Annotation Contributions: sharing derived research data

Steve Cassidy[*], **Dominique Estival**[†],
[*]Macquarie University, Sydney, Australia,
steve.cassidy@mq.edu.au
[†]Western Sydney University, Sydney, Australia,
d.estival@westernsydney.edu.au

Abstract

Many research projects make use of language resources and in the process of the research, generate derived resources such as annotations or derived features. These resources are not often shared and there is no mechanism for associating them with the original resource in a way that would help future users discover them. This paper describes a new feature of the Alveo Virtual Laboratory that allows researchers to upload these derived resources and have them associated and integrated with the original resource. This facility is designed to encourage the sharing of these derived resources by making the process of sharing easy and by providing an incentive in the form of a citeable URL that can be properly attributed to the authors in subsequent publications.

Keywords:

1. Introduction

Projects making use of shared language resources often create new artifacts as a by-product of the main research program. These may be annotations of a particular linguistic structure or derived signals such as pitch tracks, translations or summaries of documents. These new resources are often not shared or, if they are, are made available as a separate download not associated with the original language resource. This paper describes a new feature of the Alveo Virtual Laboratory that allows researchers to upload these artifacts and have them associated with the original data they were derived from. These *contributions* are then made available both as a separate resource and linked to the original resource.

The Alveo Virtual Laboratory[1] (Cassidy et al., 2014) combines a data repository with a web-based API and a workflow platform and aims to provide access to data in a way that may enhance reproducibility of research results (Cassidy and Estival, 2017). Data in Alveo is made available as a *collection* of *items* with associated meta-data and one or more *documents* (audio, video, text, annotations, etc). Users can query the data store for items relevant to a particular study, creating an *item list*; this list can then be fed into a data processing pipeline either by downloading data as a zip file or writing scripts against the API that access documents directly. Each item list has a unique URL and can be shared such that other researchers can access the same items (subject to access restrictions) and hence use these in reproducing or extending the original work.

Meta-data in Alveo can be associated with collections, items and documents. The system does not mandate a fixed schema but supports a range of existing vocabularies (OLAC, DCTERMS, etc.). Until recently, all of the data ingested into Alveo has been legacy collections and so the decision was taken to accept any meta-data that was available while providing a mapping to common properties where possible. As a consequence, the Alveo meta-data store is flexible and able to incorporate any meta-data that might be available for a collection (Estival, 2016).

1.1. Sharing Derived Resources

Annotations and other derived resources are often produced as a side-effect of research by language researchers but we are not aware of any repository which accepts resources like these for distribution that associates them in a useful way with the original resource.

As an example of the current state of play in distribution and sharing of annotation we can look at the popular Switchboard corpus published by LDC (Godfrey and Holliman, 1993). The LDC catalog page describing this resource includes a link to the Switchboard Dialog Act Corpus[2], a separate collection of annotations that is distributed as a download from the documentation directory of the corpus on the LDC site. In addition, LDC also publishes the NXT Switchboard Annotations (Calhoun et al., 2009) which combines a number of layers of annotation using the NXT format developed in a collaboration among researchers from Edinburgh University, Stanford University and the University of Washington. On Github we find the *Switchboard Dialog Act Corpus with Penn Treebank links*[3] which builds on the earlier dialog act annotations.

