Linking and sharing data in the humanities and creative arts: building the HuNI Virtual Laboratory

Toby Burrows
Manager, eResearch Support
University of Western Australia
toby.burrows@uwa.edu.au

Deb Verhoeven
Professor of Media and Communication
Deakin University
deb.verhoeven@deakin.edu.au

Abstract:
The Humanities Networked Infrastructure (HuNI) is one of the national Virtual Laboratories that are being developed as part of the Australian government's National e-Research Collaboration Tools and Resources (NeCTAR) programme. This paper examines the methodologies and technical architecture being deployed by HuNI to link and share Australian data in the humanities and creative arts.

Introduction

The Humanities Networked Infrastructure (HuNI) is one of the national "Virtual Laboratories" which are being developed as part of the Australian government's NeCTAR (National e-Research Collaboration Tools and Resources) programme. NeCTAR aims to integrate existing capabilities (tools, data and resources), support data-centred workflows, and build virtual communities to address well-defined research problems. This is a particularly challenging set of technical requirements and problems for the humanities and creative arts, which cover an extensive range of different disciplines and are characterised by complex and highly heterogeneous collections of data (Burrows 2011). User requirements and use cases are also very varied and complex.

HuNI is being developed by a consortium of thirteen Australian institutions, led by Deakin University. It is bringing together data from thirty different Australian datasets, which have been developed by academic research groups and collecting institutions (libraries, archives, museums and galleries) across a range of disciplines in the humanities and creative arts. These datasets include Design and Art Australia Online, the Australian Dictionary of Biography, AustLit, AusStage, the Dictionary of Sydney, and the PARADISEC linguistics archive (see Appendix 1 for a full list). These datasets contain more than 2 million authoritative records, capturing the people, places, objects and events that make up the country’s rich heritage.

HuNI is ingesting data from all these different Australian data providers, mapping the data to an overall data model, and converting the data for inclusion in an aggregated store. HuNI is also assembling and adapting software tools for using and working with the aggregated HuNI data. Two existing tools are being extended to interface with HuNI: LORE (Literature Object Reuse and Exchange), developed at the University of Queensland as part of the AustLit service, and Heurist, developed at the University of Sydney.

Fundamental to HuNI’s architecture was the decision to build a central aggregate, rather than designing the Virtual Laboratory functionality (e.g., federated searching, browsing and annotating) to work with the individual data sets in a distributed way. A central aggregate adds significant value to the disparate data sources by maximising the links between them, and by putting them into a much broader interdisciplinary context. It also enables researchers to work with data from a variety of different sources in a much more effective way and on a much larger interdisciplinary scale.

HuNI is part of the rapidly growing global Digital Humanities initiative, which is producing many innovative applications and services aimed at expanding the use of digital technologies in humanities research. In Australia, this saw the formation of the Australasian Association for Digital Humanities (aaDH) in March 2011, its formal incorporation in March 2012 and its acceptance into the international Alliance of Digital Humanities Organisations (ADHO). There is a significant overlap in membership between the AADH Executive Committee and the HuNI Steering Committee.

Data Integration

The thirty Australian humanities data sets that are being incorporated into HuNI are managed and maintained by a variety of different institutions, including various universities and government agencies like the Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS). Data providers also include several national consortia in specific humanities disciplines, including AustLit (managed by the University of Queensland [Kilner 2009]) and AusStage (managed by the Flinders University of South Australia [Bollen et al. 2010]). The data from some of these services conform to standard metadata schemas like Dublin Core, EAC-CPF (Encoded Archival Context - Corporate bodies, Persons and Families) and MARC-XML (Machine Readable Cataloguing – eXtensible Markup Language). Many of the other data services are in a customised format specific to that dataset.

Appendix 1 lists the various datasets, their metadata schema, and their custodian or owner. It also shows the type of data contained in each dataset.

Setting up the HuNI harvesting process has required development iterations across two levels of technology deployment. The first relates to the technology needed for data providers at the partner sites to publish their data in XML (eXtensible Markup Language) format. HuNI provides three options for supplying: jOAI and OAIcat for those who are exposing their data via OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) and a custom-built non-OAI solution that requires very little work to integrate at a provider's site. The second aspect of the deployment relates to technology for harvesting updated content from the partner XML data feeds and transforming the data into forms suitable for ingestion into a Solr search server.

