**BERGHOLD — Alveo**

<0 figures>

**Alveo: Software Engineering with Humanities Researchers**

Berghold, J., Hammond, J., Cassidy, S., Estival, D., Burnham, D. and Sefton, P.

This paper explores the interactions between software engineers and researchers in developing large-scale digital humanities platforms. In particular, we focus on how a diverse, cross-discipline group contributed to the development and deployment of the Alveo1 virtual laboratory (Cassidy et al., 2014). We present an overview of how the tool works, its impact through transforming methodologies within the language sciences, and a summary of what digital technologies underpin its framework. We show the process of the Agile methodology2 in a humanities context, good practice for scoping and implementation of highly technical features, and how stakeholder consultation is key to successful software development processes.

Alveo is an online system designed specifically for research in Human Communication Science (HCS). Human Communication Science encompasses the areas of speech science, speech technology, computer science, language technology, behavioural science, linguistics, music science, phonetics, phonology, and sonics and acoustics. HCS research depends upon datasets (collections) of speech, music, text, video, sounds, and specialised tools by which to search, analyse, and annotate these data. Alveo provides a platform to access these datasets and use the specialised tools. The Alveo system consists of two major components: a data discovery interface and workflow engine. These work independently, but are compatible so that one feeds into the other.

The data discovery interface gives both a Graphical User Interface (GUI) and an Application Programming Interface (API) that allow for browsing and searching of data, viewing collections and their metadata, and previewing and downloading documents (text documents, audio files, video files, images, etc.). The data discovery interface supports the creation of stable ‘item lists’ that can be shared with other users and transferred to third-party tools for analysis. The workflow engine uses Galaxy,3 a scientific workflow platform originally developed by Penn State University for data-intensive biomedical research. Alveo uses Galaxy to provide its users with easy access to a variety of tools for analysing and manipulating human communication datasets. The innovation of the Galaxy component is that it allows for workflow construction independent of data. For example, a researcher builds a workflow that uses PsySound for an acoustic analysis combined with an independent (but linked) analysis using the ParseEval tool on the same language data. The workflow can then be reused on different collections (or selected subsets), shared with collaborators or archived for publication.

The technical complexity of the project is seen in the range of components that underpin the system. These include: Hydra framework comprised of Blacklight4 to provide a discovery interface and Apache Solr5 for search; JSON (JavaScript Object Notation) and JSON-LD formatted metadata to build human-readable text to transmit data; and Sesame framework6 for storing and accessing metadata that is compliant with the Resource Description Framework (RDF) specifications. A core part of the engineering contribution was to evaluate these technologies, and provide advice and consultancy so that outcomes of the project were achieved with maximal functionality but without undue overheads.

During the development of the digital laboratory we utilised work practices adhering to the Agile software development process. Specifically, we used the Scrum software development framework.7 Scrum is an iterative and incremental process that encourages close collaboration with clients, daily face-to-face communication between team members, and a flexible approach to changing requirements. During the development phase of Alveo, team members met with the major stakeholders on a weekly basis. Meetings alternated between demonstrating the latest version of the system (resulting from the previous two-week period of development) and planning work for the next two weeks. Ad hoc communication with stakeholders took place on an almost daily basis via email, video conference, and an online issue tracking system. By collaborating closely with stakeholders during the development process, there was a greater likelihood that the resulting software system would meet the needs and expectations of those stakeholders.

We argue that large digital projects have better outcomes when software engineers work not *for* but *with* humanities scholars (Teehan and Keating, 2010; Dombrowski, 2014; Bradley, 2009). The contribution of software engineers to digital humanities projects is not only to implement a digital or computational solution to a problem but to also be highly engaged with the discipline, to introduce options and advice, and to react rapidly and flexibly to changing requirements at all stages along a project timeline.

**Notes**

1. http://alveo.edu.au/.

2. http://agilemethodology.org/.

3. https://usegalaxy.org/.

4. http://projectblacklight.org/.

5. http://lucene.apache.org/solr/.

6. http://rdf4j.org/.

7. https://www.scrum.org/Resources/What-is-Scrum.

**References**

**Bradley, J.** (2009). What the Developer Saw: An Outsider’s View of Annotation, Interpretation, and Scholarship. *New Paths for Computing Humanists,* **1**(1).

**Cassidy, S., Estival, D., Jones, T., Burnham, D. and Berghold, J.** (2014). The Alveo Virtual Laboratory: A Web-Based Repository API. *9th Language Resources and Evaluation Conference (LREC 2014)*,Reykjavik, Iceland, 26–31 May 2014.

**Dombrowski, Q.** (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing,* **29**(3): 326–39, doi:10.1093/llc/fqu026.

**Teehan, A. and Keating, J. G.** (2010). Appropriate Use Case Modeling for Humanities Documents. *Literary and Linguistic Computing,* **25**(4): 381–91, doi:10.1093/llc/fqq026.