Open WebSearch
.eu

OpenWebSearch.EU
"Piloting a Cooperative Open Web Search Infrastructure to Support Europe's Digital Sovereignty"

Community Programme:
1st call for proposals
– Information for applicants –

Open Web Search

The Project is funded by the EU under GA 101070014

Funded by
the European Union

NGI OPENWEBSEARCH.EU

Table of Contents

1   Introduction

1.1   OpenWebSearch.eu Project

Web search has become an essential technology and infrastructure, driving not only future innovations but forming a backbone for our digital economy. However, currently few non-European gatekeepers control Web search, which creates a biased, one-sided information access centred around economic success rather than the needs of individuals or European values and jurisdiction. This one-sided ecosystem puts pressure on the many small contributors to the Web in science, economy, art, culture, media and society to optimize their content to be taken-up by the few gatekeepers and thereby greatly feeding the gatekeepers' monopolies. A vicious cycle, leading to lock-in effects and to a closed search engine market.

To open the search engine markets and to provide a true choice for users in selecting the search engine based on their preferences, OpenWebSearch.eu proposes to develop and pilot the core of a European Open Web Index (OWI) and the foundation of an open and extensible ***European Open Web Search and Analysis Infrastructure (OWSAI).*** Our approach is based on four objectives, namely (1) to develop a core suite of search, discovery and analytics services to create, maintain and utilize the OWI; (2) to develop relevant search engine verticals and new search paradigms demonstrating the impact of the OWI; (3) to establish a network of European HPC infrastructure, research and business organisations to pilot the OWSAI based on Europe's values, principles, legislation, ethics and standards and (4) to stimulate an Ecosystem around the OWI.

The envisioned infrastructure will not only contribute to Europe's sovereignty for navigating and searching the Web, it will also allow Europe's researchers, innovators and business to systematically tap into the Web as business and innovation resource, without paying huge upfront costs.

More information on the conceptual approach can be found in the Project Background (section 9) below and on the homepage https://openwebsearch.eu/.

1.2   OpenWebSearch.eu Community Programme

The OpenWebSearch.eu project consortium is eager to identify new project teams to be funded under the OpenWebSearch.eu Community Programme, on-board and integrate them into ongoing and future activities for sustainable Research and Development on Open Web Search.

Third-party activities are driving forces in the OpenWebSearch.eu Community Programme, with funding provided by the OpenWebSearch.eu project (funded by the EC under the GA 101070014). In spring 2023 the OpenWebSearch.eu Community Programme is coming to life. The Community Programme is comprised of activities in support of third-party project initiation, execution and successful completion.

Consequently, the OpenWebSearch.eu project will call for proposals to help building an open and independent Web-index. ***The first call will open on 1st of March 2023*** inviting researchers and innovators to submit their applications to contribute to the mission of OpenWebSearch.eu project. ***Applications will be accepted until 28th of April 2023, 17:00.***

The first call consists of two tracks:

Track 1: Conceptual studies on legal or economic aspects of Open Search,

Track 2: Technical approaches to legally compliant data acquisition considering societal constraints.

Candidate third-party projects should address topics closely related to the project. They should aim at widening and enriching the existing R&D activities as well as suggesting new ones, complementary to the project goals and aims.

Successful applications can request funding between 25,000 and 120,000 EUR in this first call, for a funding period of up to 12 months (see the track specific details for funding and budget break down).

In particular, we are targeting smaller companies (e.g., SMEs, start-ups), individual innovators, individual researchers or research teams (e.g., doctoral or post-doctoral researchers) from renowned universities. Eligible applicants are individuals residing in EU Member States or Horizon Europe Associated Countries or organisations registered in EU Member States or Horizon Europe Associated Countries.

1.3   Overall Procedure

The Community Programme starts with the **openly published calls for proposals,** which are specified with a clearly defined application procedure. The applicants will be asked to submit a short and concise proposal addressing the specified call topics. The call topics are developed within the consortium, partially with the help of external experts. Applications will be reviewed by the members of the consortium as well as by the invited external experts chosen prior to the calls; the procedure is comparable to a review procedure by a programme committee of a scientific conference.

The objective is to select projects that have the greatest potential to maximize the impact and success of the OpenWebSearch.eu project. These projects should possess a well-defined research or development component aimed at promoting Open Search as a whole.

