



ALVEO The Human Communication Science Virtual Laboratory

building on HCSNet (an ARC research network)

Presented by: Peter Sefton

Steve Cassidy Dominique Estival* Peter Sefton*, Jared
Berghold*** Denis Burnham***

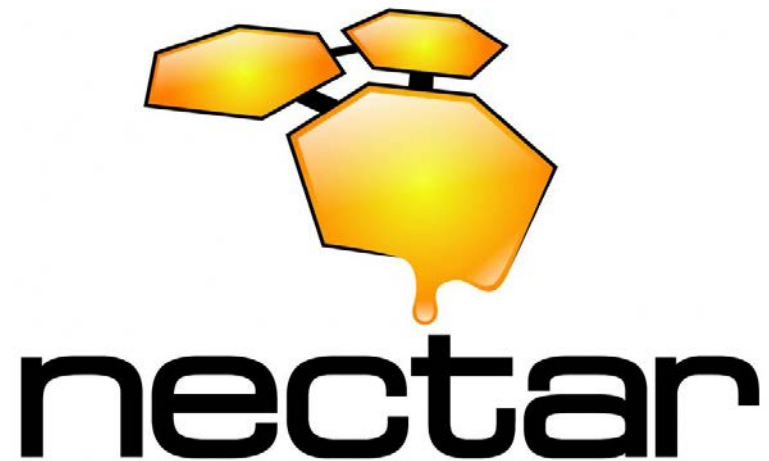
*University of Western Sydney, Australia;

** Macquarie University, Australia;

***Intersect Australia



Funding



Alveo acknowledges funding from the NeCTAR project

<http://www.nectar.org.au> NeCTAR is an Australian Government project conducted as part of the Super Science initiative and financed by the Education Investment Fund.

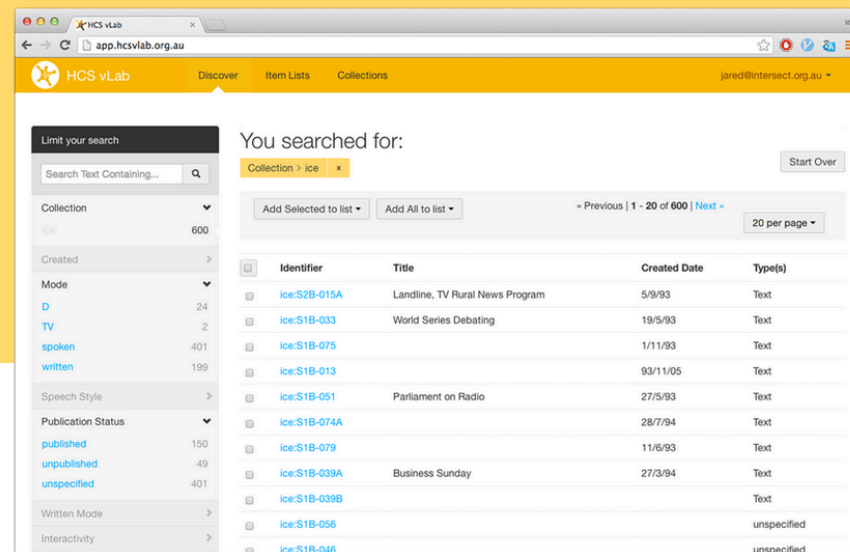


Above and Beyond Speech, Language and Music

A Virtual Lab for Human Communication Science

[Get Started with Alveo](#)

The Alveo provides on-line infrastructure for accessing human communication data sets (speech, texts, music, video, etc.) and for using specialised tools for searching, analysing and annotating that data.



Data Discovery Interface

Browse and search collections, view documents and create lists of items for further analysis. The Data Discovery Interface provides the jumping-off point for further analysis using the Galaxy Workflow Engine, the NeCTAR Research Cloud, the R statistical package or any other preferred tool or platform. A fully featured API underpins the Data Discovery Interface, providing opportunities to extend the functionality of the Virtual Laboratory.

