

# The Australian National Corpus Initiative: Technical and Legal Issues

Michael Haugh<sup>1</sup>, Dennis Alexander<sup>2</sup>, Linda Barwick<sup>3</sup>, Denis Burnham<sup>4</sup>, Kate Burridge<sup>5</sup>, Steve Cassidy<sup>6</sup>,  
Michael Clyne<sup>7</sup>, Anne Fitzgerald<sup>8</sup>, Cliff Goddard<sup>9</sup>, Jane Hunter<sup>10</sup>, Bruce Moore<sup>11</sup>, Simon Musgrave<sup>12</sup>,  
Pam Peters<sup>13</sup>, Roly Sussex<sup>14</sup>, Nick Thieberger<sup>15</sup>

<sup>1</sup>Griffith University, Brisbane, m.haugh@griffith.edu.au, <sup>2</sup>Department of Education, Employment and Workplace Relations, Canberra, Dennis.ALEXANDER@deewr.gov.au, <sup>3</sup>University of Sydney, Sydney, linda.barwick@gmail.com, <sup>4</sup>University of Western Sydney, Sydney, Denis.Burnham@uws.edu.au, <sup>5</sup>Monash University, Melbourne, Kate.Burridge@arts.monash.edu.au, <sup>6</sup>Macquarie University, Sydney, steve.cassidy@mq.edu.au, <sup>7</sup>University of Melbourne/Monash University, Melbourne, mgclyne@gmail.com, <sup>8</sup>Queensland University of Technology, Brisbane, am.fitzgerald@qut.edu.au, <sup>9</sup>University of New England, Armidale, cgoddard@une.edu.au, <sup>10</sup>University of Queensland, Brisbane, jane@itee.uq.edu.au, <sup>11</sup>Australian National University, Canberra, bruce.moore@anu.edu.au, <sup>12</sup>Monash University, Melbourne, Simon.Musgrave@arts.monash.edu.au, <sup>13</sup>Macquarie University, Sydney, Pam.Peters@ling.mq.edu.au, <sup>14</sup>University of Queensland, Brisbane, sussex@uq.edu.au, <sup>15</sup>University of Melbourne, Melbourne, nick.thieberger@gmail.com

## THE AUSTRALIAN NATIONAL CORPUS

The Australian National Corpus Initiative involves a concerted push by linguists, applied linguists and language technologists to establish a massive online database of spoken and written language in Australia in all its forms and diversity. It was resolved at a meeting at the Australian Linguistics Society Annual Conference in July 2008 that the Australian National Corpus must be freely accessible and have cross-disciplinary acceptability and uptake [1]. At a subsequent HCSNet-sponsored workshop in December 2008 it was resolved that the Australian National Corpus should consist of a distributed set of multimodal and multilingual resources which meet leading-edge technical standards [2]. By multimodal it is envisaged that not only plain text, but also visual texts, audio and audiovisual language data will feature in the corpus, while by multilingual it is intended that the corpus incorporate significant collections of English in Australia (including Australian English and migrant Englishes), indigenous languages, community languages, and Australian Sign Language. These collections of language data will include both historical collections (that is, language data sets that have already been gathered) as well as being a point of departure for current and future collections of relevant language data. This aim contrasts with other existing national corpora which are text-only, monolingual, and relatively static. At an ensuing workshop sponsored by the Australian Academy of the Humanities in May 2009 it was resolved that the various legal and ethical issues associated with sharing language data require the development of a legal framework for the Australian National Corpus.

A national repository of language data would have significant value as eResearch infrastructure for a number of research communities in Australia and overseas, thereby increasing access to Australian language data and widening the global integration of research on language in Australia. First, it would facilitate collaborative ventures in collecting new language data to support multimodal research in human communication. Second it would consolidate presently scattered and relatively inaccessible collections of historical language data where possible within the Australian National Corpus. Such data is of interest not only to researchers in linguistics and applied linguistics, but also to members of the wider Humanities and Social Sciences (HASS) and informatics research communities who have an interest in Australian society. Second, it would enable the emergence of the practice of data citation in language sciences [3]. Third, such a large annotated language dataset would provide invaluable training data for work in natural language processing, speech recognition, and the further development of semi-automated annotation.

This emphasis on providing widespread access to language data is consistent with the overall aims of the Australian National Data Service [4]. However, in order to maximize the potential for the Australian National Corpus to enable data sharing amongst researchers with an interest in Australian languages and society, the complex technical and legal issues that arise when attempting to share (historical) language data need to be addressed.

