is used by several large document collections at CERN, namely CERN's Engineering Data Management Service (EDMS), and by CERN's large information space made of more than 14000 Web sites.

Since it was announced [1], the major highlight was the addition of a new content extraction Application Programming Interface (API) - using *Tika* and *Celery* workers, and deployment templatization of *Citadel* instances with *Helm*[5], simplifying the adoption for new users.

*Citadel* focuses on transparency, efficiency and empowering its "user systems" by giving them control over their data and associated querying strategies. Built with collaboration in mind, it strives to offer to users a flexible solution for structured searching with document level Access Control Lists (ACLs), and structured data obtained with a web crawler. Custom-tailored data models can be created for different information sources, and fine grain access control is provided, as to obtain relevant results to search queries.

Our approach details how *Indico* overcame the challenges involved in designing a new search module on top of *Citadel*, adapting several heterogeneous records into structured documents, namely: events, contributions and materials. In addition, we explain how the search engine was adapted to cover multiple use-cases from a simple text-based search to advanced aggregations and filters.

[1] https://docs.microsoft.com/en-us/sharepoint/dev/general-development/search-in-sharepoint
[2] https://www.elastic.co/what-is/elasticsearch
[3] https://tika.apache.org/
[4] https://inveniosoftware.org/
[5] https://helm.sh/

## REFERENCES

[1] *Citadel Search: Open Source Enterprise Search*, Open Search Symposium 2019, Garching / Munich, Germany, Oct. 2019. https://doi.org/10.5281/zenodo.3581157

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6139956



# OPEN SEARCH @ DLR - TOWARDS TRANSPARENT ACCESS TO WEB-BASED INFORMATION IN SCIENCE

S. Voigt[1], German Aerospace Center, Oberpfaffenhofen, Germany
T. Hecking, German Aerospace Center, Köln, Germany
D. Jankowski, OFFIS Institute for Computer Science, Oldenburg, Germany
J. Möller, University of Oldenburg, Germany
M. Schwinger, German Aerospace Center, Oberpfaffenhofen, Germany

*Abstract*

Data is the raw material of the 21st century - for research, innovation, economy and society. Digital sovereignty requires free, uninfluenced & traceable access to information - in other words, open Internet search and systematic access to web data. Currently, there is a monopoly in information search: In Europe, more than 90% of all Internet searches are conducted via a single commercial and advertising-optimized search engine. This holds immense potential for intentional or unintentional manipulation in access to data, information, technology and knowledge (cognitive/economic bias). Especially for science, new concepts for a distributed and open Internet search infrastructure are needed.

The wealth of data and information on the web must be rendered more accessible through uninfluenced discovery of scientific data and information, since it is the basis for free research and innovation. Against this background, the German Aerospace Center (DLR) is contributing to the European Open Search Initiative, formed by science and computing centres. Within the Open Search @ DLR project, existing in-house capacities and know-how in data access and search are identified and pooled to set-up a cooperative crawling, indexing and search capability to web data repositories – internal and external to DLR. Furthermore, dedicated pilot applications in areas such as information retrieval, knowledge management or information evaluation and transparency, making use of the infrastructure, are developed in the project.

A primary focus of the Open Search @ DLR project is networking of in-house expertise as well as connecting with the Europe-wide Open Search Initiative.

Within this talk we present the project layout and findings during the first project phase. This includes inventorying of in-house data and heterogeneous information repositories, coordinated crawling, indexing and searching. We present architecture and set-up of a testbed for cooperative crawling, where single crawling nodes communicate URLs to crawl in a peer-to-peer fashion as basis for joint assembly of large corpora of web data.

In a second part of the talk scientific pilot applications of an open search infrastructure are discussed, including the use of georeferenced data from web- and database sources, e.g. for monitoring of news, events, geospatial analysis and early warning. Furthermore, open search approaches for exploring, linking, and indexing of information from heterogeneous scientific data sources and public web content are particularly being addressed. This includes access to (semi-)structured information in databases as well as information extraction from texts, e.g. automatic geo-tagging. In this context, especially the establishment of geographical connections between scientific, structured databases and human-readable content from the Internet play an important role.