[Go to Alveo web app »](#)

Galaxy Workflow Engine

Initially targeted at genomics researchers, Galaxy is a scientific workflow system which is largely domain agnostic. The Galaxy Workflow Engine provides Alveo users with a user-friendly interface to run a range of text, audio and video analysis tools. Workflows defining a sequence of steps in an analysis can be created and then shared with other researchers.

[Go to Galaxy »](#)

Accessible

Accessible to non-technical researchers via workflow tools, stored protocols, and interactive GUIs, while maintaining high standards of data integrity and security.

Interoperable

Interfaces are provided to the UIMA Java framework, Python and NLTK and the Emu/R environment. Annotations are stored using RDF following a model

Sustainable

13 universities, 3 organisations, and 47 key investigators have provided support for sustained operational development and further capability development.



Contributing Partners

Gold	Silver	Bronze	Contributor
Flinders Macquarie RMIT UNE UWS	ANU Intersect UC UMELB USYD	Griffith La Trobe NICTA UNSW UWA	ASSTA UTAS

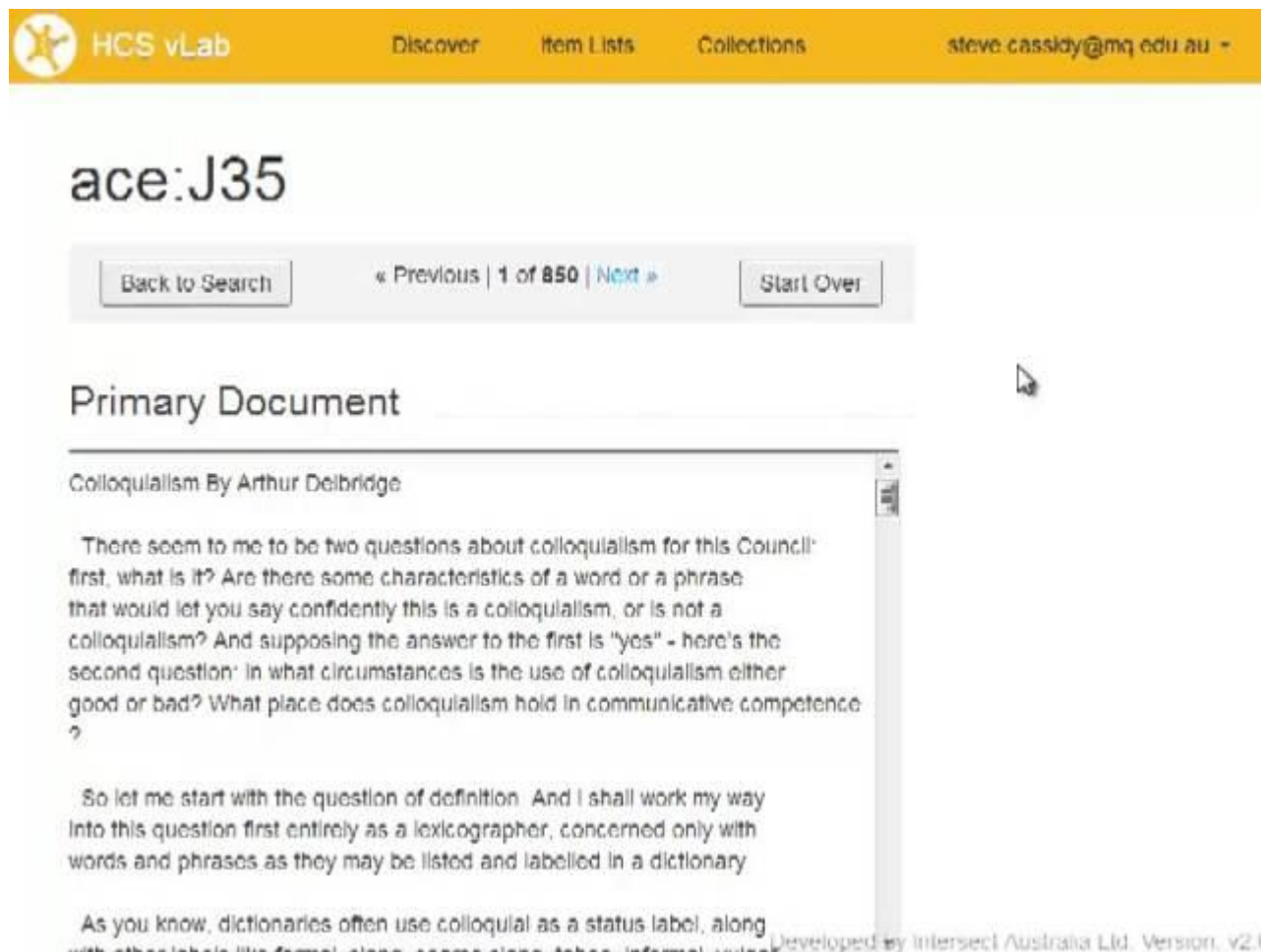
Development partner: Intersect

Intersect Development Team:

Ilya Anisimoff
Jared Berghold
David Clarke
Georgina Edwards
Karen El-Azzi
Gabriel Gasser
Matthew Hillman
Chris Kenward
Nasreen Sharique
Kali Waterford
Elyse Wise
Marc Ziani de Ferranti
Shuqian Hon
Sean Lin
Theeban Soundararajan
Vincent Tran
Stanley Hon
Pierre Estephan
Simon Yin



Watch the video!



The screenshot displays the HCS vLab web application. The top navigation bar is orange and contains the HCS vLab logo, links for Discover, Item Lists, and Collections, and a user email address: steve.cassidy@mq.edu.au. Below the navigation bar, the page title is 'ace:J35'. A search bar is present with buttons for 'Back to Search', '« Previous | 1 of 850 | Next »', and 'Start Over'. The main content area is titled 'Primary Document' and displays a document titled 'Colloquialism By Arthur Delbridge'. The document text is as follows:

