CCP-AHC Roadmap Open Draft

Collaborative Computational Project for Arts, Humanities and Culture

Prepared by: CCP-AHC Delivery Team
Status: Preliminary DRAFT
Date: 2025-09-11

Note: This is an ongoing and open draft summary of the roadmap for CCP-AHC, building on the outcomes of our engagement events so far including the CCP-AHC Town Hall 2025. It begins with the core of the roadmap: the requirements for the new community. Please send your comments and corrections to ccpahc@durham.ac.uk, or provide them as Hypothes.is annotations online.

Summary of CCP-AHC Requirements

Preface

In what follows, topics of interest to the community have been distilled from contributions made during collaborative, round-table elicitation exercises at the CCP-AHC Town Hall 2025. **We are aware that there is a large and growing body of academic research, research policy, and planned work in the area of (digital) research infrastructures for the arts and humanities. We do not review this literature here.** For a more on the background to the CCP-AHC initiative, start with the About page on the project's website.

Each topic is followed by a set of actionable activities. Not all of these can be taken forward by CCP-AHC. Readers are invited to reflect on the relative importance of these proposed actions and to identify further topics and activities as necessary.

Contributions are not attributed to specific individuals. Where specific individuals, institutions, research projects or infrastructures are named, this is done to illustrate more general points and does not signal an endorsement or commitment to support by CCP-AHC.

Uses of large-scale compute in AHC research

A number of specific projects, investigators, and institutions who have made use of or were planning to make use of large-scale compute or other UK digital research infrastructure (DRI) were mentioned (in no particular order):

- Rosa Filgueira (EPCC) has developed, over many years, **distributed computing frameworks to support text analysis of digitised collections at scale**, including

geoparsing applications led by Beatrice Alex (University of Edinburgh), which allows **researchers to convert placenames to geospatial co-ordinates**
- Researchers and technologists at The National Archives use AI techniques **to develop semantic search for document selection and discovery**, and steward many terabytes of government-originating data as part of the UK Web Archive (UKWA)
- Pieter Francois (Oxford University) examines **the spread of ideas over large spans of space and human history** with the help of new geospatial humanities datasets.
- A team of researchers led by Pontus Stenetorp (University College London) has **fine-tuned large language models (LLMs) on a multilingual corpus of British languages** and developed new benchmarks for their evaluation
- Leonardo Impett (University of Cambridge) uses AI technologies **to conduct digital art history and to explore the history of seeing**
- Ruth Ahnert (Queen Mary University of London) has led a multi-year project (Living with Machines) producing **many software outputs for large-scale analysis of AHC data**
- Researchers at the British Library are using **large language models (LLMs) to turn unstructured data into structured data**
- Will Lamb (University of Edinburgh) and others are working with the National Library of Scotland on **machine learning technologies to support the Gaelic language (*Gàidhlig*) including subtitling, transcription**, and other tasks
- Iain Emsley (University of Warwick) has used **distant "sonification" as a critical computational method as an alternative way** of analysing data
- Daniel Chávez Heras (King's College London) uses **LLMs and vision-language models (VLMs) for moving-image (e.g. films, advertisements) understanding** at scale

These projects have benefited from various sources of funding over the last ten years and more, including UKRI, Research England, HEIs and IROs themselves, Jisc (which no longer plays a major role), as well as the in-kind personal contributions of those unfunded by external sources. The details of their use of large-scale compute or other DRI can sometimes be found in publications or supplementary materials but rarely have these been consolidated.

The value of producing a set of case studies or use cases, which illustrate the use of a given infrastructure to produce a given set of AHC research results by a given research project, was recognised. It was felt that such write-ups are more useful upon project conclusion, and that it is also important to collect applications and workflows conceived

of - but not realised - during a project. It was also noted that a structured approach to recording use cases could also be useful earlier in the lifecycle of a project using large-scale compute. The relevant details could be collected and expressed in many forms/formats (e.g. blogs, technical notes, podcasts etc). Whatever the case, there is a need for the co-ordination and collection of use cases.

The precise technical requirements for these projects nor the details of the compute infrastructures used to realise them were not always at hand to the participants at the Town Hall (many projects were reported second-hand). However, most attendees were comfortable articulating computational requirements in the more abstract language of workflows. This provides an opportunity to elicit real or imagined applications from researchers with various levels of skill. Workflows can offer a level of detail allowing the community to identify similarities between applications and the use of shared components.

