Alveo: making data accessible through a unified interface – a pipe-dream?

Dominique Estival
MARCS Institute, Western Sydney University
Locked Bag 1797, Penrith NSW 2751, Australia
E-mail: d.estival@westernsydney.edu.au

Abstract

This paper addresses an old issue in corpus management which is still problematic in real-life systems: to allow users to explore and access data from various sources using a single simple interface, thus creating a tension between ease of use and over-simplification. This is then mirrored in the similar difficulty encountered with a simple data upload facility. In Alveo, the Virtual Lab for Human Communication Science, the original unified interface was sufficient for most of the datasets but proved inadequate in some cases. This paper is intended to facilitate a discussion on best practice with developers who may propose different solutions and with researchers who may have other requirements for their own datasets. We describe specific challenges posed by some datasets for Alveo, issues faced by users, identify the problems with the current state of development and propose several solutions.

Keywords: data access, data discovery, data upload, facets, hierarchy

1. Alveo

This paper addresses an old issue in corpus management which is still problematic in real-life systems: to allow users to explore and access data from various sources using a single simple interface, thus creating a tension between ease of use and over-simplification which is mirrored in the similar difficulty for a simple data upload facility. In Alveo, the Virtual Lab for Human Communication Science, the original unified interface was sufficient for most of the datasets but proved inadequate in some cases.

1.1. Aims of the Alveo project

The Alveo Virtual Lab (Estival, Cassidy, Sefton, & Burnham, 2013) was designed to:

- facilitate access by Australian and international researchers in Human Communication Science to a range of data and tools across the HCS disciplines (i.e. speech science, linguistics, psycholinguistics, computational linguistics, social sciences and musicology);
- afford new tool–corpus combinations, for instance, allow musicologists to discover speech science tools (and vice-versa) or computational linguists to access little-known historical text corpora;
- allow analysis and annotation results to be stored and shared, thus promoting collaboration between institutions and disciplines;
- improve replicability and reusability by moving local and idiosyncratic desktop-based tools and data to an accessible, in-the-cloud, environment to standardise, define and capture procedures and data output, so that research publications can be supported by re-runnable re-usable data and coded procedures (see e.g., www.myexperiment.org/).

1.2. Current status

Alveo uses Australian national infrastructure, such as data storage (RDS) and research computing services (NeCTAR Research Cloud). The platform itself is composed of 2 main parts. A Web discovery and search interface, through which users can explore the available datasets manages the licenses for each dataset, [1] also enables the construction of item lists across datasets. Item lists can then be imported in a Workflow engine derived from Galaxy (Goecks, Nekrutenko, Taylor, & Team, 2010) which offers a range of analysis and visualisation tools for easy use by researchers with limited technical background.

All access to data, including search via the Web interface, is mediated via an authorisation layer, and all data and services are made available via a RESTful web API (Cassidy, Estival, Jones, Burnham, & Berghold, 2014). The entities in the system (collections, items, documents, annotations, etc.) are identified via a URI and, following the principles of Linked Data, that URI resolves to a representation of that entity. The API enables more advanced users to build new services using the facilities of the core Alveo platform, and so far has allowed the

[1] Datasets currently available: Current datasets: PARADISEC, the Pacific and Regional Archive for Digital Sources in Endangered Cultures, including Indigenous languages music, and speech [13TB] (Thieberger, Barwick, Billington, & Vaughan, 2011); AusTalk, audio-visual speech corpus from the Big ASC project [34TB] (Burnham, Estival et al. 2011); AusNC, the Australian National Corpus, incorporating the Australian Corpus of English (ACE), Australian Radio Talkback (ART), AustLit, Braided Channels, Corpus of Oz Early English (COOEE), Email Australia, Griffith Corpus of Spoken English (GCSAusE), International Corpus of English (Australia contribution is ICE-AUS), the Mitchell & Delbridge corpus, and the Monash Corpus of Spoken English [5TB] (Musgrave & Haugh, 2009); AVOZES, visual speech corpus [13GB] (Goecke & Millar, 2004); CJI, Colloquial Jakartan Indonesian corpus (early 1990's) audio and text, ANU [32.5GB]; PixarMusic, music excerpts from films, expressing different emotions, UNSW [7.2MB]; RIR, room impulse responses, Sydney U. [816MB]; Emotional Prosody, sung sentences using different prosodic patterns Macquarie U. [30MB]; The ClueWeb09 dataset [100TB] (/lemurproject.org/clueweb12/); LLC, the Liberated Learning Corpus [81GB] (Bain, Basson, Faisman, & Kanevsky, 2005).