There seem to me to be two questions about colloquialism for this Council: first, what is it? Are there some characteristics of a word or a phrase that would let you say confidently this is a colloquialism, or is not a colloquialism? And supposing the answer to the first is "yes" - here's the second question: in what circumstances is the use of colloquialism either good or bad? What place does colloquialism hold in communicative competence?

So let me start with the question of definition. And I shall work my way into this question first entirely as a lexicographer, concerned only with words and phrases as they may be listed and labelled in a dictionary.

As you know, dictionaries often use colloquial as a status label, along with other labels like formal, slang, coarse slang, taboo, informal, vulgar.


Developed by Intersect Australia Ltd. Version: v2.0



[Watch](#)



Data Discovery

 HCS vLab

DiscoverItem ListsCollections

georgina@intersect.org.au ▾

My Account

[Edit Account Details](#)

[Change Account Password](#)

Licence Agreements


[Report An Issue](#)

Review and Acceptance of Licence Terms

Manage your subscriptions to collections.

Collection or Collection List	Collections	Owner	State	Actions
AusNC	8	jared@intersect.org.au	Not Accepted	Preview & Accept Licence Terms
PARADISEC	2	jared@intersect.org.au	Not Accepted	Preview & Accept Licence Terms
avoze	1	jared@intersect.org.au	Accepted	Review Licence Terms
austalk	1	jared@intersect.org.au	Not Accepted	Preview & Accept Licence Terms


Discover data via metadata facets

 HCS vLab

DiscoverItem ListsCollections

georgina@intersect.org.au ▾

Limit your search

Search Text Containing 

Collection >

Created >

Mode ▾
D 24
TV 2
spoken 45682

Speech Style >

Publication Status >

Written Mode >

Interactivity >

Communication Context >

Communication Medium >

You searched for:

