AGREGACIÓN DE LOS DATOS PARA VINCULAR SOCIAL EN LAS HUMANIDADES Y ARTES CREATIVAS: LA *HUMANITIES NETWORKED INFRASTRUCTURE (HUNI)*

AGGREGATING DATA FOR SOCIAL LINKING IN THE HUMANITIES AND CREATIVE ARTS: THE *HUMANITIES NETWORKED INFRASTRUCTURE (HUNI)*

Toby BURROWS
King's College London, and University of Western Australia
toby.burrows@kcl.ac.uk

Deb VERHOEVEN
Deakin University
deb.verhoeven@deakin.edu.au

Resumen: Este documento informa sobre el desarrollo de la *Humanities Networked Infrastructure (HuNI),* un servicio que agrega datos de treinta bases de datos de Australia y los pone a disposición por los investigadores a través de las humanidades y las artes creativas. Se discuten los métodos utilizados para datos agregados, así como el marco conceptual que ha dado forma al diseño del modelo de datos de HuNI. Dos de las principales funciones disponibles para los usuarios de HuNI – la construcción de colecciones y la creación de vínculos – se discuten, junto con sus fundamentos de diseño.

Abstract: This paper reports on the development of the *Humanities Networked Infrastructure (HuNI),* a service which aggregates data from thirty Australian data sources and makes them available for use by researchers across the humanities and creative arts, and more widely by the general public. We discuss the methods used by HuNI to aggregate data, as well as the conceptual framework which has shaped the design of HuNI's Data Model around six core entity types. Two of the key functions available to users of HuNI – building collections and creating links – are discussed, together with their design rationale.

Palabras clave: Agregación de datos. Humanidades. Artes creativas. Vinculación social.

Key Words: Data aggregation. Humanities. Creative arts. Social linking.

0. INTRODUCTION

The Humanities Networked Infrastructure (HuNI)[1] is one of the Virtual Laboratories developed with funding from the Australian Government's NeCTAR (National e-Research Collaboration Tools and Resources) programme[2]. The general parameters for these Virtual Laboratories were defined by NeCTAR. They focused on integrating existing e-research capabilities (tools, data and resources), supporting data-centred research workflows, and building virtual research communities to address existing well-defined research problems. This framework was designed with *big science* in mind; other Virtual Laboratories were funded in areas like climate science, geophysics, astronomy, genomics, characterisation and marine science.

The *data-centred* nature of the framework presented a challenge for the humanities research community. It was clear that NeCTAR expected something more than a service built around a collection of digital images or digital texts; a digital library or a *Europeana*-type service was not what was envisaged. To address this, the HuNI consortium developed and applied a definition of *data* which would be relevant to a wide range of humanities researchers and which would also meet NeCTAR's expectations.

In the humanities, *data* is a term that is not always well understood or agreed upon (Burrows, 2011). Collections of source material, whether physical or digital, are often described as *humanities data* (Borgman, 2007: 215-217), usually accompanied by *metadata* descriptions of these sources. HuNI has taken a different approach to defining data. For HuNI, *humanities data* consists primarily of the *semantic entities* referenced by the products of the humanities research process, whether these be books, articles, artworks, annotations, tags, reviews, ratings or other types of content. HuNI is not a collection of digital texts or images, nor is it built around catalogue records for these kinds of resources. Instead, HuNI focuses on the people, places, events and concepts referenced and discussed by humanities researchers.

What this means in practice is that HuNI does not contain a comprehensive catalogue-style record for a book like Richard Flanagan's *The Narrow Road to the Deep North* or for a movie like Baz Luhrmann's *Australia*. Instead of combining information into one record about the people involved with these works (authors, directors, actors, producers), their titles, their themes, and their locations, HuNI separates these out into individual entity records. There are individual entities for Flanagan, Luhrmann, Hugh Jackman, Nicole Kidman, *Australia*, *The Narrow Road to the Deep North*, and so on. This approach was taken because these entities (and the relationships between them) are fundamentally what humanities researchers want to discuss, analyse and talk about.

