Access to High Performance Computing Call: Application Form

Closing date for applications to EPSRC: 30th April 2021 at 16:00

Closing date for the technical assessment (needs to be included in your application): 2nd April 2021 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 appendix 1 in the Access to High Performance Computing call document).

Provided the technical assessment endorses the proposal, **applications should be submitted via smartsurvey (**https://www.smartsurvey.co.uk/s/1GBOOJ/) **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

LURID: Longitudinal study of URI deployment

Start Date and Duration

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

Proposed start date:	1 September 2021
Project duration:	1 year

Requested resource

Service you are applying to:	**Cirrus**
Units requested (e.g. CU, core hours, GPU hours)	**360,000 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 ☒

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 extend both the granularity and of our previous longitudinal studies based on the Common Crawl web archive (CC). Efficiently processing the 50TB (compressed) monthly CC archives requires large scale data-parallelism as provided by Cirrus.

Objective 1: Measure the shift from `http:` to `https:`

- Is the rate of shift decreasing? This has important implications for the long-term value of (existing approaches to) web caching.
- Are particular types of web pages more resistant to change than others? If so, targeted information campaigns become possible.

Objective 2: Measure the uptake of persistent identifiers (PIDs)

- Is there any sign of significant usage outside the scholarly journals?
- What are the relative trends as between single-scheme resolvers (e.g. `doi.org` for DOIs, `hdl.handle.net` for handles) vs. generic resolvers (e.g. `n2t.net`, `identifiers.org`)

If either or both of the above suggest significant impending growth, support for resolvers and agreement about resolver protocols may need to move from private to public hands, or at least become publicly-supported.

Objective 3: Measure the increase in fanout (the number of HTTP requests required to render a single web page)

- Prioritising aspects of HTTP protocol development/replacement depends in part on predicting this.

Objective 4: Cross-validate all the above analyses by splitting the 10 months of data from five years available to us and comparing the results from the two five-year longitudinal samples

- This will give some measure of the extent to which CC is a ‘good’ sample of the Web.

Description of proposed research and its context (up to 2 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 much larger computational resource Cirrus would provide to extend both the granularity (by exploiting page header timestamps) and extent (from 3-year to 5-year) of our previous longitudinal studies based on the Common Crawl web archive.

Common Crawl (CC) is a very-large-scale web archive, containing petabytes of data from more than 60 monthly crawls, totalling over 100 billion web pages. Collection began in 2008 with annual crawls, expanding steadily to the point that since 2017 crawls happen monthly. Recent months contain over $3x10^9$ pages, about 50 Terabytes (compressed). Together with Edinburgh colleagues we have created local copies of 7 months of CC in a petabyte store attached to Cirrus, and this project would enable us to add 3 more. Earlier work was slow because we were restricted to streaming access from CC originals stored on Amazon S3 in California, but we now, with 10 months attached to Cirrus, we can do much larger scale analysis in more detail.

Research: The focus of our work is on using the *content* of the Web to help track and predict the *functioning* of the web. By looking at web page content, in particular at the kinds of links that pages contain, and where in the pages they occur, we can learn what aspects of Web architecture are gaining or losing popularity.

Our existing work with CC data has been limited in a number of ways which the proposed work will address:

- Our previous work used one one-month CC archive from each of three years, streamed over the Internet. Processing even this limited amount of data in this way with existing Informatics resources took weeks for a single pass. The proposed work will be able to process two one-month archives from each of five years, in less time.
- We used CC's tabulation of the links in web pages, but this covers only HTML pages. Cirrus resources will enable us to extract links from PDF pages as well.
- We compared changes in web page properties over time by reference to the month the pages were collected. This is a coarse measurement: in the proposed work we will supplement this by reference to the Last-Modified HTTP header of each page for which this is available.

Dataset: The three studies described below will all start from the same dataset: 10 one-month CC archives held on BeeGFS, every six months from February 2017 through August 2021. In the case of PDF pages, we'll augment the CC data by extracting all the links found in each page, and we'll augment all link tabulations (from HTML and PDF pages) with Last-Modified dates when possible.

Specific studies: 1) `http:` vs. `https:` Since the launch of the "Let's Encrypt" campaign in 2015, there has been a significant shift towards encrypted access to web pages. We will explore the recent time-course of this trend, comparing the trends as found in HTML and PDF pages. We're particularly interested to see if the trend is showing signs of levelling off, possibly only in some application areas.

2) Persistent identifiers (PIDs): Academic journals have begun requiring the use of Digital Object Identifiers(DOIs) in the articles they publish. We suspect that our previous study of DOI use significantly under-estimated it, because it only looked at links in HTML pages, whereas virtually all journals publish online using PDF. We can now remedy this, and compare PID use between HTML and PDF. We'll also compare the use of generic versus single-scheme resolver hosts in PID links, which is of critical importance for projecting server demand.

3) Fanout: The rapid growth of script-based HTML page design has driven a need for support for multiplexing of HTTP requests, as manifested by HTTP 2.0 and a range of exploratory alternatives. By looking at the growth of links within the HTML head, and making use of the additional information about redirects provided by CC, we can plot the this growth in detail.

4) Validation: 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. But the sampling procedure used by CC *is* designed to be representative of the *hosts* that serve the Web, and that number is finite.

The three studies described above will provide a basis for testing this design, and thus to some extent the validity of their results. We will begin that work as part of this project if time allows, otherwise in a subsequent project.

Doing this will take advantage of the unprecedented scale of the data and processing power we will have from Cirrus. Each CC month is structured is such a way as to allow for the creation of sub-samples which *should* have the same properties as the whole. By dividing each of our 10 CC months into two equal-size sub-samples, and comparing the distributions of results we get by re-running the above studies, we can get measurements, with error bars, of their similarity to each other, and to the whole dataset.

Either way, we will learn something important.

- If we get mostly negative results, that suggests that Common Crawl's sampling approach is flawed, and should be rethought.
- If we get mostly positive results, we know that in those cases we probably have learned something true about the phenomena. It will then also be useful to further sub-divide. By plotting the error-rate as the size of the samples decreases, we can provide a guide for subsequent research: for no more than X amount of error, you need at least Y% of a CC month.

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.

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, 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

Outcome 4, Validation, whichever way it turns 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/

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 filestore, which is accessible from Cirrus nodes.

Resource Management (up to 1½ pages)

The key resources for this project are the BeeGFS filestore and the compute nodes of Cirrus HPC facility. Data parallelism is the key to effective exploitation of very large scale web archive datasets, and we have developed tools and methodologies which make it straightforward to tabulate the kinds of distributions required for each of the outcomes described above. With the resources described in the accompanying Technical Assessment, we can process 10 months of CC data in 24—36 hours.