Mode > spoken x

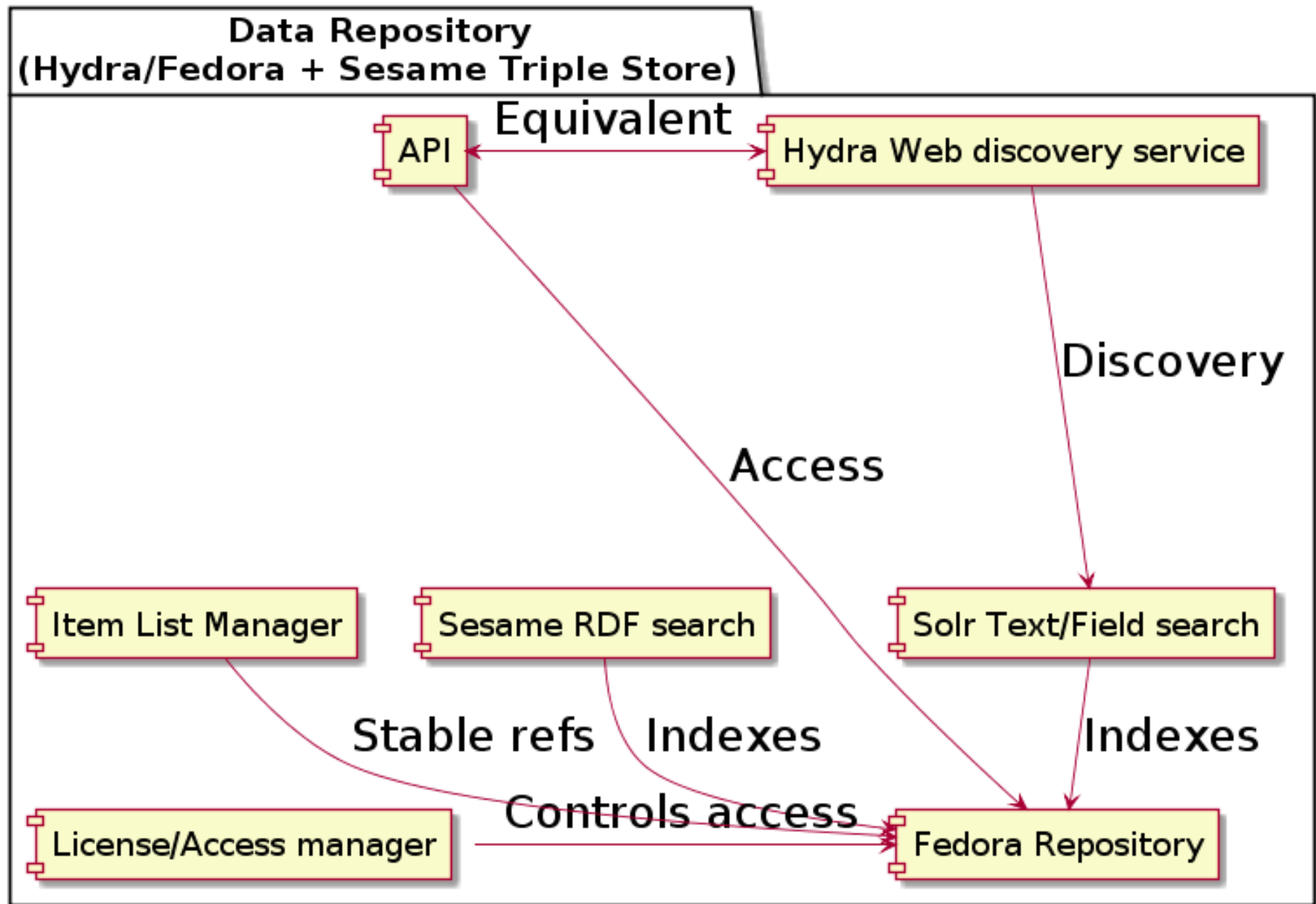
Start Over

Add Selected to list ▾Add All to list ▾« Previous | 1 - 20 of 45682 | [Next »](#)20 per page ▾

