IIPC Discretionary Funding Program 2021-2022

User-Friendly High Fidelity Browser-Based Crawling System for All

Proposal submitted by

Royal Danish Library, Anders Klindt Myrvoll, Programme Manager Netarkivet - the Danish Web Archive, ankm@kb.dk

Sept 15. 2021

Fill it in the form & the budget spreadsheet, download both files and send them as an attachment to projects@iipc.simplelists.com

PROPOSED PROJECT START DATE: January 1, 2022

LEAD IIPC INSTITUTION: Royal Danish Library

2ND IIPC INSTITUTION: UK Web Archive

3RD IIPC INSTITUTION: University of North Texas Libraries

4TH IIPC INSTITUTION: National Library of New Zealand

OTHER INSTITUTIONS OR CONSULTANTS: Webrecorder Software LLC - ilya@webrecorder.net

PROJECT LEAD CONTACT: ankm@kb.dk

PROJECT SECONDARY CONTACT: Andrew.Jackson@bl.uk

PROJECT 3RD CONTACT: lauren.ko@unt.edu

PROJECT 4TH CONTACT: ben.obrien@dia.govt.nz

TOTAL REQUESTED FUNDING FROM IIPC (IN USD)**: 50,000**

BRIEF ABSTRACT OF THE PROJECT: (Maximum 250 words)
[Provide a brief, succinct overview of the project.]

The goal of this proposal is to create a flexible, browser-based high fidelity crawling system driven by a full-featured user interface and accessible to curators and web archivists at any institution. The crawling system will focus on enabling the capture of complex, dynamic websites which can not be adequately captured with existing crawling tools such as Heritrix. The system will be built to be 'cloud native' and support running in the cloud, as well as in a local institutional environment. The core crawling engine will extend the Webrecorder Browsertrix Crawler system (https://github.com/webrecorder/browsertrix-crawler) for browser-based crawling, and the system will be built in a modular way to allow for future extensibility and customizations.

Netarkivet - the Danish Web Archive at the Royal Danish Library will lead the project, in partnership with the UK Web Archive, the University of North Texas Libraries and the National Library of New Zealand to establish design requirements. Development will be implemented by Webrecorder Software, while IIPC partners provide concrete use cases, help guide the development and contribute to testing and local deployment of the crawling systems. This group of partners will help ensure that the system can meet the varying needs of IIPC members, both libraries crawling on a national scale as well as smaller institutions. We hope this project will push the limits of browser-based crawling and provide a more concrete understanding of what is feasible and where the limits may be with a browser-based crawling approach.

GOALS, OUTCOMES AND, DELIVERABLES: (Maximum 150 words)
[Provide a description of deliverables that will be produced by the project.]

The deliverables will be an open-source, high-fidelity browser-based crawling system featuring:

- A well-defined REST API (OpenAPI spec) for crawl management
- A well-defined JSON-based crawl specification for defining crawl parameters and schedules.
- An intuitive user-interface for defining crawls (converted into the JSON-based spec)
- An intuitive user-interface for starting and monitoring crawls, and observing crawl logs/reports
- Instantaneous replay of crawled content during or after crawl
- A UI-based workflow for logging into websites via a remote browser, saving a browser profile, and launching crawls with pre-existing profiles.
- An automated QA system for evaluating the completeness of a crawl
- An automated patching mechanism, for running a crawl against an existing replay endpoint.
- A manual patching mechanism, for interactively patching via a remote browser.
- Support for deployment under Docker and Kubernetes.
- Standardized crawl artifacts (definitions, logs, screenshots)
- Documentation for deploying this system

HOW THE PROJECT FURTHERS THE IIPC STRATEGIC PLAN: (Maximum 250 words)
[Explain how the outcomes of the project support the mission of the IIPC detailed in its Strategic Plan.]

This project closely aligns with many of the goals in the IIPC strategic plan. A key goal of this work is to enhance the capability to archive complex and difficult web content in a consistent and easy-to-use way. Another goal is to work collaboratively to create a modular open-source system which provides a common feature set for browser-based crawling, and which can be used not only by the IIPC partners on this proposal, but any IIPC member, and the web archiving community at large. The partners on this project represent both national libraries as well as smaller archiving institutions with the goal of ensuring this work can meet the varied needs of different web archiving institutions based on use case driven development. In addition to creating a working product, we hope the approaches taken in this project will further the shared understanding of browser-based crawling, it's limits and possibilities and foster further work in this area.

DETAILED DESCRIPTION OF THE PROJECT: (Maximum 1,000 words)
[Provide a detailed description of the project. Include information on aspects such as phases of work, what participant is responsible for what activities and deliverables, how the project will manage and measure its progress and performance, any technologies used or developed, communications and data sharing plans, any risks or perceived threats to project success, and expect impact upon the field.]

The growing complexity of the web makes browser-based crawling essential for institutions to be able to capture many modern websites, including most social media sites. To date, no integrated, open source solution exists to run crawls that are comprehensive, scheduled, and browser-based (to enable automated capture of sites that are not possible to fully capture with Heritrix and other traditional approaches). During several of the IIPC open source community calls, the participants reached a consensus that browser-based crawling is a top priority, or even “needed yesterday”.

The Webrecorder project has specialized in developing high-fidelity capture tools, focusing on interactive browser-based capture. Webrecorder has also built the Browsertrix crawler system, which currently provides a low-level browser-based crawler inside a single Docker container.
Browsertrix Crawler can now be launched via command-line to run a single crawl at a time with a variety of low-level configuration options, including configuring crawl scope, number of browser workers and optional full text search extraction.