In the last part of the talk first ideas and concepts for a long-term activity of science and computing centres to set up an open Internet search ecosystem are discussed. Such a shared activity should be based on cooperative computing, open-source software stacks and public moderation and should involve distributed scientific high-performance computing and cloud facilities forming a cooperative open search infrastructure to warrant a long-term, public and open web search environment.

As long as the digital sphere – the web – exists, free and unbiased orientation therein has to be ensured to guarantee free and unbiased access to information for science, economy and society as a whole.

1 stefan.voigt@dlr.de

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6205256

# UNDERSTANDING WEBSITES

M.A.C.J. Overmeer† MSc (Mark), MarkOv Solutions, Arnhem, The Netherlands
К. Беров* MSc (Krasimir Berov), Studio Berov, Byala Slatina, Vratza, Bulgaria
R. Lam‡ MSc (Ronny), HNW.NU, Gasselternijveenschemond, The Netherlands

*Abstract*

Bluntly crawling for all the web-pages you can get has many disadvantages. Unfortunately, it is an approach where most existing crawl projects are stuck. Being smart during crawling and smart while processing the results is really difficult to achieve: this requires a wide variety of data which are not readily available at the moment.

To name a few quality enhancing contributions: how can you stay away from phishing sites, exclude erotic content, avoid SEO-spam networks, or keep fake news from polluting your collection? How can you crawl behave kindly for the website, so you are not locked-out by robots.txt rules? Which websites are better, compared to websites on the same subject? What are the aliases of a website, which you do not need to crawl again? How often should we visit a page, to have fresh data?

We need to know more about a website before we crawl it. We need to know more about a website to be able to present it to a non-research audience.

## GOOGLE'S ADVANTAGE

The current situation is that Google has direct contact with the owners of millions of websites via their "Console" interface. This gives them a big advantage over the competition.

Part of our project, is an investigation whether Google or its competitors (like Bing), are willing to participate in sharing their contacts with website owners. Ideally, this triggers a wide cooperation.

In any case, a fully open platform will improve the quality of internet for everyone. We can get rid of `robots.txt` and other ad-hoc information suppliers.

## BUILDING AN ALTERNATIVE

In the "Open" spectrum, fact collectors can help with pieces of the puzzle. Some players are expected to be eager to contribute to the effort. For instance, Amnesty International can be expected to maintain a list of hate speech, security organisations to publish phishing sites, media organisations to publish is list of "do not visit unless license".

This project provides an orchestration of website knowledge contributed by many sources: website owners, their ISPs, detected by crawl related intelligent processes, added by copyright holders, website users, and so forth.

Collected data is published unbiased: it is up to the user to interpret it. The use of the data is fully open: also commercial companies are allowed to use it. Preferably, they contribute to the required infra-structure, development, and maintenance.

This design must be capable to work (potentially) on the full internet scale of a few hundred million websites, hence on a distributed cluster of servers. It needs to provide

- an interactive interface for website owners: to configure optimal crawling, to see what is published about their website(s);
- an interactive interface for internet provides: to configure optimal crawling defaults, to monitor the websites they host;
- an interface for bulk uploads from fact providers: contributing knowledge. Data will expire when not maintained;
- an interface for querying for website facts, for crawlers and seach engine help, to improve quality;
- an infrastructure to "deep query" fact providers, for instance to inspect crawl failures and performance;
- an interface for website visitors to flag and view website meta-data, so they can see which links are unsave to follow, see what other people reported, or add tags; and a
- website-owner authentication facility.

Development will focus on designing a save, pluggable cluster implementation. Most effort is spend in shaping an open community. The interfaces will be created as demonstration, to be extended with sub-projects in 2022.

## PRESENTATION

The presentation will convince the audience how important detailed website knowledge is to improve the required search engine quality.

The general design will be presented on a mild technical level, with some examples of applications. The algorithmic and legal challenges will be discussed.

