



OpenWebSearch.EU

**“Piloting a Cooperative Open Web Search Infrastructure
to Support Europe’s Digital Sovereignty”**

Report

**License Aware Web Crawling for Open
Search AI (LAW4OSAI)**

Version 1.1

Open Web Search 

The Project is funded by the EC under GA 101070014



Funded by
the European Union



OPENWEBSEARCH.EU

Table of Contents

| | | |
|-----------------|--------------------------|-----------|
| 1 | Introduction | 3 |
| 2 | Results | 4 |
| 2.1 | MS 1 Data Sets | 4 |
| 2.2 | MS 2 Workshop | 5 |
| 2.3 | MS 3 Models | 5 |
| 2.4 | MS 4 Library | 6 |
| 2.5 | MS 5 Legal White Paper | 6 |
| 3 | Dissemination | 8 |
| 3.1 | Conference Presentations | 8 |
| 3.2 | Website and Social Media | 8 |
| 3.3 | Publications | 9 |
| 4 | Conclusion | 10 |
| Appendix | 11 | |

Preliminaries

i. Project Info

| | |
|---|--|
| Project number | 101070014 |
| Project acronym | ows.eu |
| Project name | OpenWebSearch.eu – Piloting a Cooperative Open Web Search Infrastructure to Support Europe's Digital Sovereignty |
| Call | HORIZON-CL4-2021-HUMAN-01 |
| Topic | HORIZON-CL4-2021-HUMAN-01-05 |
| Type of action | HORIZON-RIA |
| Responsible unit | DG CNECT |
| Project starting date / Duration | 01/09/2022 |
| Project reporting period | 1 |
| Project Coordinator | Prof. Dr. Michael Granitzer, University of Passau |

ii. Project Partners

| Acronym | Partner |
|-----------------|---|
| LAW4OSAI | Daniel Braun, University of Twente Baltasar Cevc, fingolex Bernhard Walzl, Liquid Legal Institute |

iii. Deliverable Info

| | |
|------------------------------------|------------|
| Due Date / Delivery Date | 11/10/2024 |
| Deliverable Lead | NN |
| Deliverable type | Report |
| Dissemination level | NA |
| Document Status / Version | V1.0 |
| Work-package / Lead Partner | NN |

1 Introduction

Users will expect more than a simple list of results from the next generation of search engines. Conversational search and receiving direct answers, rather than websites that contain answers, are just two trends that are already visible today. All these next generation search applications are based on Machine Learning (ML) models that are trained on vast amounts of data.

Collecting sufficient data for the training of such models is an important endeavor, particularly in the light of a sovereign European infrastructure. The legal situation when it comes to the training of such models is in many aspects still unclear and subject of a lively debate among legal scholars and practitioners. Even once the legal debate is settled, it remains a moral question whether we want to accept that authors and content providers potentially don't have a say in how their intellectual property is going to be used.

The goal of the "License Aware Web Crawling for Open Search AI" (LAW4OSAI) project was to enable license-aware crawling of content, particularly text and images, by automatically identifying and retrieving content licenses to enable open web search filtered by licenses and, more importantly, the development of open ML models for next-generation search technology that respect licenses and copyright.

We investigated both technical and legal aspects of the topic: We annotated a data set for the detection of content licenses and developed a resource-efficient Python library that is able to not only detect licenses on webpages but also map those licenses to assets on the webpage to distinguish, e.g., between different licenses for images and text on the same page. We also investigated the legal situation and developed a framework to assess content licenses with regard to their compatibility with the training of ML models and reviewed popular open licenses (like Creative Commons) based on this framework, in order to allow both content providers and parties interested in training ML models to make informed decisions about which license to choose.

The results of our project are publicly and openly available and we continue to disseminate the results in the form of publications and talks.

2 Results

In this chapter, we will describe the results achieved by the project. The report is structured along the six milestones that we outlined in our third-party project proposal. All but one milestone (MS2 Workshop, see Section 2.2) have been achieved as planned. The sixth milestone is this report and therefore not separately listed in this result section.

2.1 MS 1 Data Sets

In order to be able to train ML models and evaluate them as well as the python library, we first needed to collect and annotate data. Based on the feedback from the consortium we decided to focus completely on standardised license instead of also collecting individual licenses that we did not plan to analyse further anyway.

