

# HiTHeR (High Throughput Humanities e-Research)

## 1. What research is being done by the researcher(s)

The Nineteenth Century Serials Edition (NCSE) corpus contains circa 430,000 articles that originally appeared in roughly 3,500 issues of six 19th Century periodicals. Published over a span of 84 years, materials within the corpus exist in numbered editions, and include supplements, wrapper materials and visual elements. Currently, the corpus is explored by means of a keyword classification, derived by a combination of manual and automated techniques. A key challenge in creating a digital system for managing such a corpus is to develop appropriate and innovative tools that will assist scholars in finding materials that support their research, while at the same time stimulating and enabling innovative approaches to the material. One goal would be to create a “semantic view” that would allow users of the resource to find information more intuitively.

However, the advanced automated methods that could help to create such a semantic view require processing power that is currently not available to CCH researchers. Gerhard Brey has implemented a simple document similarity index that would allow journals of similar contents to be represented next to each other. The program used the [lingpipe](http://alias-i.com/lingpipe/) (<http://alias-i.com/lingpipe/>) software to calculate similarity measures (specifically, the TF/IDF similarity measure on character n-grams) for articles within the corpus. A test using 1,350 articles, requiring a total of  $910,575 (n * (n-1) / 2)$  separate comparisons, was executed on a Mac Mini, which took 2 days to process 270 documents, that is to perform  $270 * 1,349 = 364,230$  comparisons. Assuming the test set was representative, a complete set of comparisons for the corpus would take more than 1,000 years!

## 2. What are the issues that ENGAGE could address?

We see ENGAGE as an opportunity to start building the e-infrastructure required to support advanced research in the (digital) humanities. One driver is the NCSE project, but, in parallel with this, King's is planning to set up a new e-infrastructure for its researchers. CeRch is tasked with doing this work, and in particular Mark Hedges, Tobias Blanke and Richard Palmer are taking the lead in setting up a Campus Grid which will provide King's researchers with the processing and data resources as well as a link to the National Grid Service for additional resources. Because of its connections with humanities computing, CeRch has chosen to demonstrate this approach in the first instance with a humanities-based project. ENGAGE would provide us with the means to realize the first steps towards such an infrastructure and to facilitate further engagement of humanities academics with e-infrastructures.

In addition, support from ENGAGE would help us to develop a software solution for a genuine problem in humanities research, which would have a lifespan beyond the initial project and could provide a plug-in for future similar digitisation projects. We plan to come up with a prototype toolkit that can carry out document similarity processing and other advanced text analysis, which can be run on either campus or national computation grid resources.

## 3. Future benefit to user community

1. The digital resources available to humanities researchers are increasing rapidly in size, as result of digitisation programmes and the born-digital nature of more recent research outputs. Researchers in the (digital) humanities have been looking for more effective means of exploiting these resources, adding value to them and generating new knowledge from them. The use of High Throughput Computing (HTC) provides one such means, and the project will demonstrate to this community both the utility of

such an approach and an effective and affordable means of realising it, by using grid technology such as Condor. This could result in the increased uptake of HTC in the humanities.

2. The project will provide humanities researchers with well documented ways of linking to the NGS and will enable researchers to use its resources in their projects.
3. The project will develop a prototype tool addressing a real research problem in the humanities, specifically in the field of textual analysis of large corpora.