Actionable items

- Develop a community engagement approach that facilitates attendance by digital research technical professionals (dRTPs) with a working knowledge of the technologies and services used to deliver projects
- Collectively design and release a standard “use case” template to increase the likelihood that funded projects using large-scale compute disseminate their experiences (positive and negative) within the community and beyond
- Task CCP-AHC staff resource with collecting and preparing use cases, via interview or other techniques, with a view to publication within and beyond the project

Technical requirements for AHC users of large-scale compute

Nevertheless, technical requirements were occasionally discussed. The technical requirements for data and compute infrastructures are different for AHC applications than for traditional HPC users. Participants described limitations of ill-fitting solutions. One project made heavy use of commercial cloud provision (e.g. Amazon Web Services, Microsoft Azure, Google Cloud Platform) to execute various computationally intensive tasks over the lifetime of a multi-year project.

This infrastructure choice was largely one of convenience, since the host institution had already procured institution-wide access to this resource. This has caused issues for the reproducibility of computational pipelines developed during this project, now that institutional support has ended. “Bureaucracy” is also cited as a hurdle to be overcome in securing access to large-scale compute for AHC research and innovation.

More generally, the value of avoiding supplier or vendor lock-in was noted, both at the level of the project and the DRI (e.g. Tier 2 HPCs). Software and hardware are affected. A trade-off of time taken to rewrite code to avoid such dependencies was commented on, as was the prohibitive cost of data egress that would be required to migrate the extensive collections of one institution from a given cloud provider.

Notebook-style programming environments are a commonplace in AHC software development, particularly in early stages and in training contexts. With the increasing interest in GPU-intensive applications (AI and deep learning) and the scant availability of GPU resource under interactive or urgent computing interfaces, this has led to the use of "free tier" compute offerings from commercial hyperscalers (e.g. Google Colab).

Though the variety of applications used is significant, there are shared challenges with data. For instance, some struggled with moving large numbers of relatively small files - of a variety of formats with their own rich and often structured metadata - into certain HPC systems. This is a data ingress scenario that not all infrastructures are equally prepared to support, since many services have been designed to work with large, homogenous scientific datasets. Relatively fundamental implementation decisions, such as specific firewall rules and information security practices, can hinder computationally intensive AHC research.

Relatedly, it can happen that scalable data processing applications are developed by AHC research teams using tools that were not designed to run in "academic" shared-computing environments. This can happen when AHC researchers are not aware of the existence of such infrastructures.

Actionable items

- Ensure that "lessons learned" describing project shortcomings are accepted and values as contributions to the community as much as success stories
- Develop guidelines for selecting commercial cloud and software solutions which may not be well-suited to existing and future state-funded large-scale compute
- Clearly identify opportunities within CCP-AHC where cloud providers may be appropriate, noting the community's strong commitments to open research and concerns about the public digital good and value for money

Discoverability of DRI

Funders have a responsibility to ensure that the existence of DRI with a significant funder stake is known to beneficiaries of research funding. This is particularly important for AHRC-supported DRI, since that funder is a relatively late entrant in the arena. A "checklist" of compute DRI for AHC research was envisaged. Any such register should be

supported by advocacy on the part of service owners. The notion of including the use of UK-funded DRI as a funding criterion was mentioned.

The uneven geographical coverage of large-scale compute in the UK was raised, perceptions of which hinder access to resource. State funders (e.g. NSF) who support schemes that broker access to infrastructure and help to identify the most appropriate compute resource (e.g. ACCESS-CI) were positively mentioned. Other DRI initiatives related to federated access to compute seem relevant to this interest.

Some participants were unfamiliar with flagship UKRI investments (e.g. the AI Research Resource) and the routes to access them (at the time, the "expression of interest" portal). This suggests much more needs to be done to route information and opportunities through CCP-AHC, via high-influence members, and onward to wider stakeholder groups. The importance of engaging with calls for compute resource was mentioned, since – as with many shared services – such calls are not only used to allocate resource but also to measure demand. Where appropriate, this can eventually establish the need for new DRI services.

Actionable items

- Produce a register of DRI relevant to HPC/large-scale compute to arts, humanities, and culture research
- Elicit requirements and measure demand for a compute brokerage process to help widen access to the most appropriate form of DRI for a given project at a given stage of maturity

What AHC users want…from artificial intelligence?

Participants highlighted many current and potential applications of artificial intelligence (AI) to the humanities, much of which can be supported by HPC and large-scale compute. Speech-to-text, optical character recognition (OCR), document layout analysis (DLA) and related computationally intensive workflows consistently emerge as recurring applications. Each of these tasks are now well established within the relevant computational literature. However, the performance of established solutions to these tasks on AHC data (especially low-resource materials) is frequently inadequate for responsible downstream use by researchers, at least not without manual intervention.

