

Success

Application received by UKRI.

Submitted by Charis Wilson, from University of Edinburgh Research Office on 23 April 2024, 2:50pm.

APP39557: LURID3: Longitudinal studies of the World Wide Web

Application overview

This is what your application will look like.

LURID3: Longitudinal studies of the World Wide Web

1. Details

Summary

Longitudinal study of usage patterns of URIs in web pages allows us to detect trends that are important for managing the evolution of Web Architecture. The proposed work will extend our work on Common Crawl data in the LURID and LURID2 projects (EPSRC Tier 2, 2021-9 thru 2022-9 and 2022-12 thru 2023-04). We'll expand this work in two directions in order to provide the starting point for a larger subsequent research project:

1. Producing extended index files containing Last-Modified timestamps for multiple Common Crawl months

2. Filling in the PDF resources that are unusable in multiple Common Crawl months because they were truncated

Common Crawl has been a key provider of training data for AI text generation systems such as ChatGPT. Our research aims to both validate the representativeness of Common Crawl as a sample of the HTML content of the Web and to begin to fill in a significant gap in its coverage of the Web as a whole, namely its PDF content.

Start date

July 2024

Duration

9 months

2. Core team

Name	Role	Organisation
Henry S Thompson	Project lead	University of Edinburgh

3. HPC Service requested

Application question

Please name which HPC service are you applying to.

You may also name a reserve choice of service, providing you are willing to take up the offer of this service, should your first choice be oversubscribed and you are awarded time on your reserve choice.

The services available for this opportunity are:

- ARCHER2 (EPSRC remit only)
- Baskerville
- Bede (NICE)
- CSD3

- Cirrus
- Kelvin-2 (NI-HPC)
- JADE
- Sulis

Cirrus

4. HPC resources requested

Application question

Please enter the units and quantity of HPC resources that you are requesting.

Please enter the units (for example; CU, GPU hours, CPU hours) and quantity of HPC resources that you are requesting.

400,000 CPU Hours

Dataset storage: 400TB disk storage, already in place on /beegfs

Working storage: Cirrus standard provision

5. High End Computing (HEC) consortia

Application question

If you are a member of a currently funded HEC consortium, and are applying for ARCHER2 compute, please enter the name of the consortium and briefly explain why you are not applying for time through this consortium.

If you are a member of a currently funded HEC consortium, and are applying for ARCHER2 compute, please enter the name of the

consortium and briefly explain why you are not applying for time through this consortium.

N/A

6. Vision and Approach

Application question

What are you hoping to achieve with and how will you deliver your proposed work?

For the Vision, explain how your proposed work:

- is of excellent quality and importance within or beyond the field(s) or area(s)
- has the potential to advance current understanding, generates new knowledge, thinking or discovery within or beyond the field or area
- is timely given current trends, context and needs
- impacts world-leading research, society, the economy or the environment

Within the Vision section we also expect you to:

- identify the potential direct or indirect benefits and who the beneficiaries might be
- include how it would contribute to computational science, for example, through:
 - generating new codes
 - development of existing codes
 - increased computational efficiency
 - opening up HPC for new scientific areas and industrial sectors

For the Approach, explain how you have designed your work so that it:

- is effective and appropriate to achieve your objectives

- is feasible, and comprehensively identifies any risks to delivery and how they will be managed
- if applicable, uses a clear and transparent methodology
- if applicable, summarises the previous work and describes how this will be built upon and progressed
- will maximise translation of outputs into outcomes and impacts
- describes how your, and if applicable your team's, research environment (in terms of the place, and relevance to the project) will contribute to the success of the work

Within the Approach section we also expect you to:

- Include details of the codes and methods to be used and clarification of how they are appropriate for the work being proposed

References may be included within this section.

See attached

[APP39557 Vision and Approach.pdf](#)

0.07MB

7. Applicant and team capability to deliver

Application question

Why are you the right individual or team to successfully deliver the proposed work?

