Metadata of the chapter that will be visualized in SpringerLink

Book Title	Metadata and Semantics Research	
Series Title		
Chapter Title	Aggregating Cultural Heritage Data for Research Use: the Humanities Networked Infrastructure (HuNI)	
Copyright Year	2015	
Copyright HolderName	Springer International Publishing Switzerland	

Author	Family Name	**Verhoeven**
	Particle	
	Given Name	**Deb**
	Prefix	
	Suffix	
	Division	School of Communication and Creative Arts
	Organization	Deakin University
	Address	Melbourne, Australia
	Email	deb.verhoeven@deakin.edu.au
Corresponding Author	Family Name	**Burrows**
	Particle	
	Given Name	**Toby**
	Prefix	
	Suffix	
	Division	Department of Digital Humanities
	Organization	King's College London
	Address	London, UK
	Division	School of Humanities
	Organization	University of Western Australia
	Address	Nedlands, Australia
	Email	toby.burrows@kcl.ac.uk toby.burrows@uwa.edu.au

Abstract

This paper looks at 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. 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.

Keywords (separated by '-') Data aggregation - Humanities - Creative arts - Social linking

Aggregating Cultural Heritage Data for Research Use: The Humanities Networked Infrastructure (HuNI)

Deb Verhoeven[1] and Toby Burrows[2,3](✉)

[1] School of Communication and Creative Arts, Deakin University, Melbourne, Australia
deb.verhoeven@deakin.edu.au
[2] Department of Digital Humanities, King's College London, London, UK
toby.burrows@kcl.ac.uk
[3] School of Humanities, University of Western Australia, Nedlands, Australia
toby.burrows@uwa.edu.au

Abstract. This paper looks at 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. 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.

Keywords: Data aggregation · Humanities · Creative arts · Social linking

1 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, as defined by NeCTAR, 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. Most of the other Virtual Laboratories were funded in big data areas of science, including 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 other 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 had to develop and apply a definition of "data" which would be relevant to humanities researchers but would also meet NeCTAR's expectations.

In the humanities, "data" is a term that is not always well understood or agreed upon. [3] Collections of source material, whether physical or digital, are often described as "humanities data", usually accompanied by "metadata" descriptions of

E. Garoufallou et al. (Eds.): MTSR 2015, CCIS 544, pp. 1–7, 2015.
DOI: 10.1007/978-3-319-24129-6_36

these sources. [4] HuNI has taken a different approach. 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.

This means that HuNI does not contain catalogue-style records for books like Richard Flanagan's *The Narrow Road to the Deep North* or for movies like Baz Luhrmann's *Australia*. Instead of combining information into one record about the people involved with these works (authors, directors, actors, producers), the titles of the works, 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 it is these entities – and the relationships between them – which are the fundamental focus for the discussions, analyses and conversations of humanities researchers.

The user community for HuNI is, effectively, the entire range of humanities and creative arts researchers in Australia and beyond. This was reflected in the composition of the various project teams and working groups, as well as in the disparate sources of data. Thirteen different institutions actively contributed to the project – including universities, government institutes, and e-research service providers. HuNI is 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.

2 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. 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. The datasets, for the most part, have been developed as ongoing services by consortia involving researchers and cultural institutions, usually with government funding.

HuNI harvests records from these datasets 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 approach positions HuNI somewhere 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. [5] 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.

HuNI was not intended to replace the underlying datasets, which continue to exist and develop within their disciplinary context. As a result, 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 unsuitable. [6]

Instead, the HuNI team implemented a very generic framework for categorization, with the aim of acknowledging disciplinary perspectives while providing a level of interoperability between them. Because of this, the HuNI Data Model is deliberately restricted to six core entities: concept, event, organization, person, place, and work. 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. As of May 2015, HuNI contained more than 750,000 entities, categorized as follows:

- Concept (5,970)
- Event (76,015)
- Organization (45,276)
- Person (289,458)
- Place (10,828)
- Work (322,818)

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 intend to replace the underlying datasets by imposing its own version of the original information or its meaning. Records are ingested on the HuNI side and displayed in the HuNI service 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. This is done by harvesting only those source records which can be matched to one of HuNI's six basic categories. In some cases, this is straightforward; "Person" records in the AustLit database, for example, map to the HuNI "Person" category. In other cases, the mapping is more indirect; "Venue", "Company", and "Film" records in the CAARP database map to the HuNI categories "Event", "Organisation", and "Work" respectively. These mappings are hard-coded into the harvest and ingest process, and are based on a thorough comparison between the data models of the source datasets and the HuNI data model.

Currently the HuNI ingest process only picks out one entity from each incoming record from each of the source datasets. 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 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. Also currently under development is a data ingest pipeline for entity references identified in the text of the Australian digitized newspapers hosted by the National Library of Australia's Trove service. [7]

3 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. [8] It exposes a search API, allowing matching documents to be returned. It is a read-only resource.
- 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's JavaScript, HTML components, stylesheets, and images to the user's browser client. [9] The RESTful API allows the client application to query and manage the user profile information, links, and collections. [10] 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 the HTTPS protocol.

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 Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH). [11] 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.

4 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. These 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) has been modified to export its datasets to HuNI. [12] 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. [13]

The next stage in developing upload functionality for HuNI is being explored in the context of Omeka, the Open Source collections and exhibitions publication platform. [14] As part of a national e-research project, a "publish to HuNI" plug-in will be developed for Omeka. This feature will be incorporated into a hosted version of Omeka, which will be available to all Australian university researchers.

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".

No central editorial control is imposed by HuNI on the creation of links. Nor do the creators of links have to be recognized "experts" in a particular disciplinary area. Any registered user of HuNI is able to create links and publish them into the HuNI network graph. The creator of each link is recorded and publicly identified, enabling subsequent users to see the source of the link and assess its authoritativeness. This approach recognizes the critical importance of contestation and plurality in humanities-based frameworks for knowledge formation. [15]

The graph of links between entities can be browsed through a network visualization interface. Each different type of entity is identified with a distinct icon. These entities, in their turn, link outwards to other related HuNI entities, as well as to user-created collections. Selecting any of the icons representing entities in the initial network graph changes the focus of the graph. 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 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 the 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. [16]

5 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. [17] 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.

References

1. http://huni.net.au
2. http://nectar.org.au
3. Burrows, T.: Sharing humanities data for e-research: conceptual and technical issues. In: Thieberger, N. (ed.) Sustainable Data from Digital Research, pp. 177–192. PARADISEC, Melbourne (2011)
4. Borgman, C.: Scholarship in the Digital Age, pp. 215–217. MIT Press, Cambridge (2007)
5. Burrows, T.: 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
6. Evans, E.: Domain-Driven Design: Tackling Complexity in the Heart of Software. Addison-Wesley, Boston (2004)
7. http://trove.nla.gov.au/ndp/del/about/
8. http://lucene.apache.org/solr/
9. http://nginx.org/en/
10. Richardson, L., Ruby, S.: RESTful Web Services. O'Reilly Media, Farnham (2007)
11. https://www.openarchives.org/pmh/
12. https://code.google.com/p/heurist/
13. http://tugg.me
14. http://omeka.org
15. Rowland, S.: The Enquiring University. McGraw-Hill, Milton Keynes (2006)
16. Verhoeven, D., Burrows, T.: Crowdsourcing for serendipity. The Australian Higher Education Supplement, December 10, 2014
17. Verhoeven, D.: 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. University of Minho, Minho, Portugal (2012)