This is unsurprising, as many AHC researchers work closely with texts (in a general sense) - be they already digitized as machine-readable text, as audio, or as (scanned) documents originally intended for print. The existence of open-source codes and trained models that implement these applications, as well as their strong potential to accelerate standard

workflows across many fields, strongly suggest that CCP-AHC focus on these applications.

Each might be thought of as relating to one or more “scholarly primitives” (due Unsworth) in the operations of AHC research. Another, similar, perspective suggests a conceptual organisation centring common operations in data-intensive computational methods (e.g. derived corpora, annotation, [model] training, fine-tuning, evaluation). Yet more generally still, the opportunity for large-scale compute for AHC research was summarised as “getting unstructured data into structured data” (using shared or large-scale compute) and taking the outputs to a workstation for analysis and/or interpretation.

The distinction between “software” and “tools” was often drawn. The inference drawn is that tools are research software that has attained greater stability as computational science artefacts and could be made available as a service.

There is no national infrastructure providing access to large-scale compute services by AHC researchers. In fact, there is no national shared compute of any kind. The unmet demand even for simple services (such as minimally interactive web applications using standard technology stacks) arose directly and indirectly in several conversations. It remains to be seen what proportion of CCP-AHC resource should be used to develop and maintain such services that do not represent the deployment of cutting-edge (or, at least, recent) research software under active develop by the community members themselves. Complementing the possibility of picking “low-hanging fruit”, the importance of high-risk (and, ideally, high-reward) projects was also noted.

Actionable items

- Develop pilot service(s) for narrow AI applications for which and there is an established demand, using infrastructures available to CCP-AHC community members (e.g. automatic speech recognition (ASR), optical character recognition (OCR), handwritten text recognition (HTR))
- Ensure that any access to shared compute environments, tools, or hosted applications does not unfairly exclude users outside of the HEI landscape (who may not be part of e.g. existing identity and access management (IAM) federations)

The scope of the CCP, its relationship to HPC and the disciplines

The relevance of high-performance computing (HPC) to AHC research (and *vice versa*) was repeatedly questioned. When this was the case, HPC was understood to refer to large-scale simulations and (traditional) scientific computing applications. This is not surprising given the origins of CCP-AHC within a programme with a strong history of supporting these areas.

This suggests that changes in HPC and large-scale compute provision, which have seen compute services broaden their set of supported applications and that are already underway, are not well understood by the community. These contributions suggest that an alternative to "HPC" as a shorthand for the DRI in question should be explored (e.g. large-scale compute, AI machines etc.) when engaging this community. It is accepted that this is an ongoing discussion in other computational science communities. The relevance of "small-scale" compute to the project's aims was also raised. The question was posed: is this in scope for CCP-AHC, and how can researchers who do not (yet) use DRI be included? Related concerns about the environmental impact of the initiative, and of the project of large-scale compute in general, were also raised. AHC researchers are rightly sensitive to the negative effects of the global expansion of large-scale compute (which presents commercial as "AI"). The contribution of CCP-AHC to these debates, ideally while leading by example (e.g. by reporting and contextualising resource consumption, or by not using resource at all), is an open question but one whose importance is shared by many attendees.

Actionable items

- Consult community members on their understanding of and preference for terminology with a view to maximizing recognition and engagement
- Clearly establish the proportion of the community interest and effort with respect to artificial intelligence (AI) applications
- Develop a community approach to responsible large-scale compute, encompassing environmental/climate sustainability as well as broader ethical concerns with the explicit aim of embedding best practice in community activities.

Training

Training repeatedly emerged as mechanism for fostering several desired community behaviours. As with many other areas of research. It is notable that there are AHRC/UKRI DRI investments in skills development with relevance to CCP-AHC (e.g. the most recent rounds of funding supporting DISKAH and Data/Cultures). Some of these are focused on increasing access to large-scale compute, but this is the exception rather than the rule.

Established CCPs in other research areas frequently fund the delivery of and attendance at specialist software trainings by community members. This can range from a one-day workshop to an annual summer school, usually but not always focused on a specific application or toolkit. Participants who may not otherwise have funds to attend typically have travel and subsistence supported by the CCP. Hosting and delivery costs are often co-funded.