Evidence of how you, and if relevant your team, have:

- the relevant experience (appropriate to career stage) to deliver the proposed work
- the right balance of skills and expertise to cover the proposed work
- the appropriate leadership and management skills to deliver the work and your approach to develop others

- contributed to developing a positive research environment and wider community

The word count for this section is 1,650 words: 1,150 words to be used for R4RI modules (including references) and, if necessary, a further 500 words for Additions.

Use the Résumé for Research and Innovation (R4RI) format to showcase the range of relevant skills you and, if relevant, your team (project and project co-leads, researchers, technicians, specialists, partners and so on) have and how this will help deliver the proposed work. You can include individuals' specific achievements but only choose past contributions that best evidence their ability to deliver this work.

Complete this section using the R4RI module headings listed. Use each heading once and include a response for the whole team, see the [UKRI guidance on R4RI \(https://www.ukri.org/apply-for-funding/before-you-apply/resume-for-research-and-innovation-r4ri-guidance/#:~:text=UKRI%20committed%20to%20adopting%20a,UKRI%20throughout%202022%20and%202023.\)](https://www.ukri.org/apply-for-funding/before-you-apply/resume-for-research-and-innovation-r4ri-guidance/#:~:text=UKRI%20committed%20to%20adopting%20a,UKRI%20throughout%202022%20and%202023.). You should consider how to balance your answer, and emphasise where appropriate the key skills each team member brings:

- contributions to the generation of new ideas, tools, methodologies, or knowledge
- the development of others and maintenance of effective working relationships
- contributions to the wider research and innovation community
- contributions to broader research or innovation users and audiences and towards wider societal benefit

Additions:

Provide any further details relevant to your application. This section is optional and can be up to 500 words. You should not use it to describe additional skills, experiences, or outputs, but you can use it to describe any factors that provide context for the rest of your R4RI (for example, details of career breaks if you wish to disclose them).

Complete this as a narrative. Do not format it like a CV.

References may be included within this section.

UKRI has introduced new role types for funding opportunities being run on the new UKRI Funding Service.

For full details, see [Eligibility as an individual \(https://www.ukri.org/apply-for-funding/before-you-apply/check-if-you-are-eligible-for-research-and-](https://www.ukri.org/apply-for-funding/before-you-apply/check-if-you-are-eligible-for-research-and-)

[innovation-funding/eligibility-as-an-individual/#contents-list](#)).

Images:

Within this section you can also demonstrate elements of your responses in visual form, such as images, if relevant:

- use images sparingly and only to convey important information that cannot easily be put into words
- insert each new image onto a new line
- provide a descriptive legend for each image immediately underneath it (this counts towards your word limit)
- files must be smaller than 5MB and in JPEG, JPG, JPE, JFI, JIF, JFIF, PNG, GIF, BMP or WEBP format

Module 1 – Contributions to the generation of new ideas, tools, methodologies or knowledge

I am Professor of Web Informatics, based in the Institute for Language, Cognition and Computation in the School of Informatics at the University of Edinburgh. I have been carrying out research and supervising student projects in the area of Web Architecture since 2004.

Selected relevant publications:

1. **Thompson, H. S.** (2010b) “What's a URI and why should I care?”. *Ariadne* 65. Online journal, available at <http://www.ariadne.ac.uk/issue65/thompson/> (<http://www.ariadne.ac.uk/issue65/thompson/>).
2. **Thompson, H. S.**, J. Rees and J. Tennison (2013) “URIs in data: for entities, or for descriptions of entities—A critical analysis”. In *Proceedings of WebSci '13*, ACM, 479–482. Available online at <https://doi.org/10.1145/2464464.2532514>. (<https://doi.org/10.1145/2464464.2532514>).
3. **Thompson, H. S.** and C. Lilley, eds. (2014) *XML Media Types*. Internet Engineering Task Force (IETF). IETF Standards Track RFC 7303, available online at <https://tools.ietf.org/html/rfc7303>. (<https://tools.ietf.org/html/rfc7303>).
4. **Thompson, H. S.** and J. Tong (2018) “Can Common Crawl reliably track persistent identifier (PID) use over time?”. In *Proceedings of The Web Conference 2018*, ACM, 1749–1755. Available online

at <https://doi.org/10.1145/3184558.3191636>
(<https://doi.org/10.1145/3184558.3191636>).