implementation of interfaces with tools that make use of Python (NLTK), R (Emu), Matlab, Java (UIMA) (Estival, Cassidy, Verspoor, MacKinlay, & Burnham, 2014).
Some of the Alveo user projects under way give a taste of the range of types of data and research interests from Alveo users: *An Iterative Implementation of MAUS: A model for Australian Languages; Comparison of special speech registers (infant- /foreigner- / computer- directed speech)*; *Building a corpus of varieties of Kriol; Creaky voices in Australian English; Audio-visual analysis of emotional speech.*

2. Alveo Search Interface: facets

Data in Alveo is organised by items, with one or more document per item. In the relatively simple case of a text corpus, e.g. the AusNC (Cassidy, Haugh, Peters, & Fallu, 2012), the item usually consists of one text document, but can sometimes consist of 2 or 3 documents, for instance 'plain text', 'raw text' and 'xml'. Viewing and searching data in the Alveo Web Discovery interface is effected through facets (Rodriguez-Castro, Glaser, & Carr, 2010), which are largely based on the facets that were defined for the collections comprising AusNC (Cassidy et al., 2012).
Figure 1 shows the view of the COOEE corpus when searching for texts written between 1780 and 1789, using the 'Created' facet.

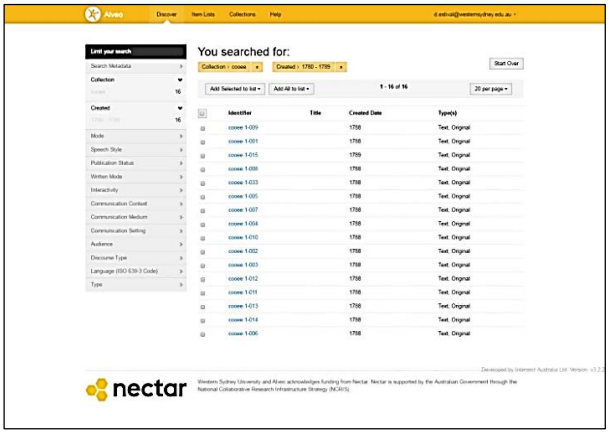
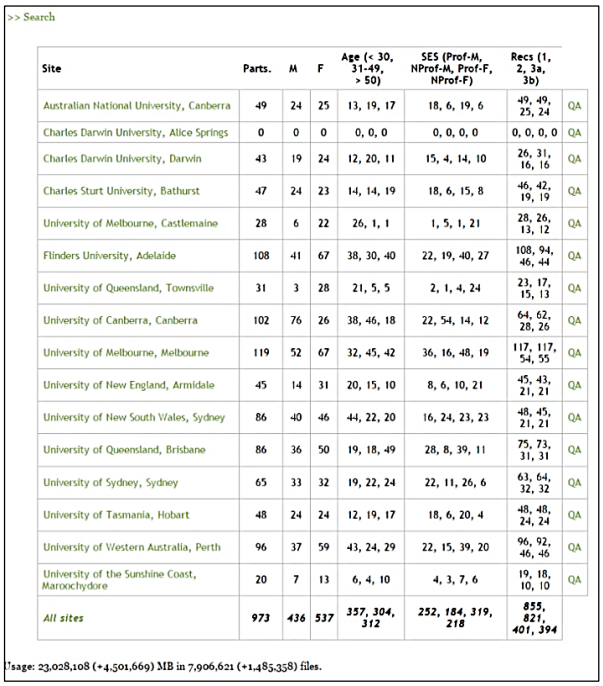