<input type="checkbox"/>	Identifier	Title	Created Date	Type(s)
<input type="checkbox"/>	ice:S2B-025B	Earthworm	9/10/91	Text
<input type="checkbox"/>	ice:S2B-006			unspecified
<input type="checkbox"/>	ice:S2B-010			unspecified
<input type="checkbox"/>	ice:S2A-070			unspecified
<input type="checkbox"/>	ice:S2A-070C		11/1/93	Text
<input type="checkbox"/>	ice:S2A-067			unspecified
<input type="checkbox"/>	ice:S2B-023A	Ockham's Razor	7/7/91	Text
<input type="checkbox"/>	ice:S2B-022B	Ockham's Razor	27/1/91	Text
<input type="checkbox"/>	ice:S2A-018			unspecified

Developed by Intersect Australia Ltd. Version: V2_candidate_02

Major functions of the repository component





Q. What do you call something that automatically prepares data for ingest?



- A: Robochef



Sivusto Rakenteilla





Architecture: Discovery leads to data

Documents

Filename	Type	Size
3-200.txt	Original	6.2 kB
3-200-raw.txt	Raw	6.2 kB
3-200-plain.txt	Text	6.0 kB



Including various media

mitchedelbridge:S1232s1

[Back to Search](#)

[« Previous](#) | 3 of 31 | [Next »](#)

[Start Over](#)

Primary Document



Item Details

Identifier:	S1232s1
Collection:	mitchedelbridge
Mode:	unspecified
Speech Style:	unspecified
Interactivity:	unspecified
Communication Context:	unspecified
Discourse Type:	interactive_discourse



Post discovery: compile your own stable “Item Lists”

Item List Actions ▾		1 - 20 of 20		20 per page ▾
Identifier	Title	Created Date	Type(s)	
cooee:2-011		1826	Original, Raw, Text	
cooee:4-288		1893	Original, Raw, Text	
cooee:2-271		1843	Original, Raw, Text	
cooee:3-022		1851	Original, Raw, Text	
cooee:2-330		1848	Original, Raw, Text	
cooee:2-259		1842	Original, Raw, Text	
cooee:4-228		1890	Original, Raw, Text	
cooee:1-103		1806	Original, Raw, Text	
cooee:3-200		1860	Original, Raw, Text	
cooee:2-280		1844	Original, Raw, Text	
cooee:3-282		1872	Original, Raw, Text	



Concordance: a tool run on an Item List

cooee:3-022 Found 1 match

more years to run out, it will then according to the general rate of letting

cooee:1-249 Found 5 matches

colony, and if they did not apply then ,he would beg to enquire when did

as the Common Law Officer. He was then told the Acts of Parliament prescribed a

were not in force in this colony, then the application of the former occasion was

penalty for his neglect. It was clear then ,that there was a legal right; it

without doubt apply here. Upon general principles then the court of sessions having power to



API Access: Get a key

The screenshot displays the user interface for obtaining an API key. At the top, a yellow header bar contains the email address `georgina@intersect.org.au` with a dropdown arrow. Below this, a white sidebar on the left contains the text "No API Key generated" and a yellow button labeled "Generate API Key". The main content area on the right is a white panel with a dropdown menu open, listing several options: "Saved Searches", "History", "My Account", "My Licence Agreements", "API Key" (highlighted in yellow with a right-pointing arrow), "Report An Issue", and "Logout". At the bottom of the sidebar, there are two more options: "Copy to Clipboard" and "Download API Key".

georgina@intersect.org.au ▾

Saved Searches
History

My Account
My Licence Agreements
API Key ▶
Report An Issue
Logout

No API Key generated
Generate API Key

Copy to Clipboard
Download API Key

Copy-paste access to data

Use Dogs in Emu/R

Copy the following code into your R environment and download the API token file. Make sure you have all the required R packages installed prior to execution.