5. **Thompson, H. S.** and Norman Walsh, eds. (2008) *Associating Resources with Namespaces*. W3C TAG Finding, World Wide Web Consortium, Cambridge. Available online at <http://www.w3.org/2001/tag/doc/nsDocuments/>
(<http://www.w3.org/2001/tag/doc/nsDocuments/>).

6. **Thompson, H. S.** et al. (2022-01-10). Thread on common-crawl Google Group: https://groups.google.com/g/common-crawl/c/AmsXrCNVBzo/m/us4j_7tpAAAJ.
(https://groups.google.com/g/common-crawl/c/AmsXrCNVBzo/m/us4j_7tpAAAJ).

7. **Thompson, H. S.** (2024) “Improved methodology for longitudinal Web analytics using Common Crawl”. To appear in *Proceedings of ACM Web Science Conference (WEBSCI '24)*, May 21–24, 2024, Stuttgart, Germany. Available online at <https://doi.org/10.48550/arXiv.2404.09770>
(<https://doi.org/10.48550/arXiv.2404.09770>).

Module 2 – The development of others and maintenance of effective working relationships

My empirical investigation of Web Architecture has provided the context for the dissertations of six generations of MSc students whom I have trained in the use of HPC techniques. Several of them have gone on to further research in this area.

Module 3 – Contributions to the wider research and innovation community

My previous HPC project **LURID2** came up with two related approaches to reduce the resources required to make use of Common Crawl datasets for Web research, based on exploiting and extending the much smaller (200GB compressed) index which is available for each archive, as follows:

- By adding Last-Modified timestamps to the index we enabled fine-grained longitudinal exploration using only a single archive;
- By comparing the distribution of index features for each of the 100 segments into which each archive is packaged with their distribution over the whole archive, we identified the least and most representative segments for a number of recent archives. Using this

allows the segment(s) that are most representative of an archive to be used as proxies for the whole.

Combining these two approaches allowed us to produce a fine-grained analysis of how URI lengths have changed over time, leading to an unanticipated insight into the how the process of creating of Web pages itself has changed.

I'm a regular contributor to the online Common Crawl user forum.

I have just published the augmented index online at <https://markup.co.uk/ccrawl/> (<https://markup.co.uk/ccrawl/>), and will be announcing this at the WebSci24 conference next month.

Module 4 – Contributions to broader research/innovation-users and audiences and towards wider societal benefit

I was a member of the SGML Working Group of the World Wide Web Consortium (W3C) which designed XML, a major contributor to the core concepts of XSLT and W3C XML Schema and a member of the XML Core and XML Processing Model Working Groups at the W3C. I have been the University of Edinburgh's representative on the Advisory Council (AC) of the W3C since we joined in 1998. I was elected by the AC to four successive terms (9 years) on the W3C Technical Architecture Group (TAG). The TAG is chaired by Sir Tim Berners-Lee, and is responsible for the stewardship of the architecture of the Web.

I regularly review draft standards (RFCs) for the Internet Engineering Task Force (IETF) in the areas of Web protocols and data formats.

I have been a contributor to the Edinburgh Science Festival, Cafe Scientifique, a British Science Association Talk/Film series, the Albertus Institute and other public understanding of science fora.

I am committed to the proposition that our professional responsibilities and the success of our intellectual pursuits both depend on keeping our humanity clearly in view.

8. Computing resources and justification

Application question

What computing resources will you need to deliver your proposed work?

Evidence of:

- full justification of the resources requested and why you are requesting the specific system
- justification of why the computing time requested is essential for the successful completion of the project
- assurance that the resources requested are appropriate and justified

Assessors are not looking for a line-by-line breakdown of all project resources. Overall, they want you to demonstrate how the resources you anticipate needing for your proposed work:

- are comprehensive, appropriate, and justified
- represent the optimal use of resources to achieve the intended outcomes
- maximise potential outcomes and impacts.

Based on experience in our previous two HPC projects, we estimate that producing an augmented index for one CC dataset needs around 10,000 CPU hours, so for our target of 11 additional such we will need 110,000 CPU hours. Storing the results only needs 2.2TB of disk space, but the complete datasets are needed to get the data for the augmentation. This will need something on the order of 300TB, depending on exactly which ones we decide to work with.