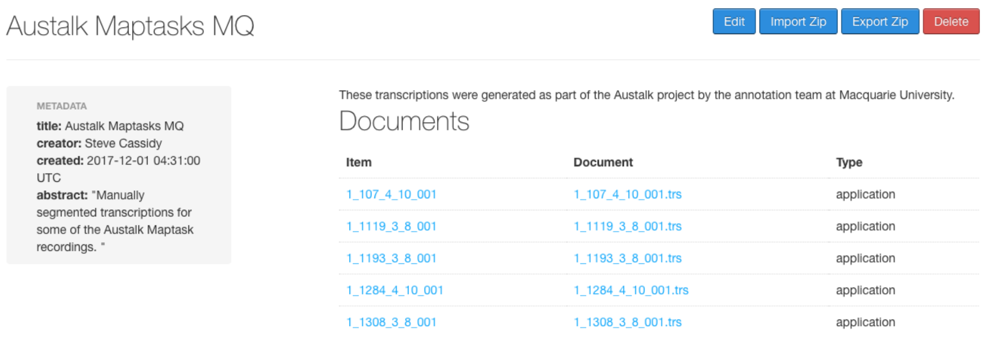
The good part of this story is that we were able to find all of these resources within a few minutes with a few web searches; at least these resources are available on the web and the links are maintained to some degree. However, we can be reasonably sure that there are additional annotations over this data (or corrections to some of these annotations) that are not turning up in our searches - for instance, data that is shared on institutional servers that have a lower page-rank and data that was never shared because the project ran out of funds or had no incentive to publish their data.

There is an opportunity, then, to improve the way that derived resources are published and shared. Language resource providers who make the original datasets available

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

[2] http://compprag.christopherpotts.net/swda.html

[3] https://github.com/cgpotts/swda



Austalk Maptasks MQ

METADATA
title: Austalk Maptasks MQ
creator: Steve Cassidy
created: 2017-12-01 04:31:00 UTC
abstract: "Manually segmented transcriptions for some of the Austalk Maptask recordings. "

These transcriptions were generated as part of the Austalk project by the annotation team at Macquarie University.

Documents

Item	Document	Type
1_107_4_10_001	1_107_4_10_001.trs	application
1_1119_3_8_001	1_1119_3_8_001.trs	application
1_1193_3_8_001	1_1193_3_8_001.trs	application
1_1284_4_10_001	1_1284_4_10_001.trs	application
1_1308_3_8_001	1_1308_3_8_001.trs	application

Figure 1: The page describing a contribution.

could develop a way for derived data to be associated with the original resource in such a way that they can be discovered easily by researchers who find the original data.

While most language data archives make collections available as downloadable archives (zip files or similar), Alveo is a more fine-grained store that exposes individual items and documents for discovery and download. Part of the motivation for this is that researchers often only need to use part of a collection in a study; Alveo supports identifying this subset as an *item list* and downloading just that subset of data. This means that we have the opportunity to associate derived resources at this fine-grained level as well. The goal of the work described in this paper is to provide a means for derived resources to be shared on the Alveo platform in such a way that they can be discovered naturally by researchers browsing the original data.

2. Annotation Contributions

The Alveo system has recently been extended to allow users to upload files derived from existing resources and have them associated with the original data they were derived from. This can be illustrated with the following scenario.

A researcher is studying Australian English vowels and identifies 10 speakers from the Austalk corpus, finding the items corresponding to their reading of the 18 hVd words (hid, had, hod, etc). They create an *item list* for these items and download the audio files associated with each. Using a forced aligner they derive TextGrid annotations for each file and hand-correct these to ensure that the vowel boundaries are correctly placed. They then derive formant tracks for each recording and carefully check that the formants are correct, some values are hand-corrected if they have been mis-tracked. This data is then used to derive vowel plots for each of the speakers and the results are written up with reference to speaker meta-data that was downloaded with the original data.

Prior to publication of the study, the researchers want to make the corrected TextGrid and formant files available so that they can be referenced in the paper and made accessible for future researchers. In the Alveo system they create a new *Annotation Contribution* and enter some basic metadata and a description of the methodology used to create the derived files. The files are then uploaded to Alveo as a zip file. The system unpacks the archive and based on the filenames, locates the original items that they were derived from (e.g. the original WAV file `2_205_1_3_001.wav` was associated with `2_205_1_3_001.TextGrid` and `2_205_1_3_001.fms`).

The new contribution is accessible via a URL which shows a page containing the meta-data and description of the contribution and a list of the items and associated files. From this page, all of the files can be downloaded as a zip file, but the page also shows direct links to the items referenced by the contribution (Figure 2).