2   Call Topics

OpenWebSearch.eu aims at building and piloting a legally compliant Open Web Index as a European infrastructure. The first OpenWebSearch.eu call asks for contributions in legally compliant data gathering and in identifying legal or economic aspects that enable or block the development and maintenance of an Open Web Index. Successful applicants will conceptually support the development of the platform as well as contribute with concept studies on the legally compliant gathering of web data and metadata.

We call for contributions in form of research studies on legal or economic aspects of Open Search (Track 1) as well as on concepts for legally compliant data acquisition, processing, considering also societal constraints (Track 2). Each track poses a set of concrete questions which should be addressed by applications.

2.1   Track 1: Conceptual Studies on Legal or Economic Aspects of Open Search

From a high-level perspective, OpenWebSearch.eu aims to gather Web data to refine it into an openly available index of the Web, so-called Open Web Index (OWI). The index will serve as a basis for search capabilities, to find websites, but also contains large parts of web-based information itself. Next to classical search, applications that can be built on such an Open Web Index include the provision of cleaned and enriched data collections for third party use (especially for search engine providers), as well as the development of AI based knowledge representation models (e.g., Neural Language Models), mobile and geospatial web services, scientific use, price-comparison services, news/trend monitoring, web-statistics, business-analytics and many more.

Building such OWI does not only include technical challenges, but also legal[1] and societal ones, especially when considering recent EU legislations like the Digital Service Act or the Digital Market Act. Furthermore, challenges for new business models or significant changes in the search engine market arise. Consequently, applicants should prove either respective legal or economic expertise as well as necessary technical expertise. Technical expertise might also be provided by members of the OpenWebSearch.eu consortium. In the context of legal, business and societal aspects of Open Search we are asking for either of the following types of contributions:

1. Legal Studies:

OpenWebSearch.eu needs to analyse and deeply understand legal constraints and requirements for building and operating an OWI. ***In particular, we are looking for research studies which analyse the legal constraints and requirements of crawling, storing, aggregating, enriching and sharing web data operationally in an OWI.*** When building the OWI, OpenWebSearch.eu will gather web-data at a large scale. The collected

---

[1] See e.g., https://link.springer.com/article/10.1365/s43439-021-00017-8

web data will be processed (e.g., cleaned, enriched) and linked with other, publicly available and internal data sources. The results will be made available to the general public as well as to industry and SMEs. Users of the OWI can potentially build commercial business on top of the provided data. *Consequently, constraints and requirements need to be better understood,* particularly regarding intellectual property rights and other related legal constrains, data protection law, data economy law (i.e., Digital Services Act, Digital Markets Act etc.) as well as constraints and requirements put forward by criminal law (e.g., through storing illegal content) and other potentially relevant areas of law.

In addition to this, accessing the OWI requires corresponding usage and license terms, with a particular emphasis on minimising the liability for operators of an OWI. OpenWebSearch.eu is thus also looking for analysis and studies on existing open source- and open data licences in regard to applying them for operating and managing an OWI. Such an analysis should include the most critical risk factors and their mitigation from the viewpoint of the OWI operator(s).

Another point of interest is the "right to de-referencing" as recognised by the General Data Protection Regulation. This right may also affect an OWI operator. However, due to the multi-stakeholder nature and the diversity of search engines to be expected to operate, based on an OWI, multiple parties will be involved. Furthermore, authenticity of claims of legal/natural persons has to be ensured, also to prevent potential abuse.

Where applicable, the analysis should also contain concrete recommendations for changes to existing or emerging laws and regulatory frame works to allow a legally compliant operation of an OWI in Europe.

Applicants are expected to support the OpenWebSearch.eu project and its goals through conceptual studies on the following topics:

a. compilation and analysis of the laws and norms that are relevant to building and maintaining an OWI, including an analysis of specific legal issues in terms of intellectual property rights and related rights, data protection law, data economy law and/or criminal law relevant for the whole or parts of the project,
b. legal assessment of technical and non-technical prevention mechanism (e.g., automated deletion of prohibited content) as well as an assessment on the degree of accurateness of automatic prevention mechanisms combined with potential manual ones,
c. legal assessment of the implications of the right to de-referencing for an OWI,
d. analysis of existing open source and open data licenses in regard to the suitability for usage in an OWI,
e. a combination of some or all of the above.

2. Economic Studies:

Setting up and maintaining an OWI as public European infrastructure will require substantial upfront investments and maintenance budgets by European member states and the European Commission. In order to justify these investments and to possibly recover some of the costs, the economic aspects and the cost-benefit dimensions of an OWI and open web search ecosystem shall be studied and better understood. It is the assumption that an OWI will be maintained as a public infrastructure and that it may recover some of its operational costs e.g., by charging service fees for commercial users,

heavily using the infrastructure, e.g., beyond a basic number of free API-requests per day (freemium-model). We ***call for studies analysing and estimating the costs associated with setting up, operating and maintaining a distributed open web index infrastructure across Europe and analysing and estimating the market potential and economic impact of such an infrastructure***.

