Access to High Performance Computing Call: Application Form

Closing date for applications to EPSRC: 18th October 2022 at 16:00

Closing date for the technical assessment (needs to be included in your application): 20th September 2022 at 16:00

Applicants should note that, unless otherwise specified, standard guidance for completion of EPSRC proposals applies.
https://epsrc.ukri.org/funding/applicationprocess/fundingguide/

All documentation should be submitted as pdf documents to aid processing. A completed and approved technical assessment form should be submitted directly to the service you wish to access as a separate document prior to the technical assessment deadline (contact details can be found in the service specification document).

Provided the technical assessment endorses the proposal, **applications should be submitted via smartsurvey** Start your application (SmartSurvey). **including the completed technical before the call deadline.**

Please refer to the appropriate call document for more information on how to fill out the application form.

Organisation:	**University of Edinburgh**
Division or department:	**School of Informatics**

Project Title

LURID2: Assessing the validity of Common Crawl

Start Date and Duration

See service specific restrictions in Appendix 1 of the Access to HPC call document.

Proposed start date:	1 January 2023
Project duration:	1 year

Requested resource

Service you are applying to:	**Cirrus**
Units requested (e.g. CU, core hours, GPU hours)	**500,000 CPU Hours**

Applicants

	Principle Investigator
Title	**Professor**
First Name	**Henry S**
Last Name	**Thompson**
Organisation	**University of Edinburgh**
Division /Department	**School of Informatics**
Address	**4.22 Informatics Forum; 10 Crichton Street; Edinburgh; EH8 9AB**
Email	**ht@inf.ed.ac.uk**

Are you a member of a currently funded HEC consortium (see call document)

Yes, consortium name: _____ **No** [X]

If yes and you are applying for ARCHER2 compute, please briefly explain why you are not applying for time through this consortium:

Please note that if ESPRC staff judge that the proposal is potentially in the remit of a HEC consortium then the proposal may be shared with the relevant consortium chair. The proposal will be shared with the relevant chair if it is successful.

For guidance in completing the remainder of this application form, please refer to the 'Guidance for writing an application' section of the Access to HPC call document.

Objectives (up to 1/2 page)

The proposed work will build on preliminary results from our just-completed project (*LURID: Longitudinal study of URI deployment*, EPSRC Access to HPC Award 2021-07-30–2022-09-01) on the Common Crawl web archive (CC), in the course of which we accomplished our three main objectives. Some preliminary progress was made on the (optional) fourth, in that we started work on a novel methodology for empirically assessing the representativeness of the 100 segments into which each monthly edition of CC is divided.

Objective 1: Establish that our new methodology's results are statistically significant

We can now measure the extent to which each of the 100 segments, each of which has about 3×10^7 request-response pairs, is representative of the archive as a whole, with respect to range of request and response properties present in the (much smaller) archive index. We need to refine this so we can give a significance figure for each measure, and then determine which property or combination of properties gives the most reliable result. We also need to confirm that sharing properties of the *index* does guarantee sharing properties of the full archive that are *not* in the index.

Objective 2: Confirm that our method scales up from one archive to many

To date we've focussed on just two monthly CC archives. We propose to expand this considerably, to each of the six January archives from 2018–2023, and all the archives since January 2021 (15 or 16). Using the results from Objective 1, we can then identify the 'best' segment(s) for each of those archives.

Objective 3: Share the Objective 2 results with the Common Crawl community

We've already received very encouraging feedback in response to the brief mention we made of our work to date on the Common Crawl mailing list **[6]**, because it will enable piloting a wide range of tasks on just a single segment from an archive of interest, knowing that the results are very likely to be replicated over the whole.

Objective 4: Move up to the meta-level

If time allows, explore the possibility of using the methodology to compare whole monthly archives. If successful, this will give some measure of the extent to which CC archives provide 'good' samples of the Web.