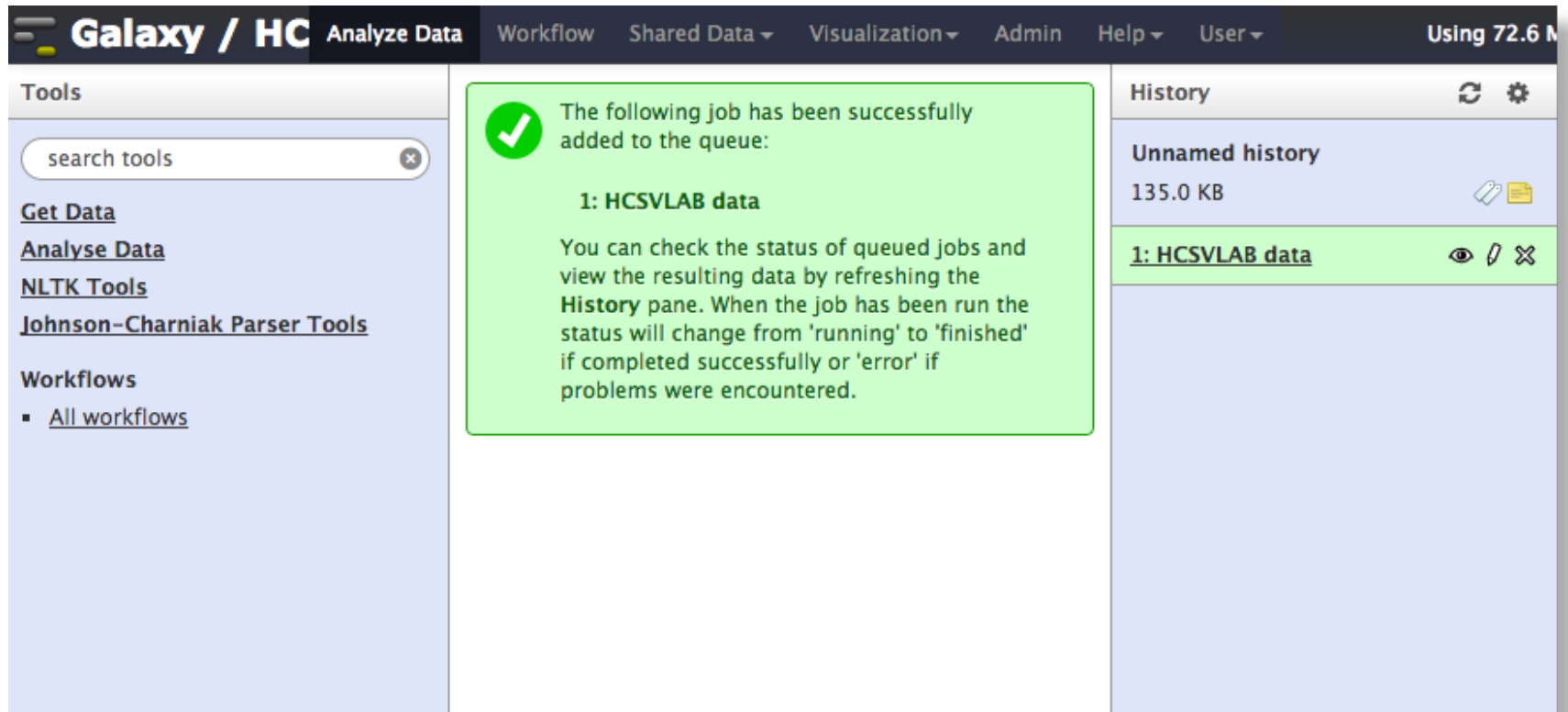
Save the following file to your home directory
Linux or Unix: /home/<user>
Mac: /Users/<user>
Windows: C:\Users\<user>

[Download API key config file](#)

```
library(emuSX)
item_list = readItemList('http://ic2-hcsvlab-staging1-vm.intersect
.org.au/item_lists/129.json')
```

Close

Workflow: Galaxy



The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo, 'Galaxy / HC', and several menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The user's memory usage is shown as 'Using 72.6 M'. The left sidebar contains a 'Tools' section with a search bar and links to 'Get Data', 'Analyze Data', 'NLTK Tools', 'Johnson-Charniak Parser Tools', and 'Workflows'. The main content area features a green notification box with a checkmark icon, stating: 'The following job has been successfully added to the queue: 1: HCSVLAB data'. Below this, it explains that users can check the status of queued jobs and view the resulting data by refreshing the 'History' pane. The right sidebar shows the 'History' panel with a refresh and settings icon. It lists 'Unnamed history' (135.0 KB) and '1: HCSVLAB data' (with view, edit, and delete icons).