Estimating the costs associated with a public OWI shall be based on the assumption that it will be set-up, hosted and maintained by existing public and private computing centres across Europe, using spare capacity or adding a few extra percent of additional computing capacity to the existing one. So relevant additional costs should be taken into account (staff, computing and storage costs, bandwidth, energy consumption, etc.), however, no building-up of new computing centres etc. shall be assumed. An initial technical scenario for the required storage volumes, bandwidth and computing capacities will be provided to successful applicants by the OWS.eu technical team and should then be further financially detailed, elaborated, documented and underpinned in this study.

Estimating the economic potential and impact of an open web index shall provide an in-depth cost-benefit analysis with respect to several dimensions: opening up the search-engine market and diversifying general web search due to the public availability of a comprehensive, up-to-date web index that can be used by start-ups, industry and business to develop and provide web-based services. The services addressed in the study range from general search, localized and specialized search, training of AI models, large scale language models, e-commerce, retail/price comparison, mobility and geolocation services, transports, communication, news and media, etc. Furthermore, the studies shall address, estimate and document the non-commercial, societal or indirectly commercial impact and benefits of such a public OWI infrastructure. Benefits to be addressed may range from improved privacy, respecting of European languages, culture and diversity in digital services, European digital sovereignty, fake-news detection, cyber safety and security, crisis monitoring, technological foresight, protection of intellectual property rights, to a wide range of possible scientific uses of such a public open web index and analytics infrastructure (non-exhaustive list).

Generally, it is expected that studies will approach the cost-benefit analysis on a macro-economic level. It is a plus if selected use-cases may also be the studies on a concrete business-economics level. Project proposals studying these economic dimensions should indicate how they plan to analyse and quantify these economic and societal benefits of an OWI and how they plan to establish a meaningful cost-benefit analysis for an OWI as public infrastructure in Europe.

Project proposals need to make clear how their conceptual considerations contribute to opening web search in general and to the planned OpenWebSearch.eu platform in particular. The applicants need to make clear, ***which components of an Open Web Index will be addressed by and derive benefit from their studies.*** Besides competences in either the legal or the economic domain, applicants have to demonstrate how they integrate the necessary technical skills, either within their team or through cooperation with the project consortium. All results of the studies must be published under an Open Access license, undergo scientific quality control (e.g., peer reviewing) and must be accessible following the FAIR principles.

Track 1 has an indicative total budget of 160.000 EUR. We expect projects in the range between 25.000 EUR and 75.000 EUR for a duration of 6-12 Months. Durations and funding might be extended for exceptionally successful projects (subject to evaluation and budget).

2.2 Track 2: Technical Approaches to Legally Compliant Data Acquisition Considering Societal Constraints

Web crawling describes the process of navigating the graph structure of the Web for discovering and fetching Web-data. Web crawling is the predominant method for web search engines to gather content for their index. From a technical perspective, crawling has become a standard technology. However, from a legal perspective, crawling remains a grey zone and needs to reflect intellectual property rights, identification of illegal content as well as data protection. While webmaster and content owners have some control over the crawling process via (de facto) standards like robots.txt and sitemaps, further control mechanisms remain proprietary, like for example Google's Search Console. ***Within track 2 we are looking for concepts and approaches for opening the proprietary components of the aforementioned issues***. In particular we are looking for

1. Vocabularies, metadata schemata or ontologies for expressing/defining legal constraints on web page usage and potential parsers / validators for such languages;
2. Services, frameworks or algorithms collecting and organising legal, technical and compliance related meta-information on websites;
3. Open datasets and machine learning models for analysing Web pages according to legal properties (e.g., filtering of illegal content, information extraction for legal purposes, genre classification etc.);
4. Services and tools for Webmasters and content owners to express and steer expected crawling behaviour and/or to specify usage rights as well as legal compliances (e.g., with GDPR) in an open and cooperative manner;
5. Legally compliant data acquisition, processing and hosting strategies that go beyond Web crawling.

Solutions have to be provided as Open Source (i.e., in an OSI[2] compliant license), collected data needs to be provided as Open Data and documentation or experimental results have to be published with Open Access.

