

# **HCSNet Data Activity**

## **Report to Steering Committee 10 August 2006**

### ***1. Background***

HCSNet has supported activity in the focus area of data almost since inception, in various phases. Outcomes of previous work can be found on the HCSNet web site under the Data Activity item, this largely consists of reports in a similar fashion to this one.

### ***2. Recent Activity***

This report essentially covers activity after the December 2005 HCSNet SummerFest event where a speed paper was presented about activities surrounding data collections.

During the first half of 2006, I have personally resourced the activity that has taken place, albeit under the umbrella of research areas covered by my academic appointment. The net cost to HCSNet has been zero in this regard. Essentially I have spent time engaged in researching collections and issues, largely aligned with other research activity. In addition to research tasks, a number of ad hoc discussions have been held where individuals concerned with Australian data collections have gathered, although no formal workshop has been held.

### ***3. Outcomes***

What then are the properties of data collections and the associated problems with sharing data more broadly between HCSNet researchers?

#### **3.1 Different definitions of data and data collections**

For example, in natural language processing and information retrieval, data and data collections are usually interpreted to mean curated collections of audio, video or text; whilst in psychology, data and data collections typically mean stimulus, the objects that provoke the elicitation of human communication which is then recorded for analysis. This diversity of views provides both potential leverage points and potential pitfalls in terms of cross-disciplinary research. Both interpretations lend themselves to data sharing at some level within specific disciplines, however, between discipline sharing is an entirely different matter.

## 3.2 Different traditions of data sharing

In certain sub-communities within HCSNet, data sharing is well established through the frameworks of shared evaluation tasks. In such frameworks, shared data collections are primarily used for benchmarking purposes in assessing advancements over the present state of the art, providing a common evaluation baseline.

In other communities, data is regarded as project specific, and never released beyond a single research project (secondary analysis, as it is called in the social sciences, is a rarity).

## 3.2 Australia-specific data collections

One of the focus areas within recent activity has been to identify specifically Australian data collections which may be of use to researchers across the human communication sciences. The motivation for this is twofold: first, there is a case to be made for the Australian context being differentiated from the rest of the world in some specific areas (e.g. pronunciation, socialization) and second, larger more common resources are typically available (albeit for a fee) from distribution agencies and we do not seek to replicate this model.

In short, we have been able to identify a small number of specifically Australian data collections which are available under some distribution scheme. These collections are largely oriented at natural language processing, and have specifically Australian language properties:

- AVOZES: stereo video corpus
- ANDOSL: speech corpus
- ACE: balanced text corpus (part of ICAME collection)

These collections are available for direct purchase via an online (or close to online process). The HCSNet Data Activity web site has more information on each of these including access information.

There are other data collections in existence at “pre-distribution” levels: for example Monash University has a collection of Australian television transcripts, but which are only available on campus through the library, yet have been used by more than one research group.

There are other data collections in various states of curation which could be turned into these types of data collections and made available for research use. Generally these collections are archival in the sense of having been collected at some time past, and committed raw data and analysis to an archive of some kind (typically a university library). For example, the Flint Archive at the University of Queensland has a wealth of Australian language data (including significant indigenous language content) recorded in

the 1960s. Preliminary work on digitizing this data, and making a catalogue available has been undertaken in the mid-late 1990s.

### **3.3 Privacy concerns**

In some disciplines, data sharing is restricted by virtue of privacy concerns largely governed by human ethics policies applied at the point of data collection. This being said, there are some other types of data, e.g. corpora specifically compiled for distribution, where this is not an issue.

## ***4. Moving forward***

In light of these findings, what does this leave us with as a community interested in sharing data to promote cross-disciplinary research? We have identified a number of areas of potential engagement by HCSNet in the area of data collections.

### **4.1 Supporting Australian participation in shared evaluation tasks**

In research areas where shared evaluation tasks are common, there are typically two barriers to entry: the cost of aligning research efforts with the shared evaluation task domain, and the cost of attending the typical evaluation workshop. Within Australia we have groups participating in large (natural language processing and information retrieval oriented) shared evaluation tasks such as those run within TREC, CLEF and NTCIR, more focused efforts in efforts such as CoNLL, and a range of smaller scale shared tasks. All of these groups could benefit from some support scheme, perhaps in the form of a student or researcher stipend to attend the evaluation workshops.

