The Human Communication Science Virtual Lab: A repository microclimate in a rapidly evolving research-ecosystem

Peter Sefton <p.sefton@uws.edu.au>, Dominique Estival <D.Estival@uws.edu.au >, Steve Cassidy < steve.cassidy@mq.edu.au>, Denis Burnham <Denis.Burnham@uws.edu.au>, Jared Berghold <jared.berghold@intersect.org.au>

Abstract

The Human Communication Science Virtual Lab (HCS vLab) represents a new kind of data driven research collaboration environment which has at its heart a repository of human communication data, drawn from research collections in a broad range of fields including speech science, speech technology, computer science, language technology, behavioural science, linguistics, music science, phonetics, phonology, and sonics and acoustics. The repository contains text, audio and video data as well as annotations which describe the data, with a discovery search/browse interface. But this is not a just a repository project - the main function of the lab is to provide a platform for doing things with data in ways that were previously difficult, via a rich programming interface (API), and via a workflow engine which allows a large cross-disciplinary research community to run combinations of tools on the data held in the repository in a reproducible way. The presentation will cover vLab's genesis as an Australian Government funded project consisting of a dozen universities and institutions led by the University of Western Sydney, with development provided by Intersect NSW. We will provide a walk-through of the functionality of the lab including showing how familiar repository functionality such as search-and-browse is linked to the creation of stable, citable collections known as "item lists", and how item lists can be processed and analysed in various ways using linguistic parsers, acoustic and phonetic analysis, and visual processing and how this fits in to the research lifecycle including how data and publications will be linked and cited, and explore the relationship between the lab and other scholarly infrastructure such as institutional repositories.

Background & Goals

The vLab[1,2] project was funded by NeCTAR, which is one of a number Australian Government investments in eResearch infrastructure over the last decade, following on from a number of repository programs[3,4], familiar to the Open Repositories audience, and the Australian National Data Service (ANDS)[5], which is well on the way to establishing research data management infrastructure including research data catalogues (AKA metadata stores) and/or research data repositories as business-as-usual systems for all universities and many other research institutions in the country. On the content side, the HCS vLab follows on from Australian National Corpus Project[6], funded by ANDS which established the core repository framework and collected the first tranche of language data to be included in the vLab. AusNC itself is one of many projects that emerged from the ARC-funded Human Communication Science Network[7] involved over 1000 Australian researchers across the disciplines mentioned above. The HCS vLab will both continue that work and take it to a new virtual accessible level that will further engender research collaboration.

The aim of the lab is to make data-driven research using a wide variety of tools more accessible to a broader range of researchers, both with and without technical skills, to facilitate new kinds of research, involving novel tool-data combinations to take place. Walters[8] surveys similar work, in "Assimilating Digital Repositories Into the Active Research Process" which is an apt title, the vLab did not grow out of a library-based publications repository and attempt to add services to it; rather, the vLab was built around a

framework for connecting data to code part of a research workflow and brought in a repository to effect this.

Architecture

The lab has a data repository at its heart, but as already pointed out, the primary aim of the lab is not to archive data, or to make it discoverable, it is to make data usable and re-usable for research, which requires that there be discovery and archiving services. This is accomplished via two main avenues, firstly via the web based API which allows services to be built on top of the core repository, and secondly via a web-based drag and drop workflow engine. The chosen workflow engine is the Galaxy system[9–11], which originated as a bio-informatics analysis platform but is now being used in a variety of other contexts, including as an image analysis tool in other NeCTAR projects[12].

Repository

The vLab has a repository component, built using the Hydra/Fedora[13] application framework. The repository has:

- *Collections (corpora) of items*, where an item is defined as a communicative event, e.g., a single text, conversation, utterance or musical performance. As with the kinds of repositories familiar to an Open Repositories audience, items can have multiple data streams representing different versions or aspects of the event, including text versions. Unlike typical institutional publications repositories the primary kind of *text* object is a plain-text stream with all formatting and analytical annotation stored as stand-off annotations.
- *Per-collection access control.* Many of the data sets have been collected from human participants under a variety of consent agreements requiring researchers to agree to a click-through license, or go through an offline process to have access approved.
- *Multiple indexes,* An Apache Solr text index and an RDF triple store are supplied as core functionality, but the intention is that other indexing services will be able to be added. The first of these is is being built using the INDRI[14] information retrieval engine; INDRI is in fact one of the research tools that has been integrated into the platform, it is used by researchers in information retrieval to experiment with new search engine architectures and features. It is now being used to build an advanced search tool for the data stored on the vLab. We expect that other search services will be implemented in future.
- *Annotations stored with items,* a core facility is the storage and management of *standoff annotations* on the data stored in the vLab. Annotations are attached to points or regions in the text, audio or video data (represented as start/end points) and may contain a simple label or a complex feature structure. Example annotations might mark a speaker turn in a dialogue, a headline in a text file or a individual phoneme in a speech recording. Researchers generate annotations manually and automatically and there are many diverse file formats and tools that work with them. The vLab implements the DADA Annotation Store[15] which represents annotations in RDF and provides a general model that unifies many styles of annotation. The API currently exposes a basic feed of annotations for each item. In the future, we expect to build comprehensive annotation query services on top of this annotation store.
- *Simple integration*: the discovery interface offers an innovative one-click service to ‘use this item-list in a workflow’ and copy-and-paste code that researchers can use to access the lab in a growing number of programming languages