Project proposals need to make clear how their solutions contribute to opening Web Search, in general, and the planned OpenWebSearch.eu platform, in particular (further details are provided under http://www.openwebsearch.eu). Note that proposals must make valid contributions to the OpenWebSearch.eu project and must be usable in the project under open source / open data licensing terms.

Track 2 has an indicative total budget of 160.000 EUR. We expect projects in the range between 60.000 EUR and 120.000 EUR for a duration of 9-12 months. Duration and funding might be extended for exceptionally successful projects (subject to evaluation and budget availability).

2.3 Activities to be Funded

Track 1 asks for contributions in form of research studies.

---

[2] https://opensource.org/definition/

Studies on legal or economic aspects of Open Search. Research studies must adhere to the scientific standards in the particular field and resulting publications must undergo peer review. Note that consultancy reports or white papers are not considered as research studies.

Studies shall review state of the art and shall reflect specifically the European market, jurisdiction, value and cultural aspects of open search. Study teams shall address at least one of the aforementioned topics and shall comprise a comprehensive literature review, review of respective legislation at European and member-state level and an analysis part as well as a comprehensive results and conclusion part within which specific findings and recommendations for the given topic on the context of open search are presented. Where relevant, empirical studies (e.g., surveys) can be suggested for answering the proposed research questions. Results must be made available as Open Access and studies have to undergo a peer reviewing process.

Track 2 asks for technical, algorithmic and/or data-related contributions, focused on **developing potential** components of the pilot open search infrastructure, relevant data sets, data acquisition techniques and/or machine learning models.

Successful applications will technically define and develop new search and discovery applications, content analysis methods, search paradigms, data products or extend the platform in relevant ways. Here we address innovators and business that (i) either build search verticals on top of our pilot infrastructure thus demonstrating its applicability, or (ii) extend the pilot infrastructure to relevant areas, like new content analysis methods, or (iii) develop interesting data products on top of the Open Web Index. Results must be made available as Open Source and/or Open Data, with documentations or experimental work published in as Open Access.

3 Target Audience

The calls target especially smaller companies (i.e., SMEs, start-ups), individual innovators, individual researchers or research teams (e.g., doctoral or post-doctoral researchers). Especially for the technically oriented calls (Track 2), the consortium will provide necessary information and training material and support as well as the necessary documentation of the on-boarding and governance process.

The eligible applicants for this opportunity are either:

- Individuals who are citizens or residents of any EU Member State or any of the countries associated with Horizon Europe; or
- Organizations that are registered in any EU Member State or any of the countries associated with Horizon Europe.

Please note that the list of associated countries may change over time, and it is recommended to check the latest list of eligible countries before applying.

The third-party calls particularly focus on the following categories of applicants:

- Academic researchers and research groups in universities or research centres or R&D focused organisations;
- Renowned experts, individuals and scholars or associations;
- High-tech start-ups, SMEs, or industry with a focus on Web technology or software development
- Outstanding individual open-source innovators / researchers and experienced individual developers / researchers;
- Other multidisciplinary actors.

Applications can also involve teams of different organisations or teams of natural persons. In case of team applications (i.e., multiple natural persons without an organisational entities or multiple organisational entities), one team member must take the role of main contact point and legally responsible party.

Following the Council Implementing Decision (EU) 2022/2506, as of 16th December 2022, no legal commitments can be signed with Hungarian public interest trusts established under Hungarian Act IX of 2021 or any entity they maintain. Affected entities may continue to apply to calls for proposals. However, in case the Council measures are not lifted, such entities are not eligible to participate in any funded role (beneficiaries, affiliated entities, subcontractors, recipients of financial support to third parties). In this case, co-applicants will be invited to remove or replace that entity and/or to change its status into associated partner. Tasks and budget may be redistributed accordingly.

Please note that in the Funding &Tender portal there is a FAQs focusing on the specific implications for Horizon Europe actions which will be regularly updated.

4 Submission and Evaluation

4.1 Important Dates

- Opening date: 1st March 2023
- Closing date: 28th of April 2023, 17:00 CET
- Notification date: 30th June 2023
- Start of projects: 1st August 2023

4.2 Submission Procedure

The applicants have to submit their appropriately formatted proposal by email to the call management (call1@openwebsearch.eu for the first call) by the given deadline. The proposal template is available for the applicants. Applicant(s) can submit at most one application per call. English is the main language for communication with the OpenWebSearch.eu consortium, and all submitted documents must be written in English.