Description of proposed research and its context (up to 2 ½ pages)

Context: Empirical evidence of how use of the Web has changed in the past provides crucial input to decisions about its future. Creative uses of the mechanisms the Web provides expand its potential, but also sometimes put it at risk, so it's worrying that there's surprisingly little empirical evidence available to guide standardization and planning more generally. Which aspects of the Web's functionality are widely used? Hardly ever used? How is this changing over time?

The kind of evidence needed to answer such questions is hard to come by. The proposed research builds on our previous work in this area **[4]**, taking advantage of the computational resource Cirrus provides to validate and expand our work on the Common Crawl web archive.

Common Crawl (CC) is a very-large-scale web archive, containing petabytes of data from more than 65 monthly archives, totalling over 100 billion web pages. Collection began in 2008 with annual archives, expanding steadily to the point that since 2017 archives are collected monthly or bi-monthly. Recent archives contain over $3x10^9$ pages, about 50 Terabytes (compressed). Together with Edinburgh colleagues we have created local copies of 8 months of CC in a petabyte store attached to Cirrus. For our purposes it is important to note that the overlap between any two archives as measured by Jaccard similarity of page checksums is less than .02 **[9]**.

Our recent work has focussed on exploiting the much smaller index for each archive. An archive index only takes about 200GB using a sharded CDX format **[7]**. It contains one line for every URI that was requested, which not only point to the corresponding request and response entries in the much larger WARC-format archive data files, but also contain several pieces of metadata about the response: length, character set, media type, status code, checksum and, for HTML responses, natural language(s) used.

The ultimate goal of our work is to use the *content* of Web requests and responses to help track and predict the *functioning* of the web. Before switching to looking at indices, we looked at request headers (for cookies and Last-Modified dates) and HTML responses (for various properties of the links therein). Each of these required processing a full 50-TB archive for each month in a longitudinal study. This imposed substantial resource requirements, both in terms of storage and compute cycles, which in turn limited the scale of study that we could undertake. The proposed work will allow us to work around this by using the much smaller indices to identify a single segment from each archive we want to study, 1% of the whole, which we can be confident has the same distribution of properties as the whole. The largest study we have been able to complete to date **[8]** compared one property (number of cookies per request) from one monthly archive from each of six years. Using the technique proposed here will enable us to expand such studies to cover

two to three times as many archives using less than 3% of the resources.

Research: There is a long-standing problem with any use of web archives to study the Web: how do we know if the results are valid? To put it another way, how can we measure the extent to which an archive is *representative* of the Web as a whole? For some time now, *inter alia* because of the extent to which many web sites build pages on-the-fly, parameterised by a wide range of contextual factors, it does not make sense to ask how big the Web is: the number of available Web pages is *unbounded*. So it is not possible to say what percentage of the Web is covered in a CC monthly archive. However, we have recently **[8]** found a solution to a simpler problem, which we believe can be extended to give an empirically-based measure of the representativeness of CC archives.

The simpler problem is this: are the 100 segments into which each CC archive is divided themselves representative of the archive as a whole? First we established that the *proportion* of certain easy-to-measure properties of the segments were very similar between the different segments, and between each segment and the whole. Easy to measure, because they were taken from the index, not the archive data itself, namely, the proportion of http: to https: URIs sampled, the proportion of English-only pages and the proportion of top-500 domains. We moved on to study the *distributions* of the languages used, using rank correlation tests, with similar positive results, and will extend that to the distributions of media types.

Dataset: The four studies described below will all use the same datasets, held on BeeGFS: 6 complete one-month CC archives, every January from 2018 through August 2023, both indices and full request/response data, as well as approximately 16 further indices, covering the rest of 2021 and 2022.

Specific studies:

For objective 1) A series of experiments on a small number of archives. Phase 1: Tuning the way we measure the similarities between properties derivable from a segment index, in order to find the best (combination of) properties in terms of statistical significance, relationships between different measures and error analysis. For example, does rank correlation between languages used pick out the same 'good' segments as rank correlation between media types? What explanation can we find for segments which are really 'bad' (and we have seen a few such in our earlier work)? Phase 2: Comparing results from the output of Phase 1 with results from properties of the response headers and bodies that are *not* present in the indices. For example, if we pick a segment based on a good rank correlation with respect to media types, do we see a good correlation with respect to Last-Modified headers.