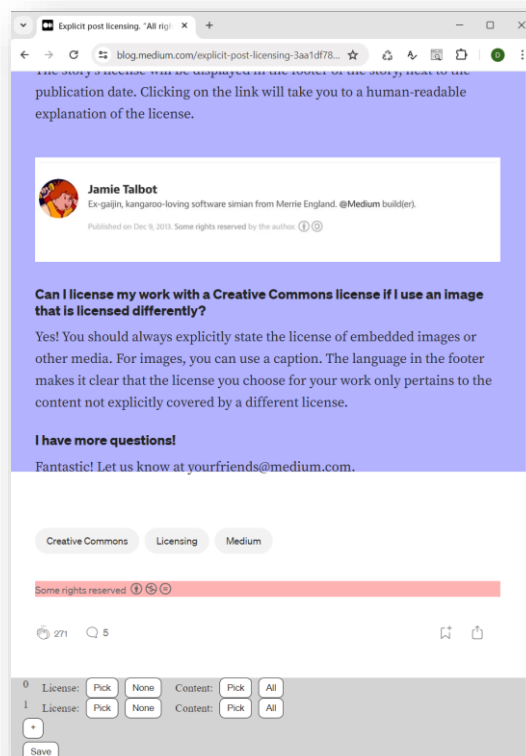


Figure 1: Chromium Browser Plugin

One of the main challenges in the detection of content licenses on website is that different parts of a website can have different licenses (e.g. images that are under a different license then the text). The data set, therefore, needed to be annotated on an HTML level in order to be able to not just annotate the (potentially multiple) licenses, but also to which parts of the page they apply. In order to enable the annotation of such a data set, we developed a browser plugin for the open-source web browser chromium. The plugin (see Figure 1) allows to annotate multiple license texts and connect them to a particular element of the page or the whole

page. The plugin adds specific HTML tags to the annotated elements that link the mentioning of the license (law4osai-license) to the element it applies to (law4osai-content) through a shared ID (law4osai-id). The annotated website can then be saved as an MHTML file which is the base for the data set. The plugin is available on GitHub¹ under the Mozilla Public License (MPL).

We used the plugin to annotate a data set of 150 pages from a wide variety of websites, including large portals like Wikipedia, Medium, and large publishers like Springer, but also small blogs and websites. Rather than focusing on size alone, we tried to create a data set that is as diverse as possible and contains as many different mixes of licenses as possible. Since the data set contains copyrighted material we cannot share it publicly, however it can be shared on request for scientific purposes.

The extraction of individual content licenses from websites, which, as mentioned before, was deemed to be of less practical relevance for the OpenWebSearch community, was still addressed as part of student work. As part of his bachelor's thesis at the University of Twente, Stefan Ilich², developed a Python tool to extract the text of individual content licenses from websites.

2.2 MS 2 Workshop

In line with the goals of the Open Web Search Community, one of our goals was to foster an exchange between legal and computer science scholars that are interested in questions surrounding content licenses and the use of licensed materials for the training of AI. We initially planned to organise an online workshop series to foster this interdisciplinary exchange and published a call for contributions in May 2024³. Unfortunately, the call did not yield a response enabling to do the workshop. However, by distributing the call in our networks alone, we were able to raise awareness for our project, the Open Web Search consortium, and the topic more broadly.

In the end, we believe we were still able to achieve the goals connected to this milestone through other dissemination activities (see Chapter 3), including presentations at the Legal and Administrative Informatics Conference and the AI Navigator Conference.

2.3 MS 3 Models

During the meetings of the OpenWebSearch community, one requirement for the extraction of content licenses that was particularly brought to our attention was computational efficiency. Originally, we were planning to mostly develop machine learning models for both the detection of mentions of licenses and connecting these licenses to the respective assets on the websites (e.g. images or texts). However, based on the importance of efficiency, to allow e.g. during web-crawling to have all websites scanned for licenses, we also tested rule-based approaches. Since it turns out the rule-based approaches were performing at least as good as ML-based approaches (often, particularly for the license detection, even better) at a fraction of the computational costs, the library we developed (see Section 2.4) relies solely on rule-based approaches.

¹ <https://github.com/LAW4OSAI/plugin-license-annotation>

² <https://essay.utwente.nl/100804/>

³ <https://www.utwente.nl/en/bms/law4osai/workshop/>

For the detection of licenses, we particularly rely on analysing outgoing links to known addresses of license texts, like <https://creativecommons.org/licenses/by/4.0/> or <https://www.gnu.org/licenses/fdl-1.3.html>. We also check for common abbreviations like “CC-BY” or “GFDL 1.2”. To avoid false positive detections, e.g. in texts that talk about Common Creative licenses but are not licensed under such a license, we use so-called shallow text features. I.e. an analysis of properties like the number of words in an HTML element or the ratio of links to text to assess whether a license is just mentioned (usually this happens in larger texts with relatively low link/text ratio) or is actually used as a license notice (usually this happens in elements with relatively little text and a high link/text ratio). Overall, the detection of the licenses is the less challenging part. The more challenging part is linking the licenses to the assets they apply to.

