**BURROWS — Enriching**

<0 images>

**Enriching the HuNI Virtual Laboratory with Content from the Trove Digitized Newspapers Corpus**

Burrows, T. N. and Davidson, A.

HuNI (Humanities Networked Infrastructure) is an innovative virtual laboratory that takes a new approach to the problems of integrating and linking data in the humanities and creative arts. It aggregates data from thirty Australian datasets, using a Data Model based on six core entity types. HuNI users are able to construct their own links between the individual entities in the HuNI service. This ‘social linking’ function is accompanied by the ability to create and annotate collections of entities, as well as to import and export data.

The initial HuNI project, funded by an Australian government grant under the NeCTAR programme (2012–2014), was completed with the launch of the HuNI service in October 2014. More than 730,000 entity records were included in this initial version of HuNI. The research community served by HuNI spans the humanities and creative arts in Australia, and includes researchers across literature, cinema and media studies, history, linguistics, Aboriginal studies, and art (Verhoeven and Burrows, 2014).

This paper will report on a project to extend the scope and content of HuNI, using the corpus of digitized newspapers assembled by the National Library of Australia and made available as part of its Trove service (http://trove.nla.gov.au/newspaper). This Trove corpus contains 13.7 million pages and 133.8 million articles, from more than 700 newspapers published between the earlier 19th century and the mid-1950s (Sherratt, 2014). An extensive crowdsourcing project has been running for several years, aimed at correcting the OCR files for the corpus.

The project, being carried out during 2014 and 2015, has two main objectives:

1. Testing and evaluating solutions and methods for carrying out Named Entity Recognition against the OCR files in the Trove digitized newspapers corpus.

2. Exporting the information about entities identified from Trove so that they can be incorporated into the data aggregate that forms the basis for the HuNI service.

This work has the potential to expand HuNI’s content significantly. It extends HuNI’s data, for the first time, to entities derived from full-text digital resources. It is enabling the HuNI team to evaluate the usefulness of the Trove corpus for this purpose and is serving as a pilot for the development of an ongoing production service of this kind.

The project is being carried out with the assistance of a grant for the Microsoft Azure for Research programme. The project is using the Azure cloud service to store a copy of the OCR files for the Trove newspaper corpus. It is also deploying the GATE software on Azure in order to test its capabilities against the Trove newspaper corpus. GATE is a full-lifecycle open-source solution for text processing, developed at the University of Sheffield and widely used around the world (Cunningham et al., 2011). GATE is already available as a packaged Virtual Machine for cloud deployment, and the project is extending that cloud deployment to Azure.

In the paper, we will report on the initial results derived from the use of GATE to identify named entities in the Trove newspapers corpus. This will include a discussion of the issues encountered, an evaluation of the results obtained, and an assessment of the effectiveness of the overall process. We will also discuss the methods being developed for packaging up these results for ingest into HuNI. This will include demonstrating and evaluating the pipeline for this process and showing how entity records added to HuNI can be used to link back to their occurrence in the Trove corpus.

The paper will also discuss the remaining work to be carried out during this project. The final stage of the project will involve deploying Azure’s own Machine Learning tools against the Trove corpus. It will then be possible to compare the results obtained from Azure Machine Learning with those obtained through the use of GATE. Progress towards the deployment of a production service for adding Trove entities to HuNI will also be discussed.

**References**

**Cunningham, H., et al.** (2011). *Text Processing with GATE (Version 6).* GATE, Sheffield.

**Sherratt, T.** (2014). Digitised Newspapers and the Varieties of Value. Presented at the Europeana Newspapers workshop, *Newspapers in Europe and the Digital Agenda for Europe*, London, http://www.slideshare.net/wragge/digitised-newspapers-and-the-varieties-of-value.

**Verhoeven, D. and Burrows, T.** (2014). Digital Scholarship in the Humanities and Creative Arts: The HuNI Virtual Laboratory. *EDUCAUSE Review Online,* May/June.