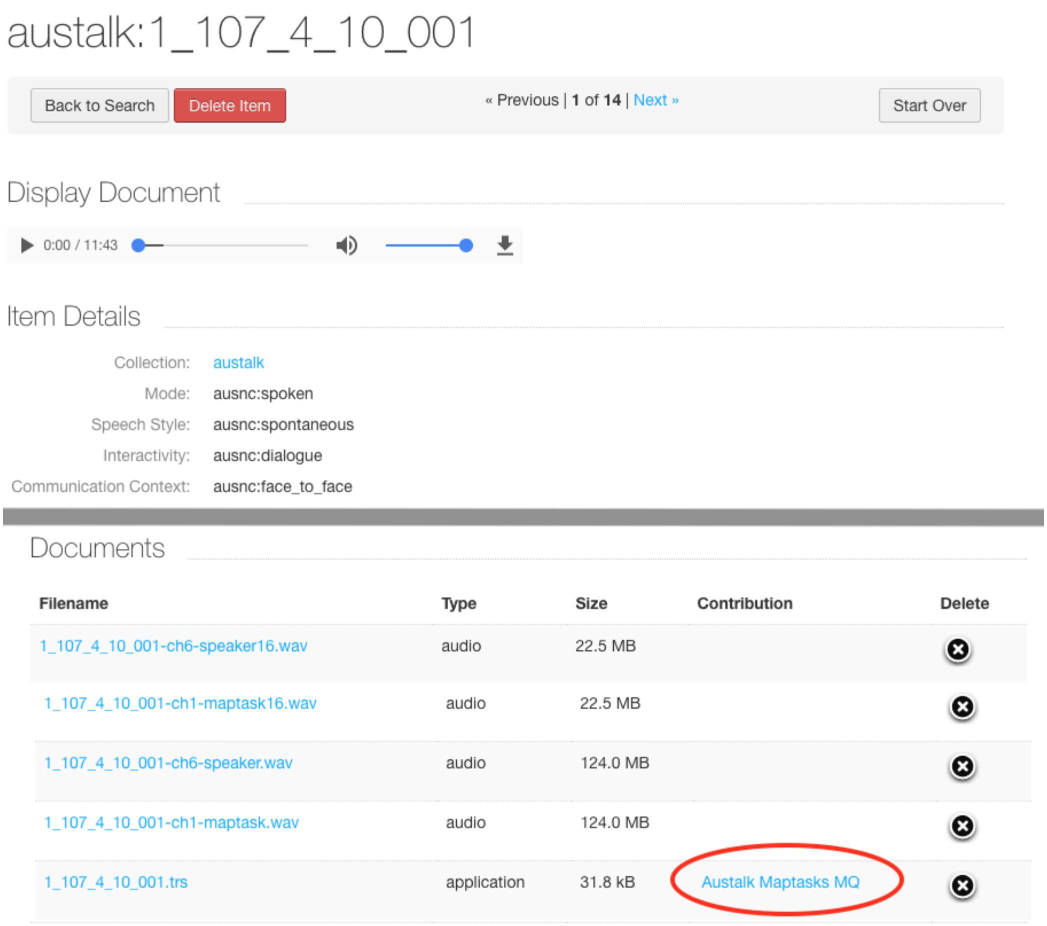
The contribution URL is a public page that can be referenced by any web user and hence is suitable for publication in work that references the resources. Users who are not registered with Alveo can see the meta-data and description of the contribution but will not be able to view the associated data or download files without registering and agreeing to the license of the original collection. At this time we do not have any provision for adding an additional license to the contents of the collection. This is something that might be considered in the future.

We are currently able to issue a DOI for *collections* in the Alveo system to facilitate citation of language resources. Since an annotation contribution is an identifiable resource within the system with a unique URL it would also be possible to issue a DOI for a contribution. This is something we will consider as users adopt this new feature.