## ACKNOWLEDGEMENTS

This project is made possible by a generous grant from the NLnet Foundation.

† mark@overmeer.net
* berov@studio-berov.eu
‡ ronlam@hnw.nu

ISBN: 978-92-9083-633-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.6045231



# FASTWARC: OPTIMIZING LARGE-SCALE WEB ARCHIVE ANALYTICS

Janek Bevendorff*, Martin Potthast†, Benno Stein*
*Bauhaus-Universität Weimar, †Leipzig University

*Abstract*

Web search and other large-scale web data analytics rely on processing archives of web pages stored in a standardized and efficient format. Since its introduction in 2008, the IIPC's Web ARCive (WARC) format[1] has become the standard format for this purpose. As a list of individually compressed records of HTTP requests and responses, it allows for constant-time random access to all kinds of web data via off-the-shelf open source parsers in many programming languages, such as WARCIO,[2] the de-facto standard for Python. When processing web archives at the terabyte or petabyte scale, however, even small inefficiencies in these tools add up quickly, resulting in hours, days, or even weeks of wasted compute time. Reviewing the basic components of WARCIO and analyzing its bottlenecks, we proceed to build *FastWARC*, a new high-performance WARC processing library for Python, written in C++ / Cython, which yields performance improvements by a factor of up to 6x.

## INTRODUCTION

The earliest open source implementations of the WARC format were provided for Java, namely Lin's *ClueWeb Tools*[3] (initially used by the research search engine ChatNoir),[4] followed by a more standards-compliant reference implementation from the IIPC.[5] Meanwhile, the IR, NLP, and machine learning communities have largely transitioned to Python, instead adopting WARCIO as a native implementation in that language. Processing large samples of the Common Crawl and web archive data from the Internet Archive, however, we observed that the library did not match our performance expectations. Even compiling it to native C code using Cython yielded only marginal improvements. Analyzing its bottlenecks, three key causes can be discerned: (1) stream decompression speed, (2) record parsing performance, and (3) lack of efficient skipping of non-response records. Our contribution is to rectify these issues with *FastWARC*, a rewrite of the entire WARC parsing pipeline from scratch.

## FASTWARC VS. WARCIO

FastWARC is a reimplementation of WARCIO in C++ with Cython, making it both fast and perfectly integrated into the Python ecosystem, yet allowing for more language bindings if required. Table 1 compiles detailed performance comparisons: On an uncompressed WARC file, it gains an overall 5x speedup over WARCIO, or 3.3x over a naively "cythonized" WARCIO. With an average processing time of 1.8 vs. 9 seconds for a single WARC file, this already saves at least 128 hours of compute time from a recent Common Crawl with 64 000 individual WARCs (62.5 TiB compressed). For better decompression speed of gzipped streams, FastWARC interfaces directly with *zlib*, achieving compute time savings of roughly 2.1 hours per TiB or 2 200 hours per PiB over WARCIO. The largest performance penalty, however, comes from the decompressor itself. While still saving about 130 hours overall on a Common Crawl, the relative speedup shrinks to only 1.5x. For this reason, we decided to add support for the more recent and much faster LZ4 algorithm and recompressed some of our WARCs. With LZ4, we can save another 215 hours on top (a speedup of 4.1x over FastWARC with GZip), or 345 hours compared to WARCIO (speedup 6.2x).