The submission will be acknowledged by the call management, and, only after the confirmation, the proposal can be considered as being submitted. It is advised not to wait till the last moment with the submission. Note that multiple submissions of the same project are not accepted. If the submission is not confirmed within max. 2 days, and provided the call deadline has not yet passed, you may contact the Grantor at call1@openwebsearch.eu to request the information and ask for re-submission.

4.3 Evaluation

Topical fit and contribution to the project has been set as the key criteria for the selected third-party projects. Thus, the call addresses specific questions in the project raised by the consortium members. The evaluation will be conducted by external experts in cooperation with the members of the project consortium.

Proposals will be evaluated according to the following categories and sub-categories:

1. **Relevance** to at least one of the call tracks and complementarity to the project, i.e.,
   - how well the proposal addresses the specific topics of the call,
   - how the proposed work complements the pilot infrastructure or repeats what has been done already;

2. Contribution to the goals of *at least one of the five main components* of the OpenWebSearch.eu project: (i) Crawling and Crawl Management, (ii) Pre-processing and Enrichment, (iii) Indexing, Search Paradigms and Search Architectures, (iv) Search Verticals, Data Products and Applications, (v) Data Storage, Provenance, and Computing Infrastructure (see also section 9 for more details). The following aspects will be evaluated in particular

- how the proposal contributes to one or more of the *four project objectives*?

- how the work in the project will benefit from the proposed work?

- to which project work-package is the proposal contributing and to what degree?

3. Importance of the target community or ecosystem, i.e.

- the size of the potential impact on the target community,

- the likelihood for successfully building up an ecosystem / community,

- the degree of importance of the target community to successfully introduce Open Web Search as alternative search system,

- access to relevant stakeholders in the OpenSearchWeb.eu community,

- plans and measures for creating an ecosystem.

4. Expected impact for Open Search in general (also beyond the pilot infrastructure);

5. Feasibility, i.e.,

- whether the planned outcome seems to be realistic,

- whether the requested resources are realistic for achieving the planned outcome

- whether the envisioned target community /ecosystem could be addressed

- sufficient past expertise of the applicants for carrying out the work

In total 100 points will be distributed among the categories 1-5. Besides the above listed criteria, formal eligibility criteria in accordance with the grant agreement and European Regulations must be fulfilled, particularly on the target group and official funding criteria.

5   Awarding Procedure and Timeline

The assessment will be conducted, after the primary screening for completeness of the proposal. After the screening, candidate projects will be evaluated by the review committee. The review board will be comprised with the selected principal investigators of the project and invited external experts. Proposal review will be conducted according to the pre-defined criteria. The evaluation will result in respective points assigned to the sections of the proposal. As a result, a ranked list of the proposals will be made and the call funding will be distributed to the top candidates of the list. ***The awarded project summaries will be published online.***

The consortium aims for prompt evaluation procedure of approximately one- or two-month duration.

6 Grant Implementation

The applicants will be informed by the Business Management Unit of the project on positive or negative outcome of the evaluation and selection process (email communication).

The successful applicants will be informed about awarding by means of an electronic funding letter. After signing a legally binding contract (a sub-grant agreement) between BADW-LRZ (the Grantor) and the (main) applicant (the Grantee), and communicating applicant's bank account information, the pre-financing will be transferred to the Grantee by BADW-LRZ on behalf of the consortium.

The Grantees will receive a pre-financing of 50% at the start of the project. The subsequent transfer will 25% of the funding will follow after the successful mid-term review, and the final payment of the 25%, after the approval of the final report.

The grantee is expected to implement the project based on effort, meaning that the promised project results are expected to be delivered within the specified timeframe. If a deliverable cannot be achieved, the grantee must inform the parent consortium in a timely manner and propose mitigation strategies. The proposed strategies will then be negotiated with the OpenWebSearch.eu beneficiaries.

The consortium members selected for the project will review the project deliverables, and the grantee will be notified of any rejection within the time specified in the grant agreement. If the Grantor does not object to the project deliverables within the specified timeframe, their silence may be considered as an implied approval of those deliverables.

An interim project report will be required for the mid-term assessment of the project progress and funding transfer. At the end of the grant implementation, the final results must be presented to the consortium for review. The final payment to the grantee will depend on the positive outcome of the project deliverable review. The grantee will receive the final payment only after acceptance of the final report.

Documents required from the Grantee:

- Completed and signed sub-grant agreement;

- Project report, including the interim and final reports;

- Financial report.

6.1 Integration and Capacity Building

The Grantor representing the consortium plans some capacity building measures directed towards seamless integration of the third-parties with the parent OpenWebSearch.eu project.