The prominence of training in CCP-AHC activities remains to be seen. Certainly, existing documentation of large-scale compute presupposes background knowledge not commonly possessed by AHC researchers, suggesting a need for wider training at the introductory level, with some preference for one-to-one or "clinic"-style support. A non-exhaustive list of potential topics for more advanced training may include: an overview of the DRI landscape as it relates to AHC research, introduction to benchmarking, introduction to containerisation (for GPU-intensive applications), storage best practices for AI/ML, preparing effective resource allocation bids and/or use cases.

Actionable items

- Identify opportunities within the remainder of the CCP-AHC project to support training or knowledge exchange activities, not envisaged by other ongoing DRI investments

The role of dRTPs in CCP-AHC

Digital research technical professionals (dRTPs), such as research software engineers (RSEs), research infrastructure engineers, data scientists, and data stewards play a crucial role in delivering computationally intensive results in many areas, and AHC research and innovation there is no exception. Interdisciplinary and cross-institutional teams empower AHC applications of large-scale compute; this entails a need to find ways to support teams often consisting of contributors with variable capacity to engage in the community (e.g. due to other daily responsibilities, organisational culture). It is also recognised that dRTPs may be engaged in research in different ways: as service partners, as collaborators, or as co-leads or leads. Each approach can be warranted, depending on context and the size of the dRTP capability. More clarity on this is needed for CCP-AHC as the initiative becomes more established, taking due consideration for the professional backgrounds and interests of the community members. dRTP contributions (either costed to the project or "donated" in kind by community members) are most effective when the nature of this relationship is clearer, and where there are clear opportunities to develop autonomy and/or leadership within the projects and services they deliver.

Actionable items

- Ensure that dRTPs with a track record of supporting access to DRI by AHC researchers are involved in community decision making
- Develop a set of working principles for dRTP effort within CCP-AHC, distinguishing clearly between different ways of working and their associated expectations on all parties, including the Delivery Team, CoSeC, and the wider CCP-AHC community

Routes to change

CCP-AHC should look to shape the direction of DRI in many ways but will not have credibility unless it offers solutions to problems that AHC researchers and innovators genuinely need solved. If there is an obvious benefit to the research ecosystem – e.g. digitisation may reduce the cost of travel to repeatedly consult the same source – it should be stated as such. A clear roadmap is expected, though there was a desire to collect and supply the evidence needed to make it credible and successful. A landscape analysis, looking at other infrastructure providers – including those offered to other disciplines – is advised.

It is also important to secure the visibility of the community within the wider DRI landscape, such as attendance at meetings that AHC researchers do not routinely attend. The practical benefits of building partnerships with key organisations already aligned to the goals of CCP-AHC (e.g. among many, the Software Sustainability Institute) were stressed. The need to seed the imagination of researchers as to what is possible with the support of large-scale compute and other DRI was underlined. This can be supported by training and the prompt communication of use cases (e.g. via the website). The importance of concrete or tangible outcomes, such as a usable demonstrator application or service, was repeatedly stressed.

Actionable items

- Produce a “case studies” resource for the CCP-AHC website, and support the production and publication of use cases (within the next 6 months), to illustrate the value of large-scale compute and other DRI to AHC research

The CCP model

The Collaborative Computational Project (CCP) model was unfamiliar to most participants. Attendees wished to know much more about the relationship between STFC, CoSeC, the existing CCPs and other (digital) research infrastructures. The focus of the proposed CCP on HPC was questioned. While the CCP's link to large-scale compute in other areas can go without saying, this is not the case for the AHC community. There was also uncertainty about the scope of CCP-AHC's vision for infrastructure provision in the long term. This is closely linked to the question of what CCP-AHC and the broader CoSeC programme bring to the community. The possibility of appetite for more than one CCP to serve AHC researchers was raised.

Actionable items

- Clearly describe the proposed working relationship between the community, CCP-AHC leadership, STFC, and staff employed by the project

- Identify the specific expertise and competencies within the RSE resource (Durham) and computational scientist/data scientist resource (STFC)

Governance and engagement

The role of project leadership in determining the direction of the community was also raised. A collaborative approach during the set-up of CCP-AHC will ensure broader engagement, but a more directive approach may also have a role. Broader engagement can be achieved by greater community involvement in management and delivery, as well as a clearly identifiable community web presence. The allocation of technical staff time during the project must be clearly prioritised and open. A community manager could be well used to broker access to RSE and data scientist resource within the project.

Cultivating advocates for the community will widen engagement. Other DRI projects have used fellowship and/or “champion” programmes to increase their recognition. The CCP-AHC community leads must be open to input and criticism. External scrutiny (e.g. from the funder) is also important. Engagement with ongoing allied DRI projects both nationally and internationally is equally important. Some community members have prior experience using and, in some cases, supporting DRI in other nations (e.g. European Union member states, Canada, the United States); it is essential that lessons learned from these experiences inform the direction of CCP-AHC.