| Comp. | Parser | Records/s | Speedup |
|---|---|---|---|
| *AMD Ryzen Threadripper 2920X (NVMe SSD)* | | | |
| None | WARCIO | 13 971.5 | – |
| None | FastWARC | 64 698.0 | 4.6 |
| None | WARCIO+HTTP | 13 570.2 | – |
| None | FastWARC+HTTP | 58 354.0 | 4.3 |
| None | WARCIO+HTTP+Checksum | 7 890.9 | – |
| None | FastWARC+HTTP+Checksum | 11 528.6 | 1.5 |
| GZip | WARCIO | 5 898.8 | – |
| GZip | FastWARC | 8 899.1 | 1.5 |
| GZip | WARCIO+HTTP | 5 986.1 | – |
| GZip | FastWARC+HTTP | 8 659.0 | 1.4 |
| GZip | WARCIO+HTTP+Checksum | 4 544.7 | – |
| GZip | FastWARC+HTTP+Checksum | 5 022.6 | 1.1 |
| LZ4 | FastWARC | 36 862.8 | 6.2* |
| LZ4 | FastWARC+HTTP | 36 327.9 | 6.2* |
| LZ4 | FastWARC+HTTP+Checksum | 10 110.0 | 2.2* |
| *Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)* | | | |
| None | WARCIO | 7 865.7 | – |
| None | FastWARC | 29 307.7 | 3.7 |
| GZip | WARCIO | 3 438.4 | – |
| GZip | FastWARC | 4 583.3 | 1.3 |
| LZ4 | FastWARC | 18 337.0 | 5.3* |

Table 1: Evaluation of FastWARC and WARCIO on two systems. Runs are (1) without payload parsing, (2) with automatic HTTP header parsing, and (3) with record checksumming. *LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

## CONCLUSION

FastWARC can speed up WARC processing significantly, saving hundreds of hours of compute time on large-scale web archive analytics. By far the largest speedup, though, can be gained from using LZ4 over GZip. Considering an additional storage overhead of only about 30–40 %, recompressing GZip WARCs with LZ4 is certainly an option to be considered, especially in cases where processing speed is more important than storage efficiency.

FastWARC is released under the Apache 2.0 license and can be downloaded from Github[6] or PyPi.[7]

[1] ISO 28500:2017; https://iipc.github.io/warc-specifications/
[2] https://github.com/webrecorder/warcio
[3] https://github.com/lintool/clueweb
[4] https://chatnoir.eu/
[5] https://github.com/iipc/jwarc
[6] https://github.com/chatnoir-eu/chatnoir-resiliparse
[7] `pip install fastwarc`

# URL FRONTIER: AN OPEN SOURCE API AND IMPLEMENTATION FOR CRAWL FRONTIERS

J. Nioche†, DigitalPebble Ltd, United Kingdom

*Abstract*

Discovering content on the web is possible thanks to web crawlers, luckily there are many excellent open source solutions for this; however, most of them have their own way of storing and accessing the information about the URLs.

This presentation introduces *[URL Frontier][1]*, a recent open-source project which aims to develop a crawler/language-neutral API for the operations that web crawlers do when communicating with a web frontier e.g. get the next URLs to crawl, update the information about URLs already processed, change the crawl rate for a particular hostname, get the list of active hosts, get statistics, etc.

Such an API can be used by a variety of open source web crawlers, regardless of whether they are implemented in Java, like [StormCrawler][2] and Heritrix or in Python like Scrapy.

The URL Frontier project also provides a reference implementation of the service.

One of the objectives of URL Frontier is to involve as many actors in the web crawling community as possible and get real users to give continuous feedback on our proposals.

After an overview of the project, we will have a quick demo of URL Frontier in action.