Galaxy / HC Analyze Data Workflow Shared Data Visualization Admin Help User Using 72.6 M

Tools

search tools

[Get Data](#)

[Analyze Data](#)

[NLTK Tools](#)

[Johnson-Charniak Parser Tools](#)

Workflows

- [All workflows](#)

History

Unnamed history
135.0 KB

1: HCSVLAB data

Chain processes on Item Lists - eg [Tokenizer] -> [Frequency List]

The screenshot displays the Galaxy / HCSVLAB web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User, along with a status indicator 'Using 72.8 M'. The left sidebar contains a 'Tools' section with a search bar and a list of tools under 'Get Data' and 'Analyse Data'. The 'Analyse Data' section lists several tools: Frequency List, Sentence Segmenter, Tokenizer, POS, Stemmer, and Collocation, each with a brief description. The main workspace shows a workflow with two steps: '1: HCSVLAB data' and '2: Tokenizer'. The output of the 'Tokenizer' step is a list of words and their frequencies, which is then processed by the 'Frequency List' tool. The right sidebar shows a 'History' section with a list of unnamed history items, including '3: Frequency List' and '2: Tokenizer'.

Galaxy / HCSVLAB Analyze Data Workflow Shared Data Visualization Admin Help User Using 72.8 M

Tools

search tools

Get Data

Analyse Data

- Frequency List Takes a text input and generates a frequency list

NLTK Tools

- Sentence Segmenter Segments the text input into separate sentences
- Tokenizer Splits the text from the input file into word tokens
- POS Generates part of speech tags from text or a list of tokens
- Stemmer Takes a list of tokens and generates a list of word stems using one of the stemming algorithms
- Collocation Generates a list of the most frequent collocations from an input sequence

the	1808
,	1704
of	977
and	767
.	745
to	675
a	560
in	472
that	318
it	283
be	235
for	232
is	232
as	216
;	208
i	206
was	184
not	178
with	160
at	153
this	151

History

Unnamed history
332.5 KB

3: Frequency List

2: Tokenizer

1: HCSVLAB data



Tools

Get Data

[Analyse Data](#)[NLTK Tools](#)[Johnson-Charniak Parser Tools](#)[PsySound](#)

Workflow control

Inputs

Workflow Canvas | PsySound



Details

Edit Workflow Attributes

Name:

PsySound

Tags:

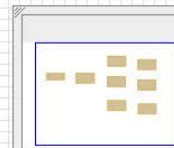
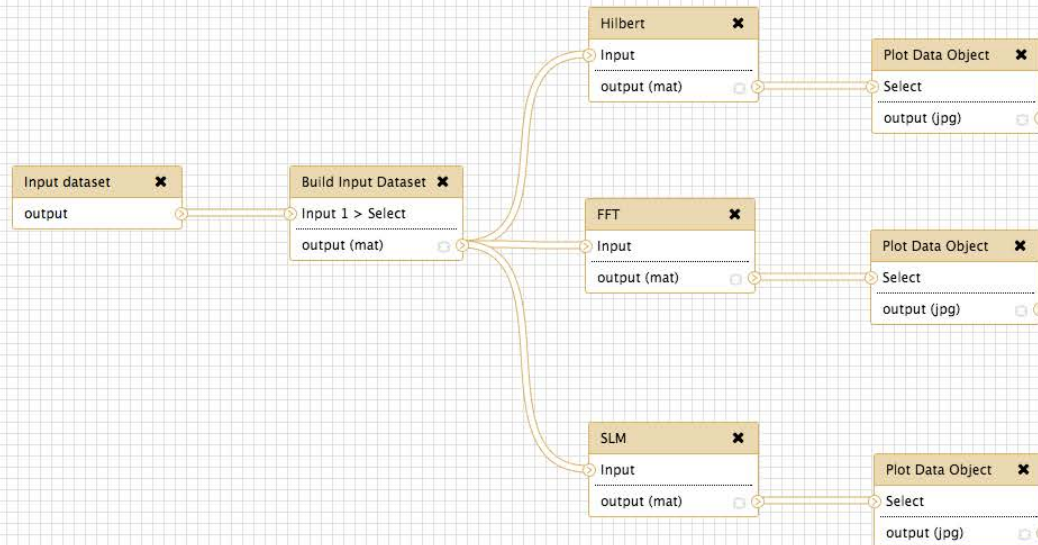


Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:

Describe or add notes to workflow

Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.