2.1. Reproducibility

Together with the other facilities of the Alveo platform, this new feature supports a research workflow that can provide for enhanced reproducibility of research outcomes. As described above, the Alveo platform allows the researcher to create an *item list* with the items that are used in an analysis. This list can be shared publicly and cited in the published research. This new feature then allows the derived resources to be shared in the same way, meaning that all of the data that feeds into the analysis in the research can be cited and is available to future researchers. Further, Alveo provides the Galaxy workflow engine (Goecks et al., 2010) as a platform for constructing and running data processing workflows over speech and language data. Galaxy workflows can also be published and cited in such a way that future researchers can reconstruct the exact sequence of the same versions of tools that were used to generate the published results.

However, straight reproduction of results is not the only de-



austalk:1_107_4_10_001

Display Document

Item Details

Collection:	austalk
Mode:	ausnc:spoken
Speech Style:	ausnc:spontaneous
Interactivity:	ausnc:dialogue
Communication Context:	ausnc:face_to_face

Documents

Filename	Type	Size	Contribution	Delete
1_107_4_10_001-ch6-speaker16.wav	audio	22.5 MB		⊗
1_107_4_10_001-ch1-maptask16.wav	audio	22.5 MB		⊗
1_107_4_10_001-ch6-speaker.wav	audio	124.0 MB		⊗
1_107_4_10_001-ch1-maptask.wav	audio	124.0 MB		⊗
1_107_4_10_001.trs	application	31.8 kB	Austalk Maptasks MQ	⊗

Figure 2: An item page showing the contributed file linking back to the contribution page.

sirable outcome. The publication of derived resources from a research project will also allow future researchers to use this work as a starting point for further investigations. This might be to build on the earlier results or to compare them with results on other data sets. Without easy access to these derived resources, our ability to compare future work with past performance is limited by the detail that is provided in published research procedures. Access to derived resources removes a significant source of variability and can save a lot of effort in repeating earlier analysis.

3. Contributions API

An important part of the Alveo system is the web-based API that provides access to both meta-data and data stored in our collections (Cassidy et al., 2014). Using the API, one can create interfaces to tools for searching and analysing data stored on Alveo. The API has been extended to cover operations on contributions so that external scripts can be written to manipulate them (Table 1).
Using this API, users can write tools as part of an automated annotation workflow to upload the resulting annotation files as part of a new or existing contribution.

4. Provenance Meta-data

The current contribution creation form only provides a very basic set of meta-data fields for the user to complete. The API is able to accept any meta-data fields in the form of JSON-LD formatted properties and values; as mentioned earlier, the Alveo system is able to store arbitrary meta-data structures associated with documents, items, collec-

`/contrib/`	GET	Get the current list of contributions
`/contrib/`	POST	Create a new contribution from a JSON meta-data description
`/contrib/<id>`	GET	Get the JSON description of a contribution including metadata and a list of document URLs
`/contrib/<id>`	PUT	Update the meta-data for a contribution
`/contrib/<id>`	POST	Add documents to a contribution from a zip file
`/contrib/<id>`	DELETE	Remove a contribution and all associated documents

Table 1: A summary of the contributions API.

tions and contributions. One of the goals for future development of the contributions feature is to be able to document the workflow used to generate annotations using the PROV-O provenance ontology.

Belhajjame et al. (2015) describe an extension of the PROV-O ontology for describing scientific workflows. The `wfdesc` ontology allows them to describe a workflow, for example, the processing tools involved and the parameters that they take. The `wfprov` ontology describes the provenance of research artifacts - the execution of workflow steps, inputs and outputs and the parameter settings used in that particular run. These might be used in combination to describe a workflow and the execution settings that generated a set of outputs.

```
{
  "@context": {
     "wfprov":
   "http://purl.org/wf4ever/wfprov#",
     ...
  },
  "@id": "<uri of contribution>",
  "@type": "wfprov:Artifact",
  "wfprov:wasOutputFrom": {
    "@id": "forrestRun13",
    "@type": "wfprov:ProcessRun",
    "wfprov:describedByProcess":
        "toolshed:g_forest/0.01",
    "wfprov:usedInput":
        "<uri of input item list>",
    "wfprov:wasEnactedBy":
        "https://galaxy.alveo.edu.au/",
    "windowShift": 5,
    "windowSize": 20,
    "nominalF1": 500,
    "speakerGender": "Male"
  }
}
```

Figure 3: An example metadata description using the WF-Prov vocabulary derived from PROV-O.