Each data set was evaluated against a set of criteria to determine its readiness for ingest:

- the capacity of the data source to connect to the HuNI ingest system (i.e., its communication protocols and technical infrastructure);
- its capacity to expose and export data in OAI-PMH;
- the availability and extent of documentation of the data models used to develop and maintain the dataset;
- the extent to which the data provider adheres to or uses standard schemas or information standards in its data structures and technical infrastructure;
- the availability of technical and data modelling advice and support from the data provider;
- the readiness of the data provider for engagement with the HuNI Project Team, their availability for liaison and problem resolution;
- the data provider's capacity for supporting data analysis and mapping; and,
- the alignment of the data with the core entities in the HuNI data model.

The need for data cleaning emerged as a significant result of the analysis of the HuNI datasets. A number of common challenges with input, processing, translation and aggregation of data (particularly from diverse sources) were identified during the HuNI ingestion process. Rahm and Do (2000:3) offer a useful visual summary of data quality problems, and of ways to improve data quality and support reuse of data.

Tools like Google Refine or Gadget can be used by data custodians to examine their data and identify any input, formatting or notation problems that need be resolved (Van Hooland and Verborgh 2012). This process enables the data to be refined at source, which is the most sustainable approach. HuNI staff worked initially with Gadget to inspect the data and to apply various cleaning techniques.

Figure 1 presents some common data cleaning issues identified during the HuNI ingestion process.

Issue	Data Types	Examples	Options
Variation in notation	Personal names; Roles	Cardell Oliver & Cardell-Oliver; ARTIST & artist	Data cleaning and conversion
Semantic equivalence	Personal names; Concepts	Smith M & Mary Elizabeth Smith; forename & first name	Linking or convergence
Semantic non-equivalence	Place of birth; Date of birth	Bendigo & Melbourne; 17 May 1914 & 17 May 1915	Preserve and flag for review

Figure 1: Examples of Data Cleaning Issues in HuNI

The integration of partner XML records into the HuNI data aggregate was dependent on successfully mapping them to a core HuNI data model. Defining this core model has been an iterative process, and has involved testing several different approaches. A range of cultural heritage ontologies (Hyvönen 2012) was initially examined as a starting-point for building a core ontology framework. These included CIDOC-CRM (Comité International pour la Documentation – Conceptual Reference Model), FOAF (Friend of a Friend), PROV-O (Provenance Ontology) and FRBR-OO (Functional Requirements for Bibliographic Records – Object Oriented). However, it became clear that there were significant technical and conceptual difficulties with this kind of approach. In part, these challenges arise from broad disciplinary shifts within the humanities, away from a traditional focus on measuring the value and meaning of cultural artefacts to recognising the import of cultural flows and the dynamic nature of cultural infrastructure (itself understood as a creative process and catalyst of social and environmental amenity).

An alternative approach involved mapping the incoming harvested XML records to a limited set of thirteen core entities. As of September 2013, a prototype Solr index was populated with 379,236 records covering 13 class entity types from 24 partner data sources. The class entity types were as follows (with the number of records in brackets):

- Act (82)
- Artefact (109,065)
- Bibliography (3,373)

- Collection (187)
- Concept (4,448)
- Event (68,079)
- Film (10,042)
- Organisation (21,426)
- Person (149,543)
- Place (519)
- Production (3,703)
- Venue (8,763)
- Videos (6)

These entity types are continuing to be refined as part of the work on analysing the user stories in the product backlog. The full HuNI data model was finalised in December 2013. It reduces the number of core entities to six: Person, Organisation, Event, Work, Place, and Concept. The HuNI data aggregate is being rebuilt by mapping the imported data to these core entities.

Deploying Tools

The HuNI Virtual Laboratory is designed to support the non-linear research methods practiced in the humanities. HuNI provides discovery tools for casual users from the wider community, but more sophisticated functionality is available to researchers who register for an account in the Virtual Laboratory. Registered researchers have their own personal workspace within HuNI. Researchers can authenticate themselves using social media logins and will be able to share their discoveries and activities through social media

Any researcher with a HuNI account can work with the HuNI data aggregate in a personalised way through a “My HuNI” interface. The following functionality will be available to researchers with HuNI accounts:

- Run a simple search across the aggregated data and browse the search results by facets based on the core entities;
- Conduct an advanced search across the aggregated data records;
- Save their search results as a private collection;
- Refine or expand a collection through additional searches of the HuNI data aggregate;
- Annotate entities within the HuNI data aggregate with assertions about the links and relationships between them (producing “socially-linked” data);
- Analyse and annotate collections with their own assertions, interpretations and commentary;
- Export collections into other digital environments for further analysis;
- Publish collections (search results and annotations) for use by other researchers;

- Share collections and annotations through social media (such as Twitter, Facebook, GooglePlus, LinkedIn).