Figure 1: Screenshot of Alveo, COOEE 1780-1789

In the more complex case of an audio-visual corpus, e.g. AusTalk (Estival, Cassidy, Cox, & Burnham, 2014), an item consists of at least one audio file and one video file, with sometimes several files to be concatenated.[2]
AusTalk is an example of a dataset where the simple view provided by Alveo was problematic. As shown in Figure 2 (from austalk.edu.au), the original AusTalk interface lists all the speakers organised by recording sites and gives the demographic distribution per site. When drilling down to each site and each speaker, we can view the data as organised in recording sessions (1/2/3), components (e.g. Read story, Sentences, MapTask, etc.) and items. At each level, it is possible to view the demographic information (age, gender, education, socio-economic status) for that speaker.


>> Search

Site	Parts.	M	F	Age (< 30, 31-49, > 50)	SES (Prof-M, NProf-M, Prof-F, NProf-F)	Recs (1, 2, 3a, 3b)	
Australian National University, Canberra	49	24	25	13, 19, 17	18, 6, 19, 6	49, 49, 25, 24	QA
Charles Darwin University, Alice Springs	0	0	0	0, 0, 0	0, 0, 0, 0	0, 0, 0, 0	QA
Charles Darwin University, Darwin	43	19	24	12, 20, 11	15, 4, 14, 10	26, 31, 16, 16	QA
Charles Sturt University, Bathurst	47	24	23	14, 14, 19	18, 6, 15, 8	46, 42, 19, 19	QA
University of Melbourne, Castlemaine	28	6	22	26, 1, 1	1, 5, 1, 21	28, 26, 13, 12	QA
Flinders University, Adelaide	108	41	67	38, 30, 40	22, 19, 40, 27	108, 94, 46, 44	QA
University of Queensland, Townsville	31	3	28	21, 5, 5	2, 1, 4, 24	23, 17, 15, 13	QA
University of Canberra, Canberra	102	76	26	38, 46, 18	22, 54, 14, 12	64, 62, 28, 26	QA
University of Melbourne, Melbourne	119	52	67	32, 45, 42	36, 16, 48, 19	117, 117, 54, 55	QA
University of New England, Armidale	45	14	31	20, 15, 10	8, 6, 10, 21	45, 43, 21, 21	QA
University of New South Wales, Sydney	86	40	46	44, 22, 20	16, 24, 23, 23	48, 45, 21, 21	QA
University of Queensland, Brisbane	86	36	50	19, 18, 49	28, 8, 39, 11	75, 73, 31, 31	QA
University of Sydney, Sydney	65	33	32	19, 22, 24	22, 11, 26, 6	63, 64, 32, 32	QA
University of Tasmania, Hobart	48	24	24	12, 19, 17	18, 6, 20, 4	48, 48, 24, 24	QA
University of Western Australia, Perth	96	37	59	43, 24, 29	22, 15, 39, 20	96, 92, 46, 46	QA
University of the Sunshine Coast, Maroochydore	20	7	13	6, 4, 10	4, 3, 7, 6	19, 18, 10, 10	QA
All sites	*973*	*436*	*537*	*357, 304, 312*	*252, 184, 319, 218*	*855, 821, 401, 394*	

Usage: 23,028,108 (+4,501,669) MB in 7,906,621 (+1,485,358) files.

Figure 2: AusTalk Interface

However when viewing AusTalk data through the original Alveo Discovery interface, as shown in Figure 3, all the files are shown at the same level. Although the file name contains information about Speaker ID, session number, component and item, that metadata is not directly available via the facets provided for searching.