The longitudinal studies we have in mind for the resulting augmented indices will need around 20,000 CPU hours each, for 60,000 in total.

Although fetching the additional PDFs will have to be done overnight on the Cirrus login nodes, because the compute nodes do not have Web access, indexing the results can be done on the compute nodes, and we estimate 10,000 CPU hours each will be required, so for 12 collections that's 120,000 CPU hours.

The longitudinal studies thereof are harder to estimate, as they will require development and testing as well as actual result runs, but

35,000 each seems a reasonable estimate, for 105,000 in total.

Based on our existing augmented PDF content collection, each augmented collection will need about 2TB, so for 11 additional such we will need 22TB.

It not only makes sense to use Cirrus for this work, as we are familiar with the environment, to which we have added a significant amount of specialised tooling, but it's also necessary, as it gives us access to the previously downloaded datasets as well as the necessary space for new data.

Overall this gives a requirement of $110K + 60K + 120K + 105K = 395,000$ CPU hours and $300TB + 2.2TB + 22TB = \sim 325TB$ of disk space, which is within the 400TB already available to us.

9. Ethics and responsible research and innovation (RRI)

Application question

What are the ethical or RRI implications and issues relating to the proposed work? If you do not think that the proposed work raises any ethical or RRI issues, explain why.

Demonstrate that you have identified and evaluated:

- the relevant ethical or responsible research and innovation considerations
- how you will manage these considerations

If you are collecting or using data, identify:

- any legal and ethical considerations of collecting, releasing or storing the data including consent, confidentiality, anonymisation, security and other ethical considerations and, in particular, strategies to not preclude further re-use of data
- formal information standards with which your study will comply

Images:

Within this section you can also demonstrate elements of your responses in visual form, such as images, if relevant:

- use images sparingly and only to convey important information that cannot easily be put into words
- insert each new image onto a new line
- provide a descriptive legend for each image immediately underneath it (this counts towards your word limit)
- files must be smaller than 5MB and in JPEG, JPG, JPE, JFI, JIF, JFIF, PNG, GIF, BMP or WEBP format

All our research to date, as well as the proposed work, conforms to the Common Crawl (CC) Terms of Use. We have sought and received from CC confirmation that our publication on the Web as part of LURID2 of a 220GB augmented index to the August 2019 CC dataset would be allowed, and this permission would also apply to the publication of similar augmented indices for additional CC datasets produced during the proposed work.

The proposed study of augmented CC PDF data raises no new legal or ethical considerations. In the first instance we will hope to *publish* our augmented PDF dataset(s) via CC itself. If this is not possible, publication will only go ahead after the University of Edinburgh legal department approves it.

10. Genetic and biological risk

Application question

Does your proposed research involve any genetic or biological risk?

In respect of animals, plants or microbes, are you proposing to:

- use genetic modification as an experimental tool, like studying gene function in a genetically modified organism
- release genetically modified organisms
- ultimately develop commercial and industrial genetically modified outcomes

If yes, provide the name of any required approving body and state if approval is already in place. If it is not, provide an indicative timeframe for obtaining the required approval.

Identify the organism or organisms as a plant, animal or microbe and specify the species and which of the three categories the research relates to.

Identify the genetic and biological risks resulting from the proposed research, their implications, and any mitigation you plan on taking. Assessors will want to know you have considered the risks and their implications to justify that any identified risks do not outweigh any benefits of the proposed research.

If this does not apply to your proposed work, enter 'N/A' in the textbox.

N/A

11. Research involving the use of animals

Application question

Does your proposed research involve the use of vertebrate animals or other organisms covered by the Animals Scientific Procedures Act?

If you are proposing research that requires using animals, download and complete the [Animals Scientific Procedures Act template \(DOCX, 74KB\)](https://ukri-tfs-prod-assets.s3.eu-west-2.amazonaws.com/AnimalResearchQuestionTemplate-BBSRC.docx). (<https://ukri-tfs-prod-assets.s3.eu-west-2.amazonaws.com/AnimalResearchQuestionTemplate-BBSRC.docx>), which contains all the questions relating to research using vertebrate animals or other Animals (Scientific Procedures) Act 1986 regulated organisms.