Two existing external software tools are also being augmented to interface with HuNI. LORE was developed at the University of Queensland to enable annotation and aggregation of digital resources using a plug-in to the Firefox browser. LORE is being adapted to work with the HuNI data aggregate.

Heurist (developed and maintained at the University of Sydney) is developing database-on-demand services for HuNI that will contribute data to the aggregate. Researchers will be able to set up a database in Heurist and then publish this database to HuNI. Heurist will serve as an exemplar of a “HuNI-compliant” database creation tool, as well as a future data source provider.

User-Centred Design

A user-centred approach to the development of the Virtual Laboratory has been deployed, with the aim of ensuring that its functionality and user interface design are in line with researchers' needs and expectations. This approach is essential to ensure that the Virtual Laboratory is adopted, used and supported by the research communities it is designed to serve.

An Agile methodology has been adopted to manage the development of the Virtual Laboratory. It includes the following basic phases:

- User Story Capture and Prioritisation;
- Development Sprints: iterative 2-week sprints to complete a user story;
- Development Showcases: monthly live demonstrations for partners.

A set of 21 user stories were identified from interviews with researchers who expressed an interest in using the HuNI Virtual Laboratory. These stories cover a range of different disciplines, and include topics like “Cultural flows of cinemas”, “Mapping narrative locations”, “Australian video art history” and “Rock art research”. Each story was mapped to twenty high-level functions, ranging from “Browse, find and display” to “Visualise” and “Publish dataset”.

A Requirements Analysis document was then derived from an analysis of the 21 stories. For each user story, it contained (1) a general analysis of the story, (2) requirements for the HuNI data model (covering entities, relationships and attributes, and (3) a number of small Agile user stories (expressed as functional requirements or features for the system). The Agile stories from all 21 cases were then compiled into a single "product backlog". The stories in this product backlog were then prioritised by the Product Owner for implementation by the solution architect and technical staff.

The Agile roles were assigned within the Project Team as follows:

- Product Owner: acts as the lead user, with a deep understanding of humanities research and technical processes; prioritises the product backlog for implementation; not associated with any of the contributing data providers;
- Product Stakeholders: senior members of the Steering Committee who assist with the prioritising of the product backlog and provide sign-off for each development sprint;
- Semantic Lead: analyses the user stories and captures functional requirements; chairs the development sprint demonstrations and showcase events;
- Solution Architect (Scrum Master): translates the functional requirements into development specifications or technical stories; carries out development work and assigns development sprint tasks to technical staff as required;
- Project Manager: takes ownership of the Agile process; ensures that development remains on track and on schedule; arranges resource allocations for development sprints as required.

The audience for the fortnightly development sprint demonstrations is the Expert Data Group (EDG), which brings together a representative group of HuNI early adopters and data custodians. Members of the EDG provide written feedback after every second demonstration. The audience for the monthly development sprint showcases is much broader, and includes members of the wider interest group (including members of Virtual Laboratory teams in other disciplines) and the international Expert Advisory Group (EAG). User Interface (UI) requirements are being captured through the same community engagement channels and fed into the development process.

User Acceptance Testing is also an integral part of the HuNI project. There are two basic purposes:

- Ensuring the software is fit for purpose for researchers;
- Ensuring that the project meets the funders' expectations.

In practice, this means that the test processes and scripts need to be normalised within the NeCTAR framework and a composite document produced for NecTAR as a formal deliverable. From NeCTAR's point of view, User Acceptance Testing must address the agreed Acceptance Criteria for each of the project deliverables. The software deliverables agreed for the HuNI project cover the user tools, the discovery interface to the data aggregate, and the processes for ingesting, mapping and integrating data.

Conclusion

The HuNI Virtual Laboratory is integrating humanities data at a national level and deploying capabilities that enable researchers to work with the aggregated data. A number of significant challenges are being addressed as part of this process:

- Integrating a variety of heterogeneous data sources;
- Defining a core data model for integrating these data sources;

- Overcoming inconsistencies in the source data;
- Building a centrally aggregated HuNI platform, rather than using a distributed or federated approach;
- Scheduling and managing the complex process of data ingest, alignment and matching from the data sources;
- Defining and building the core functionality for researchers to work with the aggregated data;
- Ensuring an appropriate level of input from users into the iterative design of the Virtual Laboratory.