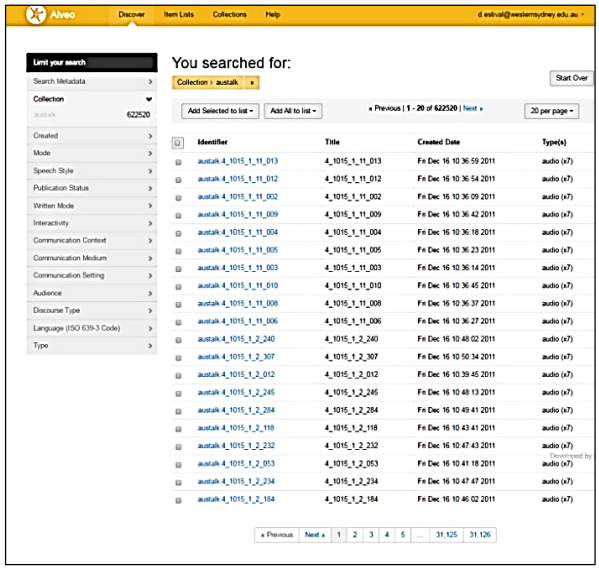


Figure 3: AusTalk through Alveo

[2] In future releases, an AusTalk item will also include one or more text file, the annotations for a phonetic or phonemic transcription of the audio (the expected prompt is currently available in the metadata).

As that information is part of the metadata, it is possible to filter the data according to Speaker, Session, Component and Item but this requires more complex search queries through the Advanced Search facility and specific knowledge about what is available for the corpus. Figure 4 shows the advanced search query that will return all the Sentences for Speaker 1_1308.

```
collection_name:austalk AND componentName:
sentences AND speaker: 1_1308
```

Figure 4: Advanced search query for AusTalk

The demographic information (gender, age, etc.) is not accessible through these queries. Therefore a new Alveo query interface was specifically designed for AusTalk, providing such filter. In Figure 5, we've selected all the female speakers born in Canberra.

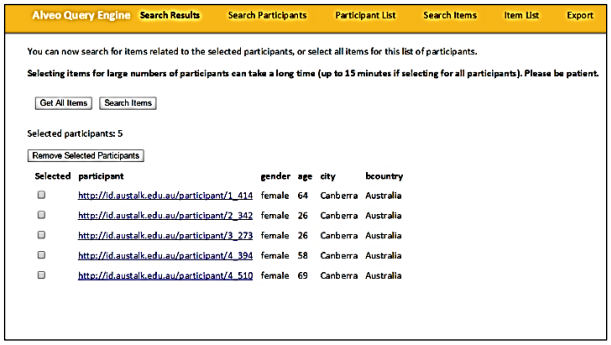


Figure 5: New Alveo search interface for AusTalk

This example points to the main issue, organising data through a hierarchy or via facets. In AusTalk, we may wish to look at the Sentence component in Session 2 for all the female speakers from Adelaide, or the MapTask component in Session 3 for all the male speakers in Melbourne. Such a view is not possible when using the facets provided by the original interface.

The AvCom corpus is another example where hierarchy is important for a dataset (Molesworth & Estival, 2015a). In that corpus (136 audio files, 6GB), each of the 17 pilot participants was recorded during 8 experimental flights in a flight simulator. Thus, we might want to look at all the 8 flights for one pilot or all instances of one experimental flight for the 17 pilots. This would not be possible with the current Alveo interface.

3. Metadata for data upload and ingest

Adding the AvCom corpus highlighted the mirror problem of specifying metadata to be provided by users who want to upload new datasets. There are two ways to add new data to Alveo: (1) with a script specifying the metadata and data to be ingested and (2) via the web user interface recently added. The earlier Alveo datasets were all ingested with scripts specific for each case. Some collections were later added via an Excel spreadsheet with columns specifying certain information about the data and metadata to be ingested, and a script making use of that information. This worked well, in particular for the Liberated Learning Corpus (Bain, Stevens, Martin, & Lund-Lucas, 2012).