[1] *http://huni.net.au.*

[2] *http://nectar.org.au.*

The user community for HuNI is, effectively, the entire range of humanities and creative arts researchers in Australia and beyond (and is envisaged as extending to include non-specialist researchers). This was reflected in the composition of the various project teams and working groups, as well as in the disparate sources of data (discussed below). Thirteen different institutions actively contributed to the project – including universities, government institutes, and e-research service providers. HuNI is consciously designed to bridge the gap between cultural heritage institutions, academic researchers, and the wider community. The design and testing groups during the project included people from all of these sectors.

1. DATA AGGREGATION

Thirty different humanities datasets have been incorporated into HuNI. The data in some of these services conform to standard schemas, but many use their own customized format (as shown in Table 1 below). A wide range of disciplines within the humanities and creative arts are covered, including history, literature, performing arts, art and design, biography, and media studies.

Table 1: List of HuNI Data Sources

Data Set	Schema	Data Type	Custodian or Owner
Australian Dictionary of Biography (ADB)	EAC-CPF	Biography	Australian National University
AusStage	Custom	Performance	Consortium led by Flinders University
AUSTLANG	Custom	Linguistic	AIATSIS
Mura	Custom	Language	AIATSIS
AustLit	FRBR-derived	Literature	Consortium led by University of Queensland
Design and Art Australia Online (DAAO)	EAC-CPF	Biography	Consortium led by University of New South Wales
Bonza	Custom	Cinema and TV	Deakin University
CAARP	Custom	Cinema	Consortium led by Deakin University
Dictionary of Sydney	Custom	History, Geography	Consortium led by Dictionary of Sydney Trust
PARADISEC	OLAC / RIF-CS	Linguistics	Consortium led by University of Sydney
Media Archives Project	Dublin Core	Media Industry	Macquarie University
Australian Media History Database	Custom	Media Industry	Macquarie University

Data Set	Schema	Data Type	Custodian or Owner
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 data sets)	Custom	Child Welfare	Consortium led by University of Melbourne
Australian Women's Register	EAC-CPF	Women	Consortium led by University of Melbourne
eMelbourne: the Encyclopedia of Melbourne	Custom	Melbourne	Consortium led by University of Melbourne
eGold: Electronic Encyclopedia of Gold in Australia	Custom	Gold Mining	Consortium led by 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	Cinema and TV documentation	RMIT University

HuNI harvests records from these sources in both XML and non-XML formats. But HuNI does not aggregate the incoming records by normalizing or mapping them to a uniform schema, as services like *Europeana* do. HuNI is not a *union catalogue* of humanities database records. Instead, the incoming harvested records are parsed to identify their primary entity type. They are then mapped to one of the six core entities in the HuNI Data Model: Person, Organization, Event, Work, Place, and Concept. This positions HuNI somehwere between a *data warehouse* in which the incoming data are first cleaned and organised into a consistent schema and a *data lake* in which the incoming data are ingested in their raw form and the responsibility for making sense of the data lies entirely with the end user.

The initial plan for HuNI envisaged that all the incoming data would be mapped to a detailed and sophisticated ontology – assembled from such sources as CIDOC-CRM (Comité International pour la Documentation – Conceptual Reference Model), FOAF (Friend of a Friend) and FRBR-OO (Functional Requirements for Bibliographic Records – Object Oriented). This approach was abandoned after fundamental conceptual and ethical difficulties were identified with it (Burrows, 2014). The HuNI team felt that it was inappropriate to attempt to impose a single, unified, complete ontological perspective across disciplines which have very different (and yet overlapping) approaches to categorization and knowledge representation. It was also decided that, as HuNI's purpose was not to replace the underlying datasets, any modelling of the data in HuNI did not need to cover comprehensively everything represented in the contributing services. And finally, as one of HuNI's key rationales was to encourage interdisciplinary understanding in humanities research, a Domain-Driven Design (DDD) process based on the recognitition and preservation of *bounded contexts* (in this case scholarly disciplines) was also deemed usuitable (Evans, 2004).