By demonstrating and testing an innovative new model for the design of humanities e-research infrastructure, HuNI is enabling larger-scale research questions to be pursued more effectively. It is also increasing the effectiveness of researchers' use of cultural collections, and working to ensure that research results are fed back into the management of these collections. HuNI's expected benefits can be summarised as follows:

- Humanities researchers can work with cultural datasets on a larger scale than previously possible;
- The systematic sharing of research data among humanities researchers is being encouraged and enabled;
- A higher level of cross-disciplinary and interdisciplinary research is being supported and promoted;
- Innovative research methodologies, which rely on large-scale datasets, are being enabled.

Through its use and development of innovative technologies and techniques, the HuNI project proposes some large questions, far beyond the specific queries of participating researchers: how might the opportunities presented by an unprecedented proliferation of networked data also challenge the unspoken assumptions and ordinary practices of conventional humanities research? Underlying the HuNI initiative is the recognition that cultural data is not economically, culturally or socially insular and, in order to fully explore its dimensions fully, researchers need to collaborate across disciplines, institutions and social locations (Verhoeven 2012). If we understand humanities research problems as comprising interdependent networks of institutional, social and commercial practices, then it follows that new kinds of 'evidence', and new ways of organising, accessing and presenting this evidence, are critical for our enquiries.

References

Bollen, J., Harvey, N., Holledge, J., & McGillivray, G. 2009, 'AusStage: e-research in the performing arts', *Australasian Drama Studies* vol. 54, pp. 178-194.

Burrows, T 2011, 'Sharing humanities data for e-research: conceptual and technical issues' in *Sustainable Data from Digital Research*, ed. N Thieberger, PARADISEC, Melbourne, pp. 177-192.

Humanities Networked Infrastructure (HuNI) 2013, http://huni.net.au

Hyvönen, E 2012, *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Morgan and Claypool, San Rafael, CA.

Kilner, K 2009, 'AustLit: creating a collaborative research space for Australian literary studies' in *Resourceful Reading: The New Empiricism, eResearch and Australian Literary Culture*, eds K Bode & R Dixon, Sydney University Press, Sydney, pp. 299-315.

Rahm, E & Do, HH 2000, 'Data cleaning: problems and current approaches', *IEEE Data Engineering Bulletin* vol. 23, no. 4, pp. 3-13.

Van Hooland, S & Verborgh, R 2012, 'Joining the Linked Data cloud in a cost-effective manner', *Information Standards Quarterly* vol. 24, pp. 24-29.

Verhoeven, D 2012, 'New Cinema History and the Computational Turn', *Beyond Art, Beyond Humanities, Beyond Technology: A New Creativity: World Congress of Communication and the Arts Conference Proceedings*, University of Minho, Portugal

Appendix 1: List of HuNI Data Sources

Dataset	Schema	Data Type	Custodian or Owner
Australian Dictionary of Biography (ADB)	EAC-CPF	Biography	Australian National University
AusStage	Custom	Performance	Consortium – Flinders University
AUSTLANG	Custom	Linguistic	AIATSIS
Mura	Custom	Language	AIATSIS
AustLit	FRBR-derived	Literature	Consortium – University of Queensland
Design and Art Australia Online (DAAO)	EAC-CPF	Biography	Consortium – University of New South Wales
BONZA	Custom	Cinema and TV	Deakin University
CAARP	Custom	Cinema	Consortium – Deakin University (in association with Flinders University)
Dictionary of Sydney	Custom	History Geography	Consortium – Dictionary of Sydney Trust
PARADISEC	OLAC / RIF-CS	Linguistics	Consortium – University of Sydney
Media Archives Project	Dublin Core	Media Industry	Macquarie University
Australian Media History Database	Custom	Media Industry	Macquarie University
Encyclopedia of Australian Science	EAC-CPF (beta)	Biography	University of Melbourne
Saulwick Polls	Custom	Social Science, Politics	University of Melbourne
Find and Connect Australia (8 datasets)	Custom	Child Welfare	Consortium – University of Melbourne
Australian Women's Register	EAC-CPF	Women	Consortium – University of Melbourne
eMelbourne: the Encyclopedia of Melbourne	Custom	Melbourne	Consortium – University of Melbourne
eGold: Electronic Encyclopedia of Gold in Australia	Custom	Gold Mining	Consortium – University of Melbourne
Wallaby Club	Custom	History	University of Melbourne
Obituaries Australia	Custom	Biography	Australian National University

Circus Oz Living Archive Video Collection	Custom	Circus	RMIT University
Australian Film Institute Research Collection	Custom	Film	RMIT University