During implementation, the Grantees will be invited to the consortium meetings and will get access to selected results of the project. If suitable, the consortium will integrate the grantee in the relevant work package and appoint a contact person from the consortium who supports them in promoting the applicant's contributions. During the grant implementation, the grant owners should present their current state of the project at the consortium meeting either in person or virtually or, if the time of the meeting is inappropriate, at a negotiated appointment.

Beyond the financial support, the awarded projects can apply for free services such as accessibility, security or license scans at NLnet[3] or take on business accelerating services from the NGI community[4]. The third-party projects as integral part of OpenWebSearch.eu project's Community Programme can greatly benefit from the NGI support measures and promotional activities such as e.g., NGI catalogue, NGI Communication Task Force and NGI Impact, Exploitation and Sustainability Working Group. The relevant integrative actions within NGI community are already in progress.

---

[3] http://nlnet.nl/ngi0/review/

[4] https://www.ngi.eu/acceleration-services/

7 Legal Part: Declarations

Note that in order to evaluate your proposal within the third-party call pipeline, the processing of personal data will take place. Personal data submitted within the proposal will be processed in accordance with applicable EU and national data protection law. Such data will be processed by the OpenWebSearch.eu Consortium partners and the external reviewers for the purposes of evaluating the proposed project and implementing, managing and monitoring the evaluation process. After the evaluation process, personal information on rejected proposal will be deleted, with the exception of reporting obligations towards the EU. All personal information from the applicants, particularly the information used in reporting, will be treated as sensitive information according to Article 13.1 of the General Model Grant Agreement[5]. Personal information for accepted proposals will be used by the consortium to prepare the sub-grant agreement, which regulates all further use of personal data in the project implementation. In case sub-grant agreement fails, the proposal will be treated as rejected.

Please check-mark the respective declarations to confirm that you have read and understood this information clause concerning processing of the personal data; that you have the legal basis for processing personal data of your team members mentioned in the application; that you will pass the information clause provided under legal part of the proposal to all team members concerned.

Respectively, in the legal context applicants will be asked to acknowledge the such declarations as e.g.:

- You have read and understood the relevant information above. You agree to participate in the OpenWebSearch.eu project;

- You understand that you can withdraw at any time without giving reasons and that you will not be penalised for withdrawing nor will you be questioned on why you have withdrawn;

- You acknowledge that the reviewers and the European Commission and its bodies and agencies may have access to the data collected under the open call;

- The data provided in the application form are true and up-to-date;

- The entity you represent meets the eligibility conditions described in call conditions;

- There is no conflict of interest between the organisation you represent and any of the OpenWebSearch.eu consortium partners or the reviewers;

- You did not make false declarations in supplying the information required, as a condition of participation in the Open Calls and do not fail to supply this information;

- Published results will required to acknowledge the funding from OpenWebSearch.eu (funded by the EC under the GA 101070014) and undergo a peer-review process.

---

[5] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/agr-contr/general-mga_horizon-euratom_en.pdf

8   Contact

If you have further questions on the application procedure and third-party call related activities of OpenWebSearch.eu project, feel free to contact us.

Administrative contact for the 1ˢᵗ call call1@openwebsearch.eu

Informal inquiries about the community program *community@openwebsearch.eu*

9 Project Background

OpenWebSearch.eu aims to develop and pilot the core for a European Open Web Index (OWI) and the foundation of an open and extensible European Open Web Search and Analysis Infrastructure (OWSAI) by bringing together strong European players, who jointly define, develop and pilot an open technological backbone for cooperative web search. The proposed pilot infrastructure will demonstrate, how search applications and web-based AI data products can be realized through cooperative crawling, analysis, storing and indexing of web content.

Four objectives are pursued by the project:

- **Objective 1** forms the technical core by creating a suite of search, discovery and data analytics services to create, maintain and utilize the OWI. The envisioned infrastructure will demonstrate the feasibility of developing search engine verticals in the long tail and corresponding Web-based data products. While adhering a full open source and open data strategy.

- **Objective 2** aims to develop sensible search engine verticals as demonstrators and for bootstrapping a new search engine and web-data product market. Examples include search verticals for science search and mobile search as well new search technologies, particularly new search paradigms and knowledge representation models (symbolic or sub-symbolic).

- **Objective 3** aims to establish a network of European HPC-infrastructure, research and business organisations for jointly piloting the developed infrastructure while adhering to Europe's values, principles, legislation, ethics and standards. Europe has strong public computing infrastructure and research organisations in its member countries.