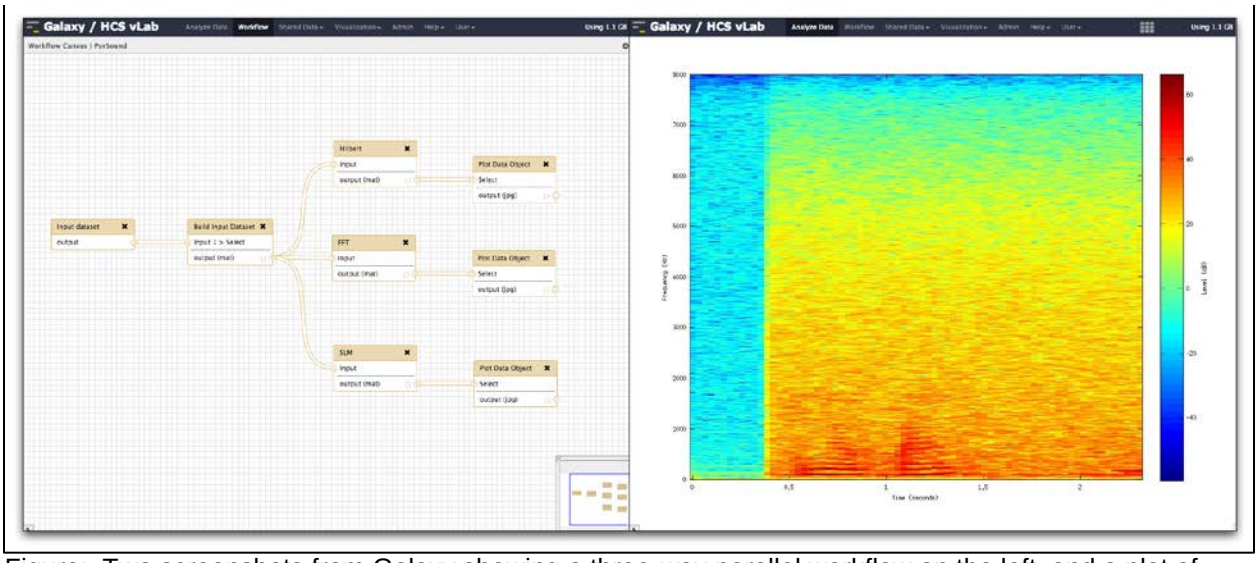

Figure: Two screenshots from Galaxy showing a three-way parallel workflow on the left, and a plot of acoustic data on the right.