A more recent ingest, that of a snapshot of the Trove newspaper archive (Holley, 2010) has shown that, although the size of the dataset itself posed a number of problems, it was possible to make the individual documents available over the web while also providing efficient support for processing large chunks of data (Cassidy, 2016).

The much smaller AvCom dataset presented a different set of difficulties, because of its different metadata. The spreadsheet facility did not work for AvCom, firstly because there was a *Pilot* filed instead of *Speaker*. This may seem trivial, as it is obviously possible to use *Speaker* instead of *Pilot*, but other information of importance for this dataset (e.g. pilot qualification, flight hours, or native language) was not catered for either. Thus a new script needs to be written for ingestion of this dataset via a spreadsheet. The original Alveo interface however will not provide this corpus-specific information unless new facets are introduced.

Figure 6 shows the process of adding the AvCom collection through the new web user interface.

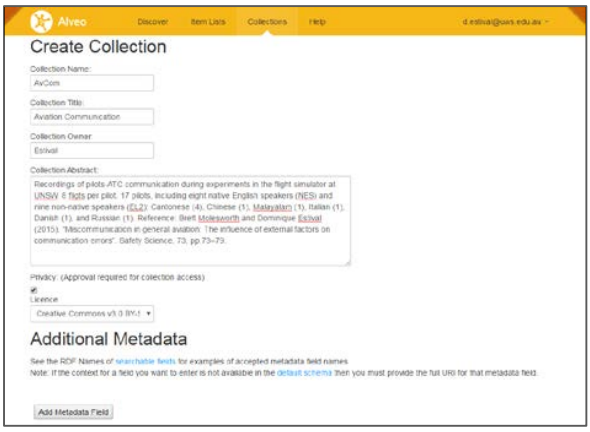


Figure 6: Adding a collection in Alveo

Once a collection has been created, items can be added one at a time, as shown in Figure 7.

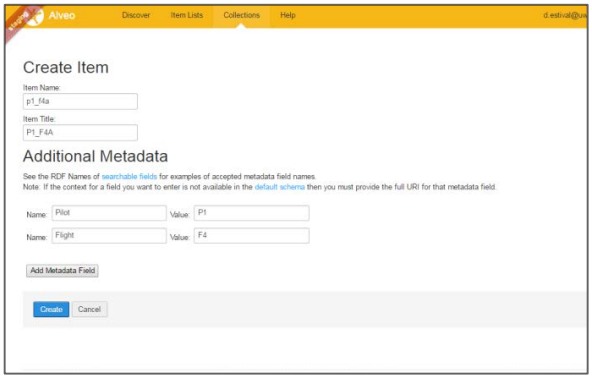


Figure 7: Adding one item

The interface allows the user to specify their own facets (e.g. *Pilot* 1 and *Flight* 4 for item P1_F4 in Figure 7), but as these fields are not yet part of the metadata recognised by the system, they do not appear in the Item details shown in Figure 8.

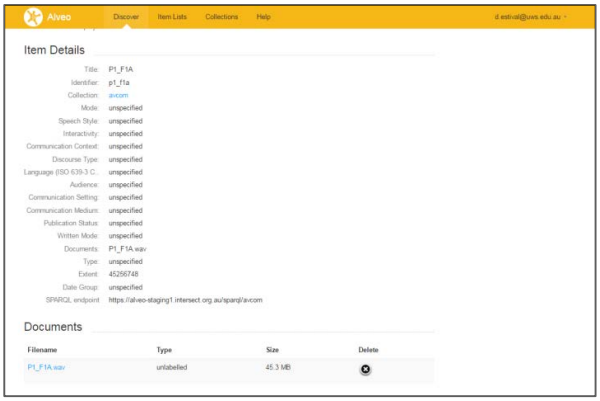


Figure 8: Item Details

The metadata that is created automatically only includes facets which had been considered useful for other datasets (based primarily on AusNC, and extended for AusTalk and PARADISEC). These facets may or may not be appropriate for a new dataset. In addition, the list presented to the user (see Figure 9 for a small selection) is currently difficult to navigate and interpret.