- **Objective 4** aims to stimulate an Ecosystem around the Open Web Index (OWI) which consists of innovators, researchers, computing centres, policy and decision makers and developers.

The development is structured along data processing chain for search and discovery systems while considering different stakeholders and services, as shown in the figure 1 below.

CCM - Crawling and Crawl Management (1 + 2) aims to conduct own web crawls and to coordinate independent crawling efforts and to gather information from website-masters to obtain additional information (e.g., provenance data, legal data, license data, website classification). An **Open Website Index** of all accessible websites of the Web plus additional metadata describing the website (e.g., categorization, topics, legal status, license information and compliances etc.) will be developed to coordinate the independent crawlers and provide crawling results in

PPE - Pre-processing and Enrichment (3) will develop **extensible pre-processing and enrichment services** with some base services for pre-processing and cleaning, but also more sophisticated semantic enrichment techniques, like removal of personal data and assessing the information quality automatically. Modules from external contributors will be welcomed.

ISPSA - Indexing, Search Paradigms and Search Architectures (4) builds indexing services for **creating full-text and metadata search indices** for the high-quality web data collections. Conceptually, every index could be maintained by a different organisation / owner in order to have a clear, legal responsible person. With this approach we aim to enable small organisations and even individuals to establish their own brand on searchable content, such that web-users can select – or even define – a search engine completely tailored to their preferences.

SVDPA - Search Verticals, Data Products and Applications (4): A series of vertical search applications (verticals) will be developed that **demonstrate and prove the capabilities and flexibility**

of the OWSAI. These verticals are built upon the OWSAI and exploit its various features, in order to facilitate different use cases and serve specific requirements of different end user groups. Thus, verticals not only include specialised search engines and a core managing those search engines, but also data products, smart objects, and particularly AI models. Specific attention will be paid to enhancing user-centred features, such as **advanced trust and transparency models** that open up technical information on the documents, gathering, indexing, and ranking methods.
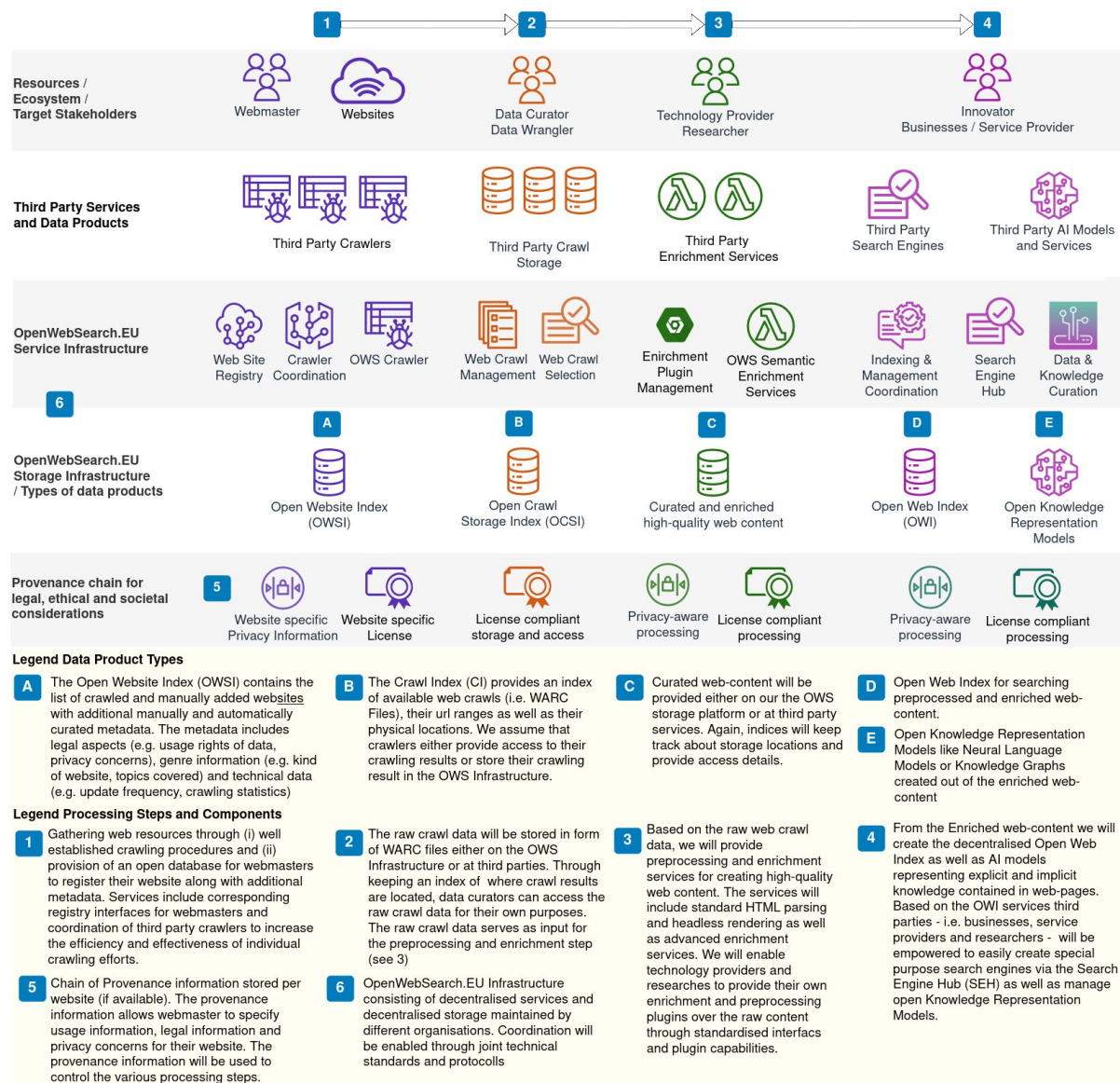