## REFERENCES

[1] URL Frontier
`https://github.com/crawler-commons/url-frontier`

[2] StormCrawler
`http://stormcrawler.net`

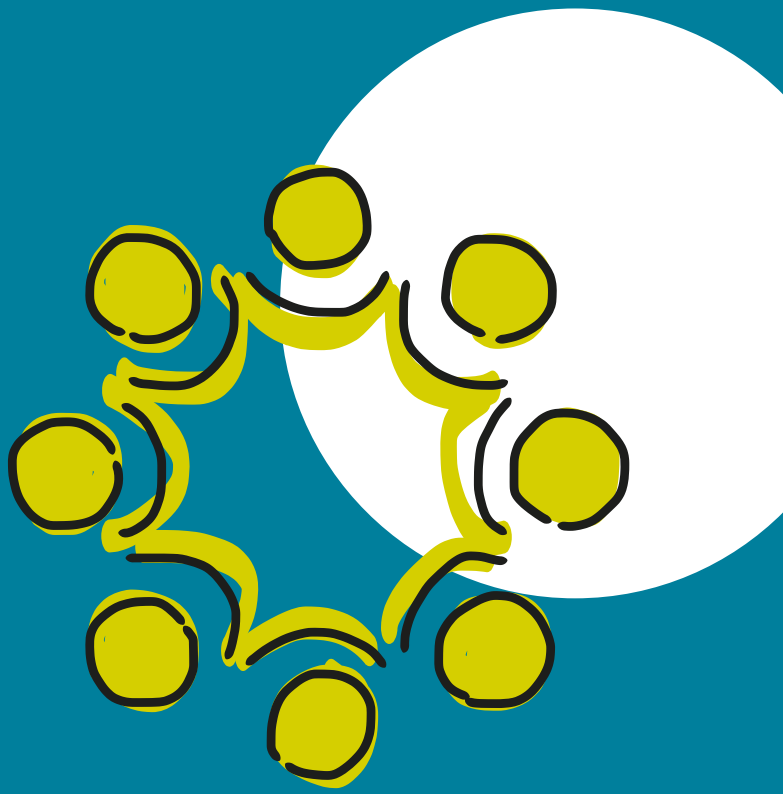
† julien@digitalpebble.com

# Appendix

## List of Autors


| | |
|---|---|
| Antunes, C. | SND-A02 |
| Behrendt, O. | FWC-A01 |
| Bensch, O. M. | SSE-P01 |
| Berov, K. | WCA-A01 |
| Berson, R. | ASA-P01 |
| Bevendorff, J. | WCA-P02, WCA-A02 |
| Bobic, A. | APP-P01 |
| Decker, A.J. | CCA-A01 |
| Engl, F. | APP-A02 |
| Ferreira, P. | SND-A02 |
| Frey, J. | APP-A01 |
| Fröbe, M. | WCA-P02 |
| Gienapp, L. | WCA-P02 |
| Granitzer, M. | APP-A03 |
| Gütl, C. | APP-P01, CCA-P01, SND-P02, WCA-P01 |
| Hagen, M. | WCA-P02 |
| Hahn, A. | SND-P01 |
| Hecking, T. | SSE-P01, SSE-A01 |
| Henrich, A. | APP-A02 |
| Hierle, A. | FWC-A01 |
| Hoyer-Klick, C. | APP-A01 |
| Jakovljevic, I. | SND-P02 |
| Jankowski, D. | SND-P01, SSE-A01 |
| Karaj, A. | ASA-P01 |
| Kern, R. | CCA-P01 |
| Kolodziejski, M. | SND-A02 |
| Lam, R. | WCA-A01 |
| Larsson, E. | ASA-P01 |
| Lourenço, P. | SND-A02 |
| Martin, L. | APP-A02 |
| Möller, J. | SND-P01, SSE-A01 |
| Mönnich, A. | SND-A02 |
| Nagel, S. | FWC-A03 |
| Niehaus, E. | SSE-P02 |
| Overmeer, M.A.C.J. | FWC-A02, WCA-A01 |
| Panero, P. | SND-A02 |
| Platz, M. | APP-P01, SSE-P02 |
| Potthast, M. | WCA-P02, WCA-A02 |
| Pujol, J. M. | ASA-P01 |

| | |
|---|---|
| Rattinger, A. | WCA-P01 |
| Russmann, S. | SND-P02 |
| Sathyanarayana, S. | ASA-P01 |
| Schröder, C. | WCA-P02 |
| Schultheiß, S. | CCA-A02 |
| Schwichtenberg, S. | SND-A01 |
| Schwinger, M. | SSE-A01 |
| Sousa, S. | CCA-P01 |
| Stein, B. | WCA-P02, WCA-A02 |
| Sünkler, S. | CCA-A02 |
| Voigt, S. | APP-A03, SSE-A01 |
| Völske, M. | WCA-P02 |
| Wagner, A. | SND-P02, SND-A02 |
| Wagner, R | WCA-P02 |



open search foundation

CERN

ISSN: NNNN-NNNN
ISBN: NNN-N-NN-NNNNNN-N