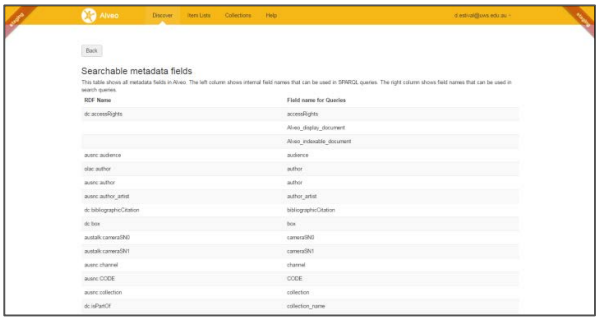


Figure 9: Alveo metadata fields

Thus the user interface would need to be modified in two respects: better navigation of current facets, and addition of new facets.

4. Solutions

At this point, it seems that a new corpus may need a specifically tailored solution for uploading and ingesting, and specific search/viewing interface. This is not a new problem and there are known solutions, e.g. CLARIN (Zastrow, Hinrichs, Hinrichs, & Beck, 2013) or ExMARALDA for spoken corpora (Haugh, Ruhi, Schmidt, & Wörner, 2014). However, they result in a lack of commonality for the search interface which undermines the original goal of unified access to different datasets and the creation of list of items from a variety of datasets for further analysis.

The solutions we currently envisage are:

1. The recently developed Alveo user upload facility, which allows users to upload one item at a time through the web app, lists all the current facets. Even if these were sufficient for a new dataset, the presentation of the allowed facets needs to be improved. In parallel, it would be good to hide the facets that are not useful for a particular corpus in the data display. This work is under investigation.
2. For corpora that are not easily amenable to the current list of facets, we can provide an interface tailored for that corpus (as is already done for AusTalk), but this means different views and different access methods for different datasets.
3. We would like to let users not only specify new facets in the upload interface (as is already possible, see Fig.7) but to have those facets appear in the data display. This would require implementing the hiding of unnecessary facets (1 above), since otherwise the screen would be too cluttered and difficult to navigate.
4. Finally, we need a new spreadsheet API connection which will let the researcher specify the facets to use as columns for ingestion of a whole dataset. This work is currently in progress.

In conclusion, the problems described in this paper are not novel, but the specific examples which are problematic for Alveo show that possible solutions detract from the original intention of the platform. Both the search interface and the upload utility are subject to constraints imposed by each dataset, and possibly by each intended research use. The issue is a more general one, which is probably common to many systems. Human Communication Science provides a restricted ontology of facets (e.g. *author*, *date_of_recording*, *composer*, *depositor*, etc.) which may seem to be adequate for most purposes and data collections but which turns out, unsurprisingly, to be at the same time too detailed (*fathers_place_of_birth*) and not sufficient (e.g. for AvCom).

5. Acknowledgements

The Alveo Virtual Lab acknowledges funding from Nectar, which is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS). Alveo also received support from 13 Australian universities: Western Sydney University, Macquarie University, the Australian National University, University of Canberra, Flinders University, University of Melbourne, University of Sydney, University of Tasmania, University of New South Wales, University of Western Australia, RMIT, University of New England, Latrobe University; and 3 organisations: NICTA (National ICT Australia), ASSTA (Australasian Speech Science and Technology Association) and AusNC (Australian National Corpus).