Fig. 1. Data processing chain for search and discovery systems involving different stakeholders and services.

Legend Data Product Types

A The Open Website Index (OWSI) contains the list of crawled and manually added websites with additional manually and automatically curated metadata. The metadata includes legal aspects (e.g. usage rights of data, privacy concerns), genre information (e.g. kind of website, topics covered) and technical data (e.g. update frequency, crawling statistics)

B The Crawl Index (CI) provides an index of available web crawls (i.e. WARC Files), their url ranges as well as their physical locations. We assume that crawlers either provide access to their crawling results or store their crawling result in the OWS Infrastructure.

C Curated web-content will be provided either on our the OWS storage platform or at third party services. Again, indices will keep track about storage locations and provide access details.

D Open Web Index for searching preprocessed and enriched web-content.

E Open Knowledge Representation Models like Neural Language Models or Knowledge Graphs created out of the enriched web-content

Legend Processing Steps and Components

1 Gathering web resources through (i) well established crawling procedures and (ii) provision of an open database for webmasters to register their website along with additional metadata. Services include corresponding registry interfaces for webmasters and coordination of third party crawlers to increase the efficiency and effectiveness of individual crawling efforts.

5 Chain of Provenance information stored per website (if available). The provenance information allows webmaster to specify usage information, legal information and privacy concerns for their website. The provenance information will be used to control the various processing steps.

2 The raw crawl data will be stored in form of WARC files either on the OWS Infrastructure or at third parties. Through keeping an index of where crawl results are located, data curators can access the raw crawl data for their own purposes. The raw crawl data serves as input for the preprocessing and enrichment step (see 3)

6 OpenWebSearch.EU Infrastructure consisting of decentralised services and decentralised storage maintained by different organisations. Coordination will be enabled through joint technical standards and protocolls

3 Based on the raw web crawl data, we will provide preprocessing and enrichment services for creating high-quality web content. The services will include standard HTML parsing and headless rendering as well as advanced enrichment services. We will enable technology providers and researches to provide their own enrichment and preprocessing plugins over the raw content through standardised interfaces and plugin capabilities.

4 From the Enriched web-content we will create the decentralised Open Web Index as well as AI models representing explicit and implicit knowledge contained in web-pages. Based on the OWI services third parties - i.e. businesses, service providers and researchers - will be empowered to easily create special purpose search engines via the Search Engine Hub (SEH) as well as manage open Knowledge Representation Models.

DSPCI - Data Storage, Provenance, and Computing Infrastructure (5 + 6): Provenance-enabled, distributed, inter-organisational storage facilities will be essential for all the above services. Augmented by computing services for the project, these facilities will **enable FAIR[6] data and**

---

[6] M.D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3,

metadata handling (and sharing), as well as high-performance data analysis. Provenance-tracking becomes especially important to understand properties like trust, privacy and security. For example, a website that has been registered by a webmaster as being marked as privacy preserving can be indexed by a vertical search engine that focuses on privacy preserving web content or for building privacy-preserving Artificial Intelligence models.

160018 (2016)