### **4.2 Repatriation and access facilitation for Australian data collections**

There are Australian specific corpora in existence elsewhere in the world (e.g. ACE as a part of the ICAME corpus), but these are not readily available within Australia and subject to distribution and licensing arrangements which present a considerable barrier to entry for researchers specifically interested in Australian data. One possible area for HCSNet to become involved is in acquiring broad access licenses for researchers perhaps by funding library subscriptions to these resources in a select number of universities.

### **4.3 Supporting Australian data collection development**

There is a need for specifically Australian data collections in a range of areas. However, the data collection and curation effort is not a cheap exercise if it is to be conducted on a large enough scale to be truly useful, and requires significant investment in human

resources as well as technical infrastructure. Elsewhere, such efforts are typically coordinated by distribution agencies, who directly or collaboratively invest in corpus creation. Getting Australian data included in such collection efforts is likely a combination of financial and political persuasion, although it does happen by accident at times. For example, in 2005 the Linguistic Data Consortium funded the CallHome2 data collection project which included Australian English as one of the target varieties. However, insufficient data was collected under this effort to result in a release of an Australian component of the corpus. It is possible with HCSNet support that this situation could be addressed.

On a different tack, supporting the development of Australian indigenous language data collections is another opportunity. For example, AIATSIS have since 1994 maintained ASEDA, an archive of indigenous language data in electronic form (particularly transcripts, wordlists contributed by academic linguists conducting fieldwork). However, AIATSIS have consistently been unable to secure funding to deliver these data collections in a standard repository framework by which researchers can access their holdings. Again, it is possible that HCSNet could provide some support to ensure availability of these resources.

#### **4.4 Supporting existing data collection enrichment and curation efforts**

Many funded research projects focus on collecting data and creating data collections. However, the overhead for tuning these collection and curation efforts to ensure broadly useful resources is small, and with timely intervention and support, a significant difference can be made between a project specific data collection, and one which is at least described adequately for other researchers to consider including in their own research. For example, “Data Grants” specifically targeted at enriching existing corpora or curating them such that they can be made more generally available. There are numerous collections in archives in Australia which could be brought back into research currency with sufficient investment in curation effort; there are others where HCSNet could add value by funding specific additional activity related to data collection and curation within existing projects.

#### **4.5 Supporting data archiving initiatives**

The archiving of data collections is becoming a significant focus of research funding agencies. In the US (NSF, NEH and others), UK (JISC, ESRC, and others) and Europe (EU) there are already mandated requirements for archiving research data collections as a condition on acceptance of research funding. The motivation is largely “public access to publicly funded research output”. In Australia we are starting to see the emergence of similar models with the support of the ARC, NHMRC and other agencies. Indeed we already have some funded archive projects which nominally cover the data collections used within the human communication sciences. HCSNet could take a leadership role in

both practical and policy areas: supporting archives in their data acquisition efforts (perhaps by funding the data deposit cycle), and engaging research funding bodies to resource discipline relevant archive establishment and in policy development to ensure archiving of research generated data collections.

## **4.6 International engagement**

In the global context, there are research networks with a data collection and curation focus. Active in the Asia Pacific region are bodies such as COCODA and WRITE, and the recently established Asian Language Resources Network (ALRN). These networks have various levels of activity ranging from funding workshops, developing standards for data collections, endorsement and sponsorship of conferences, providing training. HCSNet could engage with the research community in a similar way, and perhaps in concert with these bodies.

## **5. A Personal Note**

Until this month, my involvement in the HCSNet Data Activity has been under the auspices of a position as Research Fellow in the Department of Computer Science and Software Engineering at The University of Melbourne. However, I have taken another position within the same institution which will now occupy a high proportion of time previously committed to research, and as such my personal availability to move this effort forward will be reduced (ironically, my new position is closely related to HCSNet's Priority Area in Next Generation Search). This does not diminish my interest in seeing HCSNet Data Activity moving forward. On the contrary is likely time for practical recognition that without specifically dedicated (and possibly funded) effort, it is unlikely that anything more than systematic assessments and status reports will eventuate.

## **6. Contact Details**

Baden Hughes

Email: [badenh@csse.unimelb.edu.au](mailto:badenh@csse.unimelb.edu.au)

Mobile: 0418 466 107

Office: 03 8344 0480

VOIP: 03 9018 7408