Instead, the HuNI team preferred to use a very generic categorization, with the aim of acknowledging disciplinary perspectives while providing a level of interoperability between them. As a result, the HuNI Data Model is deliberately restricted to six core entities, defined as follows. This Data Model was derived from a thorough analysis of the types of entities present in the source datasets, in order to identify the generic common ground between them.

Table 2: HuNI Data Model: Core Entity Types

Entity Type	**Definition**
PERSON	A natural person
ORGANISATION	A company, club, trust, gallery, political party, etc.
WORK	A cultural artefact created by someone, which has some existence in its own right, either physical or digital
PLACE	A real, spatial location
EVENT	An activity that occurs in space and time and may involve people, organisations, places, works, etc.
CONCEPT	Something whose existence is primarily mental

As of February 2015, HuNI contained more than 741,000 entities, categorized as follows:

- Concept (5,965)
- Event (74,016)
- Organization (44,809)
- Person (284,912)
- Place (10,611)
- Work (321,017)

The content of HuNI will continue to expand during 2015 as a result of ongoing uploads from the source datasets. A major new additional source of content will be supplied by another NeCTAR-funded project. In collaboration with another NeCTAR

Virtual Laboratory – Alveo: the Virtual Laboratory for Human Communication Science[3] – HuNI plans to ingest entity references from the Trove digitized newspaper collection. Developed and maintained by the National Library of Australia, this collection contains more than 15.7 million page images, and accompanying OCR files, containing 150 million articles from Australian newspapers published between 1803 and 1954.[4]

No relationships between entities are imported or inferred as part of the HuNI ingest process. Initially, this was partly the result of constraints imposed by the project's timelines and resources. But there was also a conceptual reason behind this decision: inferring and creating relationships in HuNI between entities from different data sources would again be imposing an unwarranted *supra-disciplinary* perspective on disparate data. Relationships recorded in a single incoming record from a single data source can still be replicated between the resulting HuNI entities without distorting the disciplinary perspective inherent in the original data.

A deliberate decision was also made not to merge entities from different data sources into a single *authoritative* entity. The intention was to ensure that the different disciplinary contexts for these apparently duplicated entities were preserved. This also indicates that HuNI does not mean to replace the underlying datasets by imposing its own version of the underlying information or their meaning. Records are ingested on the HuNI side and displayed via huni.net.au with pointers back to the original source records. Typically a limited range of record types and entity fields are mapped from the source datasets to HuNI. Currently HuNI only picks out one entity from each incoming record from each of its data sources. This means that there is a simple one to one relationship between an incoming record and the HuNI record produced. Future iterations of HuNI will provide the ability to extract more than one HuNI entity (record) from each incoming source record. The HuNI entities have not yet been mapped to a normative vocabulary, though exposing HuNI entities to the Linked Data cloud will be tackled as part of the next stage of HuNI's development, during 2015/16.

2. TECHNOLOGIES

The HuNI Virtual Laboratory is built with Open Source technologies, and consists of four main components:

- The Solr Document Index contains the harvested and indexed partner documents. It exposes a search API, allowing matching documents to be returned. It is a read-only resource.

[3] *http://alveo.edu.au.*

[4] *http://trove.nla.gov.au/ndp/del/about.*

- The Database stores user profile information, links between documents, collection lists, and associated metadata. It is a read-write resource, allowing users to manipulate HuNI information.
- The Virtual Laboratory functionality is delivered through an Nginx HTTP server and a RESTful API service. The Nginx server sends the application Javascript, HTML components, stylesheets, and images to the client (the user's browser). The RESTful API allows the client application to query and manage the user profile information, links, and collections. It also enforces access restrictions.
- The Nginx proxy server accepts all Internet-facing requests and delegates them to the appropriate backend service. All access to the HuNI Virtual Laboratory is via HTTPS.