Actionable items

- Grow the CCP-AHC Delivery Team into a Working Group, starting by introducing an open community forum with monthly online meetings
- Design a process for the delegation of work packages to community members
- Revisit composition of advisory board to include funder representative
- Ensure that CCP-AHC committees, advisory boards, etc. reflect a diversity of career stages, as well as other relevant characteristics (including but not limited to those covered by equalities legislation).

Funding acknowledgement

This work is supported from January 2025 to December 2026 by the Science and Technology Facilities Council (STFC) on behalf of UK Research and Innovation (UKRI), under the "Collaborative Computational Communities: towards new CCPs” opportunity.

Appendix 1: Description of CCP-AHC Town Hall 2025

Overview of the event

The first CCP-AHC Town Hall 2025 was held at Delta Hotels by Marriott Durham Royal County, Durham (UK) on Thursday 22 May 2025 from 9:00 a.m. to 5:00 p.m. In attendance were 25 people from 20+ institutions, with 13 more online.

Welcome, introductions, and presentations

Following a half-hour introduction to the project from the PL Eamonn Bell (Durham University) "The story so far: The (evergreen) potential for large-scale compute in Arts, Humanities, and Culture Research", four five-minute lighting talks were given by:

- Karina Rodriguez Echavarria (University of Brighton), "The Digital Skills in Arts and Humanities (DISKAH) Network"
- Jeyan Thiyagalingam (STFC), "CCPs, CoSeC, and the STFC Scientific Computing Department"
- Martin Wynne (University of Oxford), "The Oxford Text Archive (OTA)"
- Phil Hasnip (University of York), "UKRI Living Benchmarks"

The presentations, which were also available for online attendees, had the aim of introducing the Collaborative Computational Project (CCP) model, the policy background to the CCP-AHC project, and a selection of relevant digital research infrastructure (DRI) projects.

Community consultation and requirements gathering

The focus of the remainder of the meeting, for in-person attendees only, was on gathering input into past, current and future usage of HPC and advanced computing resources by arts, humanities, and culture researchers and innovators. Questions/prompts discussed by attendees at each of five tables appear as Appendix 2 in this report.

CCP-AHC Advisory Group members served as table hosts and participants moved *en bloc* from table to table. During a debriefing meeting the following morning, they served as rapporteurs, providing further context to observations collected during the Town Hall.

Appendix 2: Discussion questions used in working sessions

Discovering HPC & AI codes, pipelines, workflows, and infrastructures

- What codes, pipelines, and workflows are being used today with HPC & AI infrastructures (and other DRI supporting large-scale compute) to produce computationally intensive arts, humanities, and culture (AH&C) research?
- How should these projects be identified, evaluated, and prioritised to promote best research software practice while covering a diversity of research domains and application types?
- What are the principal critical dependencies that AH&C codes, pipelines, and workflows require (incl. data, software, libraries, hardware, and people)?
- Which infrastructures are the most successful at overcoming the bottlenecks and common challenges faced by AH&C researchers and innovators when accessing large-scale compute?

Emerging and future users of large-scale compute infrastructure and their use cases

- What is the typical profile of the current and future AH&C users of large-scale compute, including HPC, AI research resource, and other advanced computing infrastructures?
- How can community members be supported in the development and preparation of use cases/case studies that are useful to evidence impact and the value of large-scale compute infrastructures to AH&C researchers and innovators?
- Do existing benchmarks (e.g. MLCommons benchmarks) capture typical AH&C uses, or do new benchmarks need to be defined and designed for AH&C users?
- What lessons can be learned from other communities (including the other CCPs) about who have tackled similar challenges about growing the adoption of large-scale compute in the past?

Driving and measuring positive change within the community

- What should be the key elements of the CCP-AHC roadmap through to the end of the scoping period (end 2026) and beyond (2027-)?
- How should computational and scientific resource within CCP-AHC - including the workplan for computational scientist (1.0 FTE) and RSE resource (0.6 FTE) - be prioritised during the life of the scoping project?
- Where are the key points of influence that the community can advocate for HPC, AI resource, and other DRI (e.g. in procurement, design of benchmarks, in cross-council DRI activities)?

- How should CCP-AHC ensure that community members are engaged in its decision-making and planning, with a particular emphasis on representing HEIs and RPOs of various sizes, research culture, and the project's relation to UKRI priorities (e.g. Net Zero)?