In this project, the goal will be to build on the existing Browsertrix Crawler component to provide a full-fledged user-friendly system with internationalization support, accessible to institutional curators and the web archiving community at large.

The system will be used to support the institutions in running high-fidelity crawls for a subset of seeds which currently do not work well with traditional crawling methods. The goal of the system is not to replace Heritrix, but to augment existing approaches with a new, targeted approach and explore the possibilities of browser-based crawling.

To provide maximum flexibility, the system will be ‘cloud-native’ and run in Kubernetes as well as in Docker and Docker Compose or Swarm to allow for similar local or on-site deployment.

The development of the project can be divided into four quarterly phases.

In Q1, the goal will be to create a well-defined crawling API, and ensure that the core scheduling system is operational, can run crawls, deposit WARC files, and generate basic logs. Webrecorder will establish a cloud instance of the system for testing by all of the project partners, and begin working with each institution to support local deployment in their environment.

In Q2, the core user interface will be ready for testing by non-developers, allowing for crawls to be created and run on the infrastructure by end users. The formats for all crawl artifacts, such as the crawl specification itself, the logging output, screenshots etc. will also be decided upon by this time.
Based on previous testing, a key requirement for high-fidelity crawling of social media is the ability to crawl while logged in. As part of the core crawling workflow, the ability to interactively log in to sites, save browser profiles and use them during a crawl will also be implemented.

In Q3, the focus will shift on adding improved QA features and exploring how to make automated QA more useful and reduce the amount of manual QA. We will explore various approaches, such as re-running a crawl on the archived data (the replay) and comparing results, such as missing URLs and javascript errors, producing screenshots, and creating heuristics for how to evaluate how ‘well’ a particular crawl ran. By this phase, the requirements of deploying the system at each institution will also be clearly understood.

Finally, in Q4 the focus will be on adding patching capability, including automated patching, as well as, if there is time, manual/interactive patching via remote browsers. In this phase, we will also evaluate the capabilities of the system to better understand how well browser-based crawling can (or can not) scale, and what limitations may exist. We hope to have a robust, flexible browser-based crawling system that can be used in production by many institutions. Webrecorder will continue to maintain a production service of this system for broader use. The final report will include an evaluation of browser-based crawling and identify key data points about crawl speed, capabilities and limitations of the system, to help make an informed decision on next steps for browser-based crawling techniques.

We understand that this is an ambitious project and does have some risks, which we hope to address with fully transparent development via GitHub and open communication about the progress. The development will be led by Webrecorder, and will consist of a small team, including the lead developer (Ilya Kreymer), as well as a part time UX designer and part-time frontend developer. The UX designer will help ensure that the interface design is clean and easy to use, and the frontend developer will assist with implementing the UI. The IIPC partners will be active partners in testing the system from the early stages to ensure that it meets their goals and expectations. Webrecorder will also use this as a key component for any crawling projects that it undertakes, and will seek to find supplementary funding to support development, such as through additional contract work.

We anticipate the expected impact to the field to be quite significant, as this crawling system will be designed to meet an urgent need within the web archiving community. We hope that the testing and deployment by the IIPC partners on this project will help pave the way for broader use and adoption by a greater number of IIPC members and anyone in the web archiving community interested in setting up a high fidelity crawling system.

PROJECT SCHEDULE OF COMPLETION: (Maximum 500 words)
[Provide a detailed breakdown of milestones and completion dates for all major activities of the project.]

The project schedule will involve top-level quarterly milestones, which will be roughly as follows:

Q1

- Initial design requirements are established
- Initial Crawling REST API Operational
- Crawling systems deployable in Docker and Kubernetes
- Webrecorder operates a hosted version of the system for testing.

Q2

- IIPC partners provide feedback on initial hosted deployment from Q1.
- Core UI operations implemented: logging in, creating crawl definitions, scheduling crawls, stopping crawls, viewing completed crawls, browsing replay.
- Logged-in workflow implemented (users log in to sites, then crawl with logged-in browser profile)
- Crawl formats are well-defined (crawl specification, crawl log format,, screenshot, etc..).

Q3:

- IIPC partners provide feedback on enhancements from Q2 and collaborate with Webrecorder on deployment documentation.
- Crawl QA features: screenshots, videos, automated QA testing, crawl QA reports being generated.
- Additional crawl artifacts (logs, screenshots, etc.) stored in a standardized way
- The system has been deployed successfully in different environments, local and cloud, by participating institutions and Webrecorder

Q4:

- Advanced crawling features, including patching existing crawls, or interactive patching using remote browsers
- Optimizations and benchmarks on capabilities of the system (how fast can it crawl, what are the scaling concerns, limits, etc.) produced, to be included in the final report.
- Webrecorder maintains a production ready cloud-based crawling system that will be offered to interested users.

DETAILED PROJECT BUDGET (in USD):
The detailed project budget is provided to projects@iipc.simplelists.com in a separate document as well as here: User-Friendly High Fidelity Browser-Based Crawling System for All- IIPC DFP 2021-2021 BUDGET

Regarding payment/invoicing the involved IIPC institutions have challenges around paying Webrecorder directly due to potential needed tender processes and have asked the Discretionary Funding Programme via mail to projects@iipc.simplelists.com if IIPC can pay Webrecorder, directly, with the approval of the project lead. We are waiting for an answer on this matter.

REVIEW TIMELINE:

1. See IIPC website for application deadlines.
2. Award announcements will be made by the end of November 2020.
3. All funded projects must start on January 1 and be completed by 31 January 2021.
4. A brief progress report is required six-months into the project and a final report, and all sharing of project deliverables, are required within 90 days of the completion of the project.