For objective 2) Replicate Phase 1 above across all 22 indices. Look out in particular for any trends with respect to different measures showing a change in the number of 'bad' segments. Replicate Phase 2 above across six full archives.

For objective 3) Draft a paper describing the results of the above, and circulate it to the Common Crawl community, along with a tabulation of segments we recommend for use in experiments on longitudinal change across the 22 segments we have analysed. Hold a Turing workshop to present this work and illustrate how it can be used for new longitudinal studies.

For objective 4) Compare the index-based measures longitudinally, across both years and months. Which of the measures, if any show either little change or more-or-less smooth evolution? For example, we would expect the ratio of `https:` to `http:` URIs to grow over time, but the distribution of the most common media types to be pretty stable.

If we get mostly positive results, that suggest in those cases we probably have learned something true about the phenomena for the Web as a whole.

References

7. Kreymer, I. (2015). *CDX Index Format*. Wiki page: https://github.com/webrecorder/pywb/wiki/CDX-Index-Format#zipnum-sharded-cdx

8. Chen, J. (2021). *A Survey on HTTP cookies*. MSc thesis, School of Informatics, University of Edinburgh.

9. Nagel, S. (2022). *Statistics of Common Crawl Monthly Archives*. Online at https://commoncrawl.github.io/cc-crawl-statistics/plots/craw-loverlap

Importance (up to 1 page)

As Tim Berners-Lee once said "We have to study the Web so we don't break it by mistake". There's a lot of low-level empirical data about Internet traffic, and a fair amount about Web traffic volume on a per-host basis, but much less is known about how the pages on the Web themselves are evolving. For example, the debate over the future of HTML in the early part of this century was seriously hampered by the lack of publicly available, reproducibly tabulated, empirical evidence of the use of XHTML, and indeed in our own work we have seen a recent *increase* in XHTML use.

Methodological advance

By giving concrete examples of what can be achieved by applying High Performance Computing resources to very large scale web archive data, and by explaining how this can be replicated and why it can be trusted, we will raise the bar for longitudinal study of the Web.

Policy influence

Standards bodies, in particular the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF), as well as funding bodies, can do their job better when planning is evidence-based. Our results will contribute to this, in a small way in themselves, and on a larger scale if our approach is adopted more widely.

Scientific importance

Our studies of validation, whichever way they turn out, will make a big difference to how subsequent studies based on Common Crawl are carried out and evaluated. As the most widely available and easily used very large scale Web archive, it is crucial that we know the extent to which we can trust results derived from Common Crawl datasets.

Expertise and track record of the team (up to 1 page)

Henry Thompson is Professor of Web Informatics, based in the Institute for Language, Cognition and Computation in the School of Informatics at the University of Edinburgh. He has been carrying out research and supervising student projects in the area of Web Architecture since 2004. He has been the University of Edinburgh's representative on the Advisory Council (AC) of the World Wide Web Consortium (W3C) since we joined in 1998. He 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.

Thompson is a Fellow of the Alan Turing Institute, which has provided a valuable forum for discussions about the exploitation of very-large-scale Web archives.

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/.

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

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.

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

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/

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

Other associated resources (up to 1/2 page)

Storage: As described in the accompanying Technical Assessment, we will have access to approximately 300 TB of the BeeGFS file store, which is accessible from Cirrus nodes.

Resource Management (up to 1½ pages)

The key resources for this project are the BeeGFS file store and the compute nodes of Cirrus HPC facility. Data parallelism is the key to effective exploitation of very large scale web archive datasets. The experimental work described above makes increasing demands across the lifetime of the project on the amount of data and processing power required, but we anticipate that this will peak in the 3rd quarter of the project year, as shown in the accompanying Technical Assessment and Gantt chart.