On some pages we found ten individual mentions of licenses, all applying to different parts of the websites, most to images. Sometimes, a single asset can also be licensed under multiple licenses at the same time, or multiple license notices (of the same license) can apply to a single asset. Adding to the complexity of the task is the variety of different structures that the combination of HTML and CSS allow for. Often, license notices will be close to the asset they apply to. However, while this is a convention, it is not a technical necessity. License notices for header images in blog, for example, can sometimes be found at the end of an article. In such cases, only an analysis of the actual texts, which would be dependent on the language and computationally very expensive, could reveal the correct mapping between license and asset. Our resource-efficient approach that is based on an analysis of the HTML structure and shallow text features, starting from the detected licenses, will in such cases not be able to establish the correct mapping. In most cases, however, our rule-based approach will be able to quickly retrieve the HTML element a license applies to.

2.4 MS 4 Library

In order to operationalize the approaches described in Section 2.3, we developed a Python library that is a wrapper for them. The library can take either a URL, an HTML file or HTML stores in a plain text variable as input. The library uses lxml⁴ to parse HTML, because benchmarks list it as the most efficient Python library for HTML parsing. Independent of the input format, the library returns a list of dictionaries which consist of a machine-readable representation of the licence (e.g. “cc-by-4.0”), a reference to the HTML element in which the license reference can be found, a reference to the HTML element to which the license applies, and what type of asset the license applies to (can be either text, image, or mixed).

The library is, like the plugin, available on GitHub under the MPL, where also examples for using the library can be found.⁵

2.5 MS 5 Legal White Paper

We have drafted the first version of a white paper presenting an overview of primarily copyright, to a small degree also other intellectual property rights and which role licenses play in asset choice for machine learning model training. For the purpose both of validating key issues as well as of evaluating standard licenses, we have created a list of relatively few metadata points that are often the key decision points on whether a license is

⁴ <https://lxml.de/>

⁵ <https://github.com/LAW4OSAI/license-extractor>

usable or not. The white paper aims at informing relevant actors in the context and supporting with the license evaluation.

The challenge we are facing in this context is that (1) jurisdictions are potentially world-wide (due to both the publication and the access potentially happening world-wide), (2) the law in the field of AI is often a moving target and (3) then, practically, that declarations of will are often ambiguous from a legal perspective.

To provide a scalable and adaptable framework for license assessment, we worked on capturing metadata in structured form so it would later be easier to read by machines (i.e. not as free text field but with fixed choices), partially enriched by notes intended for human consumption. This metadata can be used to make an informed yet fast decision about which material should be included in a training data set. The process included testing variations of fields and values to see how meaningful information could be condensed in a useful way. In choosing which data to capture, we strived, and have in our view been able to, keep the list short in order to create a form that enables quick addition of further licenses and avoids information overload for people who work with the data later.

We have reviewed exemplary licenses and interpreted them from a legal perspective to create metadata sheets for these licenses. We chose what, as to our knowledge were either heavily used or prototypical licenses for certain areas. On that basis, we have reviewed licenses from different families (e.g. Creative Commons, BSD), while for the moment leaving out licenses unlikely to be found useful for machine learning training; for that reason we have not included the GPL family—depending on interpretation, the copyleft provisions might lead to a duty to share the model and/or certain output generated by it under the same license, which would in our view make this unreasonable in many circumstances. Further provisions that we deem as likely problematic in many circumstances are attribution requirements, especially if they are applicable even to smallest parts of the work. While such provisions do not categorically rule out the use in context of AI systems, they significantly increase the risk. Such abstraction comes at an information loss (which might therefore make mistakes in the specific case, disregard differences between jurisdictions), but we deem that is a property of every scalable system in this context and scalability is a strict requirement. The evaluation, while not part of the white paper itself, supports it by pointing to types of clauses are likely to be problematic in practical application. From the reviewed licenses the requirements of attribution, sharing with the license and copyleft clauses proved to be major obstacles, even more where the risk is increased by these being made conditions to the license not mere duties.

Given significant legal uncertainty, we found that there is significant risk to relying on non-licensed material or material shared on-line with solely on exceptions (text and data mining exception in the EU, fair use in the USA), equally, while implied licenses to users is assumed, for example, for search engines whether and, if so, to which extent, is questionable. Using properly licensed material addresses this challenge and the white paper supports the corresponding evaluation. Besides being less risky, it is also advantageous from an ethical standpoint, fostering European values in training machine learning systems.