Data is imported into the Solr Document Index through a four-step pipeline. Each partner site makes a feed available to HuNI for harvesting on a publicly accessible location via the Internet. Each step in the pipeline results in a file on disk in the raw, clean, and final Solr format for every document ingested into HuNI. The four steps in the pipeline are as follows:

1 Harvesting: partner sites are polled daily for updates using either HuNI's custom *Simple XML* format or the OAI-PMH protocol. The harvest code uses custom Python and bash scripts.

2 Pre-processing: where necessary, the harvested data are pre-processed to ensure they can be properly transformed.

3 Transforming: Custom Python code and XSLT templates are deployed to transform the harvested data into the standard HuNI Data Model, ready for indexing by Solr.

4 Indexing: Documents created by the transformation process are submitted to a Solr instance for indexing. The result is a body of indexed documents made up of the most recently harvested versions. This can be quickly searched through an HTTP interface.

3. USING THE DATA

As well as searching the aggregated data and browsing the entities attached to each of the six core entity types, registered users of HuNI can carry out two key functions: creating collections of entities, and creating links between individual entities. User collections bring together selected entities under a heading assigned by the user. The collections can be public or private, and users can add or delete entities from their own HuNI collections at any time. User-created collections in HuNI can be exported for reuse in other software environments. The HuNI record for each entity in a user-created public collection includes the information that they are part of that collection.

Users cannot create entity records directly in HuNI; new entity records can only be added to the HuNI aggregate by the ingestion of datasets through the HuNI pipeline. But there is a way in which individual users can contribute entity records to HuNI through that pipeline. The Heurist humanities e-research tool (developed to manage individual researchers' data collections)[5] has been modified to export its datasets to HuNI. The first major dataset loaded through the Heurist tool was *TUGG: The Ultimate Gig Guide*. This dataset contains 624 records related to live music venues in Melbourne. The TUGG database documents the history of the live music scene in Melbourne from organized dance hall events, to discos and the thriving pub music scene of today[6].

Creating links between individual entities is central to HuNI's purpose and functionality. A user can select two entities to connect, can describe the nature of the relationship between the entities, and can annotate the link. This process has been dubbed *social linking*, since the links are public by default. In the initial version of HuNI, there are no pre-set vocabularies or taxonomies for describing links, and users are free to choose their own form of words – though they are prompted with pre-existing matching strings to choose from when creating a link. Multiple links can be created in both directions between two entities, both by different users and by the same user. It is also possible to assert *is not* relationships, such as *is not the sister of.* This recognises the critical importance of contestation in humanities-based approaches to knowledge formation.

The links attached to a specific entity record can be viewed in a text-based tabular form. The links for the pioneering Australian zoologist Albert Le Souef (1828-1902), for example, are displayed as follows. Each entry also indicates the type of core entity and the name of the HuNI user who created the link.

Table 3: Entity Relationship Links

Record name	**Relationship to**	**Relationship from**
Dudley Le Souef	Parent Of	Child Of
William Le Souef	Parent Of	Child Of
Ernest Le Souef	Parent Of	Child Of
Albert Le Souef	Parent Of	Child Of
Acclimatisation Society of Victoria	Associated With	Associated With
Zoo	Director of	Associated With
Albert Alexander Cochrane Le Souef	Same As	Same As