6. Bibliographical References

Bain, Keith, Basson, Sarah H., Faisman, A., & Kanevsky, D. (2005). Accessibility, transcription, and access

everywhere. *IBM Systems Journal, 44*(3), 589-603. doi:10.1147/sj.443.0589

Bain, Keith, Stevens, Janice, Martin, Heather, & Lund-Lucas, Eunice. (2012). *Transcribe your class: Empowering students, instructors, and institutions: Factors affecting implementation and adoption of a hosted transcription service* Paper presented at the INTED2012.

Burnham, Denis, Estival, Dominique, Fazio, Steven, Cox, Felicity, Dale, Robert, Viethen, Jette, . . . Wagner, Michael. (2011). *Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box*. Paper presented at the Interspeech 2011, Florence, Italy.

Cassidy, Steve. (2016). *Publishing the Trove Newspaper Corpus*. Paper presented at the LREC 2016.

Cassidy, Steve, Estival, Dominique, Jones, Timothy, Burnham, Denis, & Berghold, Jared. (2014). *The Alveo Virtual Laboratory: A Web Based Repository API*. Paper presented at the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland.

Cassidy, Steve, Haugh, Michael, Peters, Pam, & Fallu, Mark. (2012). *The Australian National Corpus : national infrastructure for language resources*. Paper presented at the LREC.

Estival, Dominique, Cassidy, Steve, Cox, Felicity, & Burnham, Denis. (2014). *AusTalk: an audio-visual corpus of Australian English*. Paper presented at the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland.

Estival, Dominique, Cassidy, Steve, Sefton, Peter, & Burnham, Denis. (2013). *The Human Communication Science Virtual Lab*. Paper presented at the 7th eResearch Australasia Conference, Brisbane, Australia.

Estival, Dominique, Cassidy, Steve, Verspoor, Karin, MacKinlay, Andrew, & Burnham, Denis. (2014). *Integrating UIMA with Alveo, a human communication science virtual laboratory*. Paper presented at the Workshop on Open Infrastructures and Analysis Frameworks for HLT, COLING 2014, Dublin, Ireland.

Goecke, Roland, & Millar, J.B. (2004). *The Audio-Video Australian English Speech Data Corpus AVOZES*. Paper presented at the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP), Jeju, Korea.

Goecks, Jeremy, Nekrutenko, Anton, Taylor, James, & Team, The Galaxy. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology, 11*(8), R86.

Haugh, Michael, Ruhi, Şükriye, Schmidt, Thomas, & Wörner, Kai. (2014). Introduction: Putting practices in spoken corpora into focus. In Michael Haugh, Şükriye Ruhi, Thomas Schmidt, & Kai Wörner (Eds.), *Best Practices for Speech Corpora in Linguistic Research* (pp. 1-19): Cambridge Scholars Publishing.

Holley, Rose. (2010). Trove: Innovation in access to information in Australia. *Ariadne, 64*.

Molesworth, Brett R. C., & Estival, Dominique. (2015a). Miscommunication in general aviation: The influence of external factors on communication errors. *Safety Science, 73*, 73-79.

Musgrave, Simon, & Haugh, Michael. (2009). *The AusNC Project: Plans, Progress and Implications for Language Technology*. Paper presented at the ALTA 2009, Sydney.

Rodriguez-Castro, Bene, Glaser, Hugh, & Carr, Les. (2010). *How to Reuse a Faceted Classification and Put it on the Semantic Web*. Paper presented at the The 9th International Semantic Web Conference (ISWC), Shanghai, China.

Thieberger, Nick, Barwick, Linda, Billington, Rosey, & Vaughan, Jill (Eds.). (2011). *Sustainable data from digital research: Humanities perspectives on digital scholarship. A PARADISEC Conference*: Custom Book Centre. http://ses.library.usyd.edu.au/handle/2123/7890.

Zastrow, Thomas, Hinrichs, Erhard, Hinrichs, Marie, & Beck, Kathrin. (2013). *Scientific Visualization as CLARIN-D Web Applications*. Paper presented at the Digital Humanities 2013.