As an example, Figure 3 shows a JSON-LD description of a single processing step that produced a collection of formant tracks using the Emu formant tracker `forest`[4]. The metadata describes the application of the formant tracker via a Galaxy (Goecks et al., 2010) tool (giving the URL of the particular version of the tool repository). In the example, the input is referenced by the URL of an Alveo item list that could be the input to this process; the output is the URL of the contribution itself. Finally, the metadata includes some of the parameter settings used in the execution of this tool.

This kind of metadata could be automatically generated from the provenance data kept by a workflow engine such as Galaxy. Storing this metadata allows a very detailed record to be made of the process used to generate the derived resources. In many cases, the process used to generate derived resources will involve manual steps such as annotation and running interactive tools. An automated capture of the provenance of the output would be difficult in this case, but a user interface for manually entering a structured description following the same format could be built to facilitate recording of this metadata.

5. Adoption by Other Repositories

We have implemented Annotation Contributions in the Alveo system to support the work of researchers working on the data that we hold. The particular design of this feature in our system depends a lot on the other aspects of the system: the individual access to items and documents within collections. However, the idea of encouraging researchers to contribute derived resources back to be associated with the original resource is one that could be adopted by other research data repositories. Even if the data is only available as an archive download, it should be possible to associate new derived datasets with existing collections via their metadata and have them exposed to researchers as they browse the holdings in the repository.

One may go further to develop a standard for linking resources between repositories if the derived resources are, for example, stored in a separate research data store as part of a larger project. As a point of reference, the W3C has recently standardised *Webmentions*[5], a way for websites to notify other sites when their work is mentioned in newly published material. Such a mechanism is built upon the HTTP standard and could be adopted as a means for repositories to notify each other of the availability of derived resources.

6. Summary

This paper has described a new feature of the Alveo Virtual Laboratory that allows researchers to share derived resources generated as part of a research project and have them associated with the original data. This is aimed at improving the sharing of this kind of data that has not been the focus of any other data repository in the past.

7. Acknowledgements

The development of the Alveo Virtual Laboratory was funded by the Australian National eResearch Collaboration Tools and Resources (Nectar) project.

8. Bibliographical References

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J. M., Bechhofer, S., Klyne, G., and Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:16–42, May.

Calhoun, S., Carletta, J., Jurafsky, D., Nissim, M., Ostendorf, M., and Zaenen, A. (2009). NXT Switchboard Annotations LDC2009T26. Web Download, Philadelphia: Linguistic Data Consortium `https://catalog.ldc.upenn.edu/LDC2009T26`.

Cassidy, S. and Estival, D. (2017). Supporting accessibility and reproducibility in language research in the alveo virtual laboratory. *Comput. Speech Lang.*, 45:375–391, September.

[4] `https://github.com/IPS-LMU/wrassp`

[5] `https://www.w3.org/TR/webmention/`

Cassidy, S., Estival, D., Jones, T., Burnham, D., and Burghold, J. (2014). The alveo virtual laboratory: A web based repository API. In Nicoletta Calzolari (conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Estival, D. (2016). Alveo: making data accessible through a unified interface – a pipe-dream? In Richard Eckart de Castilho, et al., editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 5–9, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Godfrey, J. and Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. Web Download, Philadelphia: Linguistic Data Consortium `https://catalog.ldc.upenn.edu/ldc97s62`.

Goecks, J., Nekrutenko, A., Taylor, J., and Team, T. G. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 8(11).