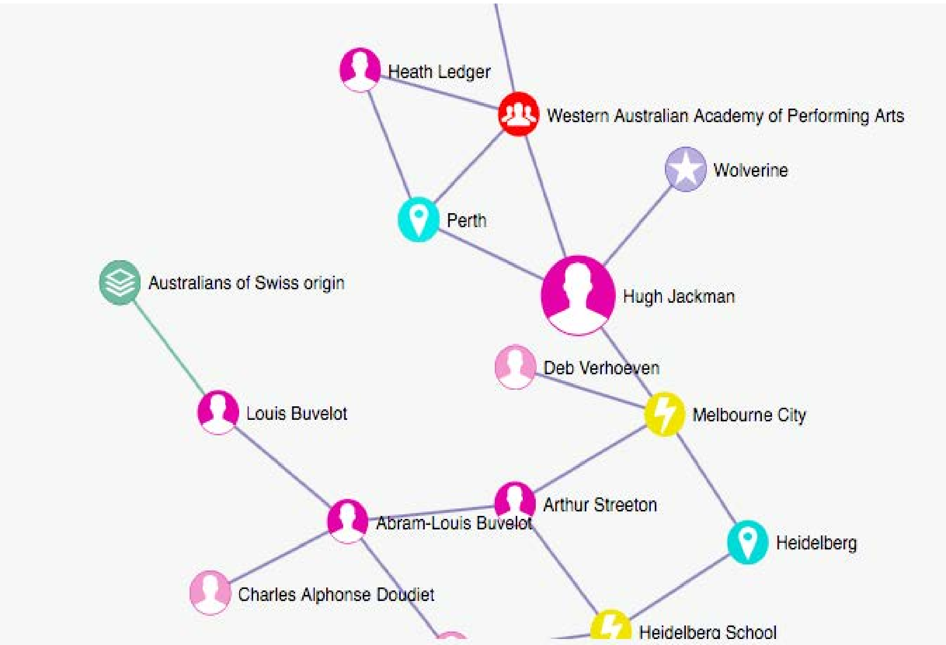
[5] *https://code.google.com/p/heurist/.*

[6] *http://tugg.me.*

Record name	Relationship to	Relationship from
Albert Alexander Cochrane Le Souef	Same As	Same As
Una Falkiner	Parent Of	Child Of

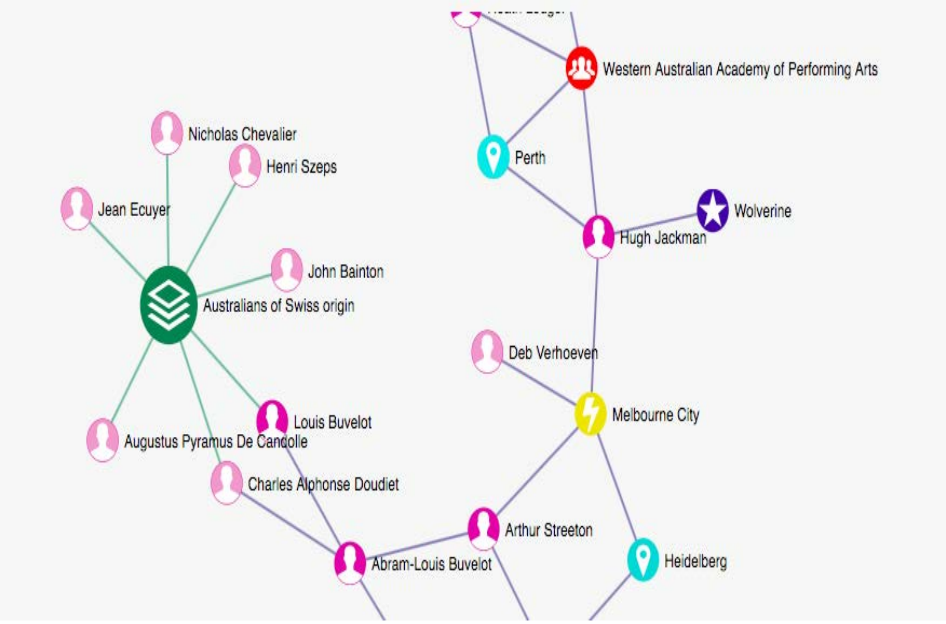
More significantly, however, the graph of links between entities can be browsed through a network visualization interface. Figure 1 shows an example of a HuNI network graph, centred on the Australian actor Hugh Jackman. Each different type of entity is identified with a distinct icon. Jackman himself is linked to a Place (Perth), a Work (Wolverine), an Organization (Western Australian Academy of Performing Arts), and a Concept (Melbourne City). These entities, in their turn, link outwards to other related HuNI entities, as well as to user-created collections. One of the Persons in this graph, the artist Louis Buvelot, is linked to a user-created Collection entitled *Australians of Swiss origin.*