3 Dissemination

We drove different initiatives to reach out to the communities and raise interest in the LAW4OSAI sub-project as well as overall OpenWebSearch project. The activities can be grouped as follows:

1. Conference presentations
2. Website and social media
3. Publications

3.1 Conference Presentations

Based on our project scope and goals, we particularly tried to also address audiences beyond the typical academic computer science conferences. We submitted two proposals for conference presentations, both of which have been accepted:

- 1) At the 7th Fachtagung Rechts- und Verwaltungsinformatik⁶ (RVI 2024; legal and administrative informatics symposium), which was co-located with the annual meeting of the German Informatics (GI) society, Informatik Festival 2024. The presentation on 26 September 2024, which included an introduction of the OpenWebSearch project and a presentation of the results of our sub-project, was held in German and inspired a lively discussion amongst the participants, which were a blend of people from technical and from legal background.
- 2) We will present at the KI Navigator conference⁷, at its first day on 20th of November 2024. This presentation will be held in English. KI Navigator is a conference hosted by Heise Medien, one of the big IT/tech publishers in Germany, de'ge'pol, the German Society for Policy Advice, and DOAG, the German Oracle User Association.

3.2 Website and Social Media

In order to have a central place to share all our project results and update, we created a project website at the University of Twente website which can be reached under <https://www.utwente.nl/en/bms/law4osai/>. The website contains a general description of the project, links to the OpenWebSearch community, a list of the involved team members, as well as a section with the news from the project and, most importantly, the links to all resources that have been created as part of the project. Furthermore, the Liquid Legal Institute has created a distinct web page on the topic matter of this project to enable sharing of relevant information, including links to valuable resources, with their community, particularly beyond also beyond the (temporal) scope of this project.

⁶ <https://fachtagung-rvi.de/>

⁷ <https://www.kinavigator.eu/en/home/>

We also used our social media reach to spread awareness about the project, particularly on LinkedIn, where the Liquid Legal Institute reaches more than 6,000 followers, and the three authors of this report have a combined reach of more than 10,000 followers.

3.3 Publications

We currently work on a number of publications based on the results of the project. Most importantly of course the white paper outlined in Section 2.5. Additionally, we have been invited to contribute an article about the project for the January issue of the legal tech magazine LTZ⁸, mostly focusing on the legal aspects of the project. In addition, we currently work on a publication focusing on the technical aspects of the project, which we aim to submit until the end of the year.

⁸ Legal Tech – Zeitschrift für die digitale Anwendung, <https://www.nomos.de/zeitschriften/ltz/>

4 Conclusion

In this report, we summarised the results and findings of the License Aware Web Crawling for Open Search AI (LAW4OSAI) project. The project results and outcome are in line with the original planning. Although the workshop could not be conducted as planned, other dissemination activities allowed us to still reach a wider audience and foster an interdisciplinary exchange between legal scholars and computer scientists. A summary of all outputs can be found in Table 1 in the Appendix. Additional publications that make the results of the project available to a larger audience are currently still in progress and will follow in the next months.

The connection with the OpenWebSearch community and the exchange with the main project as well as the other subprojects provide valuable input for the technical execution of our own project, as we were encouraged to particularly investigate resource efficient solutions that are scalable to vast amounts of pages. Thereby, our results can hopefully contribute to a European web search ecosystem that is not just open with regard to its tech stack and index, but also encourages the usage of open licenses for content and allows for the safe and lawful training of AI models for the next generation of search AI.

We would like to thank the OpenWebSearch.eu consortium for the opportunity to work on this relevant and timely topic.



Funded by
the European Union

The project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101070014 OpenWebSearch.EU project within its Cascading Funding.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Appendix

| Data Set | Data | MS 1 | On request |
|--|-------------|---------------|--|
| Browser Plugin | Software | MS 1 | https://github.com/LAW4OSAI/plugin-license-annotation |
| License Extraction Model / Aglorithm | Software | MS 3 | https://github.com/LAW4OSAI/license-extractor |
| Python Library for License Extraction | Software | MS 4 | https://github.com/LAW4OSAI/license-extractor |
| Legal White Paper | Publication | MS 5 | To be published |
| Legal Publication | Publication | Dissemination | To be published |
| Technical Publication | Publication | Dissemination | To be published |
| Website | Publication | Dissemination | https://www.utwente.nl/en/bms/law4osai/ , https://liquid-legal-institute.com/workinggroups/license-aware-web-crawling-for-open-search-ai/ |

Table 1: Summary of project outputs