A New Taste of HuNI: Redeveloping the Humanities Networked Infrastructure

Toby Burrows 1, Deb Verhoeven 2

1University of Western Australia, Nedlands, Australia, toby.burrows@uwa.edu.au
2Deakin University, Melbourne, Australia, deb.verhoeven@deakin.edu.au

DEVELOPING HUNI

This paper will discuss and demonstrate the newly redeveloped version of the HuNI Virtual Laboratory. HuNI (the Humanities Networked Infrastructure) is one of the Virtual Laboratories developed with funding from the Australian Government’s NeCTAR (National e-Research Collaboration Tools and Resources) programme. HuNI responds to several critical problems identified by humanities researchers, which are not well served by existing eResearch initiatives. These can be summarised as:

• The need for a discovery platform capable of responding to semantic search queries across different humanities disciplines;
• The need for researchers to make connections at the level of entity records and to describe these connections in their own terms;
• The need for data to be linked in complex and nuanced ways, including the ability to assert that two records are not linked;
• The need to emphasise interpretation as a core capability in making and evaluating connections;
• The ability to generate non-logical or serendipitous connections for users as they navigate the platform.

The first version of HuNI was launched in October 2014 and met these expectations. [1] HuNI’s initial content consisted of data from 28 Australian datasets in the humanities and creative arts. Incoming records were mapped to a Data Model which consisted of six core entity types: Persons, Organisations, Places, Concepts, Events and Works. The key goal was to identify all the individual entities contained in this broad range of data, and to retain the disciplinary context in which they had been placed. More than 730,000 entities were created in this initial phase. No attempt was made to remove apparently duplicated entities from different data sources, nor was there any attempt to create automatic links between entities as part of the data harvesting and indexing process.

The process of creating links was left to users of HuNI, who are able to take two entities and assert a relationship between them. Multiple links between entities can be made by different users, reflecting different interpretations and disciplinary approaches to the data. It is also possible to assert that a relationship does not exist between two entities. Users can group entities into collections, either public or private. These collections serve as de-facto classifications or categorizations of entities. The resulting set of links and collections can be visualized as a network graph, in which the nodes (entities) can serve as starting-points for exploring the wider web of connections across the HuNI aggregate of heterogeneous data.

A subsequent NeCTAR-funded project during 2015 enabled the addition of content from the Trove collection of digitized Australian newspapers. HuNI and the Alveo virtual laboratory collaborated to apply Named Entity Recognition (NER) to the Trove newspaper corpus. The entities identified were exported from Alveo to HuNI and added into the HuNI aggregate. This process added 16.8 million entities to HuNI, consisting of Trove newspaper articles mapped to the “Work” category in HuNI, together with HuNI “Person” entities mentioned in the text of these articles. As of June 2016, HuNI contained nearly 17.5 million entities.
HuNI’s original technical architecture was built using Open Source technologies, with three main components:

- A Solr Document Index which contained the harvested and indexed incoming data, and exposed a search API, allowing matching entities to be returned;
- A database which stored user profiles, links between entities, collection lists, and associated metadata;
- An Nginx HTTP server and a RESTful API service which allowed users to work with the HuNI data through a Web browser.

Researchers have been active in using HuNI to create collections and to add links. User collections have been organized around particular people and places, but have also focused on specific themes and topics such as “the secular sublime”, beauty, and former child migrants. Researchers can also create personal databases externally and import them into HuNI. The Heurist software, developed at the University of Sydney, now includes a HuNI template which can be harvested by HuNI.

IMPROVING HUNI

Since its launch, several improvements have been identified as a way of enhancing the original vision for HuNI. During 2016, a new version of HuNI is being developed with further funding from NeCTAR. The main aim of this redevelopment is a significant increase in the connectedness of HuNI – both internally and externally. Internal linking is being enhanced by revising the ingest pipeline for one of the original datasets: Design and Art Australia Online (DAAO). Links already present between entities in the DAAO records are being imported into HuNI, where they appear as linking relationships in the HuNI graph. This can be done without distorting the disciplinary perspective inherent in the original data.

For external linkages, an ingest pipeline is being developed for databases created with Omeka – widely-used Open Source software for managing digital object collections. Researchers can create instances of Omeka through the Launchpod hosted by Intersect as a node of Research Data Services, and will be able to choose a template based on the HuNI Data Model to structure their metadata, which can then be exported to HuNI as core entities.

HuNI is also publishing its entities in a form which can be shared in the Linked Open Data cloud. HuNI already had URLs for individual entities, which were automatically created as part of the ingest process and combined the data source, the entity type, and a unique reference number. For exposing these to the Linked Open Data cloud, the data are being exported in JSON format, converted into JSON-LD, and made available for harvesting and linking.

HuNI is also migrating from its customized technical architecture to a new platform built around the property graph database software Neo4j. Neo4j is a world leader among graph databases, and was recently used by the International Consortium of Investigative Journalists to present the “Panama Papers” data for exploration by the public and journalists. [2] Neo4j shares HuNI’s flexibility about data models; it has no schema as such, and a data model can be developed and defined incrementally. Neo4j also emphasizes links and patterns of relationships between nodes (or entities, in HuNI’s case). Migration to Neo4j is greatly enhancing HuNI’s functionality and sustainability, including – for the first time – the ability to search on the links between entities.

REFERENCES

2. Cabra, M. How the ICIJ used Neo4j to unravel the Panama Papers. DZone / Database Zone, 16 May 2016. https://dzone.com/articles/how-the-icij-used-neo4j-to-unravel-the-panama-pape
ABOUT THE AUTHORS

Toby Burrows is the Manager of the eResearch Support Unit at the University of Western Australia, and an Honorary Research Fellow in the School of Humanities. Between May 2014 and May 2016, he was a Marie Curie International Incoming Fellow in the Department of Digital Humanities at King’s College London. He has also held visiting fellowships at the Vrije Universiteit Amsterdam, Churchill College Cambridge and University College London. His research interests are in digital humanities and medieval manuscript studies. He is one of the Product Owners for the HuNI Virtual Laboratory.

Deb Verhoeven is Professor and Chair of Media and Communication at Deakin University. Amongst her many accolades she was named Australia’s Most Innovative Academic in 2013. From 2012-2014 she was Deputy Director of the Centre for Memory, Imagination and Invention (CMII). Until 2011 she held the role of Director of the AFI Research Collection at RMIT University. A writer, broadcaster, film critic and commentator, Verhoeven is the author of more than 100 journal articles and book chapters. Professor Verhoeven is one of Australia’s leading proponents of Digital Humanities research and served as Program Chair for the international Digital Humanities 2015 conference. She is the Project Director for the HuNI Virtual Laboratory.