Figure 1: HuNI Network Graph – Hugh Jackman



Selecting any of the icons representing entities in the initial network graph changes the focus of the graph. Clicking on the icon for the Collection *Australians of Swiss origin* displays all the entities which have been included in that collection by the user who created it and made it public, as shown in Figure 2. These newly-revealed entities can then be selected in their turn. The number of *degrees of separation* which can be displayed is only limited by the size and resolution of the user's screen. The graph in Figure 2 displays at least eight links, from Henri Szeps to the Western Australian Academy of Dramatic Arts.

Figure 2: HuNI Network Graph – Australians of Swiss origin



The two functions discussed in this section are intended to allow researchers to add their own meaning and structure to the aggregated HuNI data. The *collections* functionality allows users to create their own categories and groupings for entities. The *social linking* function allows them to create their own graph of relationships and to contribute to the growing HuNI network graph. Researchers can trace routes along these interconnected networks, as an alternative discovery process to a keyword search.

Researchers who tested the initial version of HuNI prototype commented on the benefits of this approach in enabling them to make "serendipitous discoveries through identifying points of commonality between data" and to "cross-search a significant amount of data in a single software environment and see networks of relationships" (anonymous user feedback). This reinforces HuNI's role in contributing to the design of digital resources for the humanities which foster serendipity (Verhoeven and Burrows, 2014).

4. CONCLUSION

Interpretation is at the heart of the humanities and creative arts. HuNI combines humanities data in a way which enables researchers to express, share and discuss their differing interpretations of the data. The different perspectives between (and within) disciplines are preserved and foregrounded, instead of being hidden behind a normalized, *authoritative* framework. HuNI has kept categorization and taxonomical structures to a minimum, and has provided the tools for researchers to create their own semantic frameworks for the data.

Cultural data are not economically, culturally, or socially insular. Researchers need to collaborate across disciplines, institutions, and social locations, in order to explore data fully (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 organizing, accessing, and presenting this evidence are critical for our enquiries. HuNI is designed to address this need.

BIBLIOGRAPHIC REFERENCES

BORGMAN, C. (2007). *Scholarship in the Digital Age*. Cambridge, MA: MIT Press.

BURROWS, T. (2011). “Sharing humanities data for e-research: conceptual and technical issues”. In *Sustainable Data from Digital Research*, N. Thieberger (ed.), 177-192. Melbourne: PARADISEC.

_____ (2014). “Ontologies and the humanities: some issues affecting the design of digital infrastructure”. *Digital Humanities Congress, Sheffield, UK, September 2014* (*http://www.slideshare.net/TobyBurrows/dhc2014-burrows-final*).

EVANS, E. (2004). *Domain-Driven Design: Tackling Complexity in the Heart of Software*. Boston: Addison-Wesley.

VERHOEVEN, D. (2012). “New Cinema History and the Computational Turn”. In *Beyond Art, Beyond Humanities, Beyond Technology: A New Creativity: World Congress of Communication and the Arts Conference Proceedings*. Minho, Portugal: University of Minho.

VERHOEVEN, D. and BURROWS, T. (2014). “Crowdsourcing for serendipity”. *The Australian Higher Education Supplement* 10 (December).

Recibido el 10 de octubre de 2015.
Aceptado el 30 de octubre de 2015.