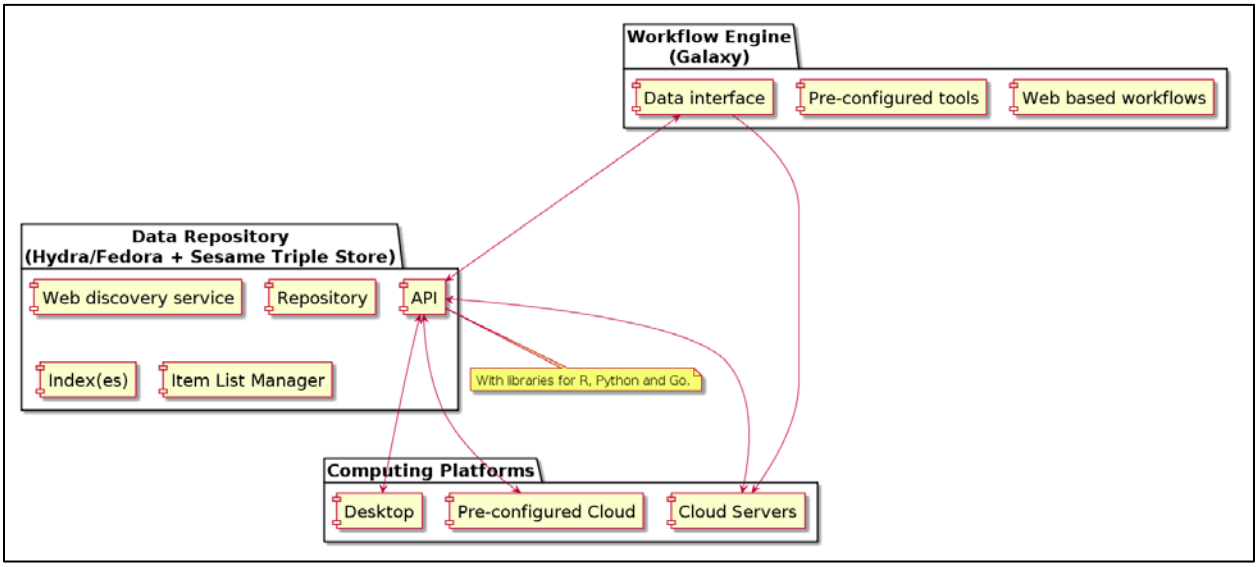

Figure: The HCS vLab architecture

Architectural issues

Using the Fedora repository and Hydra toolkit has one major advantage; the project team were able to establish very quickly a discovery interface with rich faceted search; however there have been some problems; the major one being the Fedora stack is designed around an environment where users access one repository item at a time, usually via a web interface. Hydra had poor performance on batch-access to item-lists as there is a significant overhead in the software stack for accessing an individual item. This has meant that we have had to build special purpose SOLR indices to improve access times for the core data, extending the common technique of supplementing Fedora's built in services with external index-based services.

What will vLab-enabled research look like?

Phase one of the HCS vLlab project is running between early 2013 and mid 2014, and by June 2014 version 3 will have been completed and launched. The second two-year phase will involve further development but also a much gr eater emphasis on promoting the use of the service, with a view to having research articles that use the lab, its data and tools, and successful grant applications that incorporate use of the vLab. This will involve a process of evolving ways to:

- Cite item lists as reusable data.
- Cite the products of particular processes (code, Galaxy workflows) which have been run on item lists.

One of the key goals of the vLab is to allow reproducible research and re-runnable research processes and workflows. To enable this, it is important that there are stable, referenceable entities that can be re-processed and re-used. We have yet to solve all the challenges involved in preserving item-lists exactly, and maintaining access-control.

References

1. Estival, D., Cassidy, S., Sefton, P. & Burnham, D. The Human Communication Science Virtual Lab. in (2013). at <http://eresearchau.files.wordpress.com/2013/08/eresau2013_submission_16-2.pdf>
2. Burnham, D. Above and Beyond Speech, Language and Music: A Virtual Lab for Human Communication Science (HCS vLab. (2012).
3. ARROW Project. Australian Research Repositories Online to the World. (2007). at <http://www.arrow.edu.au/>
4. APSR Project. Australian Partnership for Sustainable Repositories. (2007). at <http://www.apsr.edu.au/>
5. Sandland, R. *Introduction to ANDS*. (ANDS, 2009). at <http://ands.org.au/newsletters/newsletter-2009-07.pdf>
6. Cassidy, S., Michael, H., Pam, P. & Mark, F. The Australian National Corpus : national infrastructure for language resources. in *Proc. LREC 2012* (2012).
7. Dale, R., Burnham, D. & Stevens, C. J. (2004-2008) The Human Communication Science Network (HCSNet): Enabling Human Communication - Tough problems in human communication with bold but informed solutions drawing on sound, speech, and language research capabilities. (2004).
8. Walters, T. Assimilating Digital Repositories Into the Active Research Process. *Res. Data Manag. Pract. Strateg. Inf. Prof. Charlest. Insights Libr. Inf. Arch. Sci.* 189 (2014).
9. Goecks, J., Nekrutenko, A., Taylor, J. & Team, T. G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11,** R86 (2010).
10. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15,** 1451–1455 (2005).
11. Blankenberg, D. *et al.* Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Curr. Protoc. Mol. Biol.* 19–10 (2010).
12. Virtual Laboratories. at <https://www.nectar.org.au/virtual-laboratories-1>
13. Awre, C. *et al.* Project Hydra: Designing & Building a Reusable Framework for Multipurpose, Multifunction, Multi-institutional Repository-Powered Solutions. (2009). at <http://smartech.gatech.edu/handle/1853/28496>
14. Strohman, T., Metzler, D., Turtle, H. & Croft, W. B. Indri: A language model-based search engine for complex queries. in *Proc. Int. Conf. Intell. Anal.* **2,** 2–6 (Citeseer, 2005).
15. Cassidy, S. An RDF Realisation of LAF in the DADA Annotation. in *Serv. Proc. ISA-5 Hong Kong* (2010).