If this does not apply to your proposed work, enter 'N/A' in the textbox.

N/A

12. Conducting research with animals overseas

Application question

Will any of the proposed animal research be conducted overseas?

If you are proposing to conduct overseas research, it must be conducted in accordance with welfare standards consistent with those in the UK, as in Responsibility in the use of animals in bioscience research, page 14 (<https://www.nc3rs.org.uk/3rs-resources/responsibility-use-animals-bioscience-research>). Ensure all named applicants in the UK and overseas are aware of this requirement.

If your application proposes animal research to be conducted overseas, you must provide a statement in the text box. Depending on the species involved, you may also need to upload a completed template for each species listed.

- All named applicants are aware of the requirements and have agreed to abide by them
- this overseas research will be conducted in accordance with welfare standards consistent with the principles of UK legislation
- the expectation set out in Responsibility in the use of animals in bioscience research will be applied and maintained
- appropriate national and institutional approvals are in place

Templates

Overseas studies proposing to use non-human primates, cats, dogs, equines or pigs will be assessed during NC3Rs review (<https://www.nc3rs.org.uk/who-we-are-and-what-we-do#services>) of research applications. Provide the required information by completing the template from the question 'Research involving the use of animals'.

For studies involving other species, select, download, and complete the relevant Word checklist or checklists from this list:

- Additional questions on the use of rodents overseas (DOCX, 49.1KB)
(https://nc3rs.org.uk/sites/default/files/documents/Additional%20questions%20on%20the%20use%20of%20rodents%20overseas%20%286%29_0.docx).
- Additional questions on the use of rabbits overseas (DOCX, 49.2KB)
(<https://nc3rs.org.uk/sites/default/files/documents/Additional%20questions%20on%20the%20use%20of%20rabbits%20overseas.docx>).
- Additional questions on the use of sheep overseas (DOCX, 50.9KB)
(<https://nc3rs.org.uk/sites/default/files/documents/Additional%20questions%20on%20the%20use%20of%20sheep%20overseas.docx>).

- Additional questions on the use of goats overseas (DOCX, 47.3KB)
(<https://nc3rs.org.uk/sites/default/files/documents/Additional%20questions%20on%20the%20use%20of%20goats%20overseas.docx>).
- Additional questions on the use of pigs overseas (DOCX, 51.4KB)
(<https://nc3rs.org.uk/sites/default/files/documents/Additional%20questions%20on%20the%20use%20of%20pigs%20overseas.docx>).
- Additional questions on the use of cattle overseas (DOCX, 57.0KB)
(<https://nc3rs.org.uk/sites/default/files/2021-12/Additional%20questions%20on%20the%20use%20of%20cattle%20overseas.docx>).
- Additional questions on the use of *Xenopus laevis* and *Xenopus tropicalis* overseas (DOCX, 57.3KB)
(<https://nc3rs.org.uk/sites/default/files/Additional%20questions%20on%20the%20use%20of%20Xenopus%20laevis%20%26%20%20Xenopus%20tropicalis%20overseas.docx>).

If this does not apply to your proposed work, enter 'N/A' in the textbox.

N/A

13. Research involving human participation

Application question

Will the project involve the use of human subjects or their personal information?

If you are proposing research that requires the involvement of human subjects, provide the name of any required approving body and whether approval is already in place.

Justify the number and the diversity of the participants involved, as well as any procedures.

Provide details of any areas of substantial or moderate severity of impact.

If this does not apply to your proposed work, enter 'N/A' in the textbox.

N/A

14. Research involving human tissues or biological samples

Application question

Does your proposed research involve the use of human tissues, or biological samples?

If you are proposing work that involves human tissues or biological samples, provide the name of any required approving body and whether approval is already in place.

Justify the use of human tissue or biological samples specifying the nature and quantity of the material to be used and its source.

If this does not apply to your proposed work, enter 'N/A' in the textbox.

N/A