## TECHNICAL REQUIREMENTS OF THE AUSNC

The overall architecture of the Australian National Corpus is founded on the principle of stand-off annotation, as illustrated in Figure 1 below:

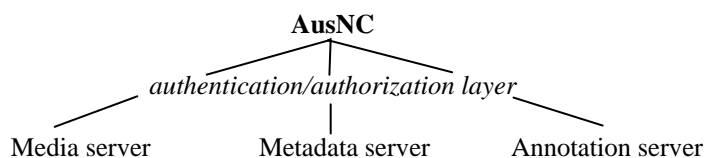


Figure 1: Proposed architecture of the Australian National Corpus [5]

While recognized annotation and metadata standards exist for various types of language data, such as the OLAC Metadata standards [6] or the Linguistic Annotation Framework [7], historical language data is often not annotated according to such standards, thereby posing sometimes significant technical challenges. In updating annotation of the Australian Component of the International Corpus of English in line with modern approaches to corpus annotation and analysis, for instance, the development of parsing tools has revealed inconsistencies in markup that require careful analysis [8]. Significant technical challenges remain, however, not only for audio(visual) data sharing, but also for annotating relevant text data mined from the World Wide Web, as seen in the recent construction of a 2 billion word corpora of British English (ukWaC) [9].

## LEGAL FRAMEWORK FOR THE AUSNC

A myriad of legal and ethical issues arise from making language data available to other researchers and more widely, including copyright, privacy and moral rights, as well as issues relating to indigenous and ethnic communities. While there has been important progress made in addressing these issues more generally for data sharing and management in the context of eResearch, for example, the Legal Framework for eResearch project [10], there are specific ethical and legal issues arising in the case of different forms of language data that remain to be explored [11]. The development of a legal framework for sharing language data in the Australian National Corpus is thus essential to ensure that not only are appropriate data management policies and plans formulated and implemented through appropriate access control [12], but that creative commons licensing of donations to the corpus as well as access agreements to data in the corpus are utilised as far as possible.

## CONCLUDING REMARKS

In being a national initiative it is envisaged that the technical and legal frameworks developed as part of the Australian National Corpus will have significant transferability to other data sharing initiatives, as well as being able to draw from existing eResearch initiatives in the Humanities and Social Sciences. In this way, we believe the Australian National Corpus initiative has the potential to contribute to the wider eResearch agenda in Australia.

## REFERENCES

1. Australian Linguistics Society, *Statement of Common Purpose: Australian National Corpus Initiative*, 4 July 2008.
2. Burridge, K., M. Haugh, J. Mulder, and P. Peters (eds.), *Selected Proceedings of the HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, 2009. Somerville, MA: Cascadilla Proceedings Project.
3. Wilkinson, R., *The Australian National Data Service*, Presentation at *UKOLN International Conference on the UK Research Data Service Feasibility Study*, 26 February 2009. Available from: <http://www.ukoln.ac.uk/events/ukrds-2009/programme/>.
4. Treloar, A. and R. Wilkinson, *Access to data for eResearch: designing the Australian National Data Service discovery services*. *The International Journal of Digital Curation*, 2008. 2(3): p. 151-158.
5. Cassidy, S., *Building infrastructure to support collaborative corpus research*, paper presented at the *HSCNet Workshop on Designing the Australian National Corpus*, UNSW, 4-5 December 2008.
6. Open Language Archive Community, *OLAC Metadata*, 5 May 2007. Available from: <http://www.language-archives.org/OLAC/metadata.html>.
7. Ide, N., L. Romary, and E. de la Clergerie, *International Standard for a Linguistic Annotation Framework*, in *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*, 2003. Edmunton.
8. Wong, D., S. Cassidy and P. Peters, *ICE Markup*. *International Journal of Corpus Linguistics*, submitted.
9. Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta, *The WaCky wide web: a collection of very large linguistically processed web-crawled corpora*. *Language Resources and Evaluation*, 2009. 43.
10. Fitzgerald, A., K. Pappalardo, and A. Austin, *Practical Data Management: A Legal and Policy Guide*, 2008. Brisbane: QUT epress.
11. The Australian Academy of the Humanities, *Workshop on The Future Australian National Corpus: Ethical and Legal Issues*, Ship Inn, Griffith University, 22 May 2009.
12. Nguyen, C., J. Dalziel and S. Cassidy, *Flexible access control, federated identity and heterogenous metadata supports for repositories*, in *Proceedings of eResearch Australasia 2008*, Melbourne.