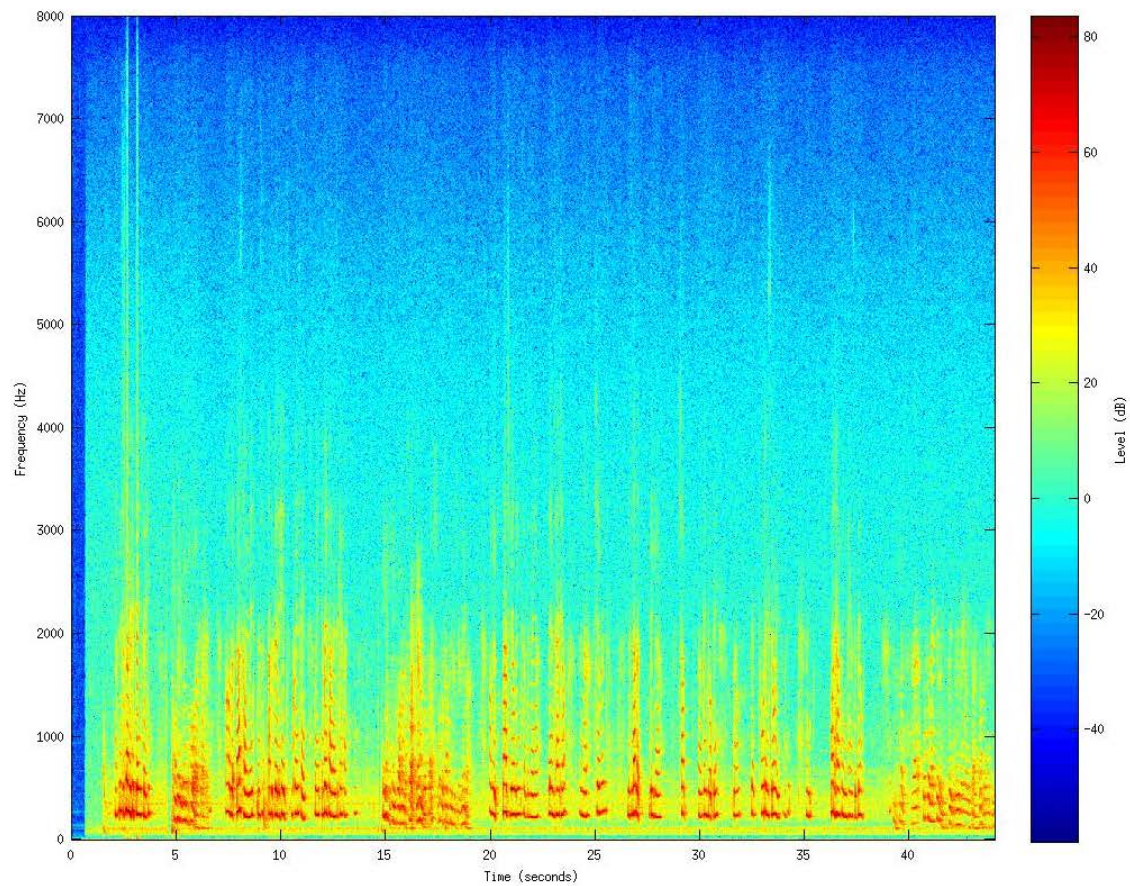
Tools

search tools

[Get Data](#)
[Analyse Data](#)
[NLTK Tools](#)
[Johnson-Charniak Parser Tools](#)
[PsySound](#)

Workflows

[All workflows](#)

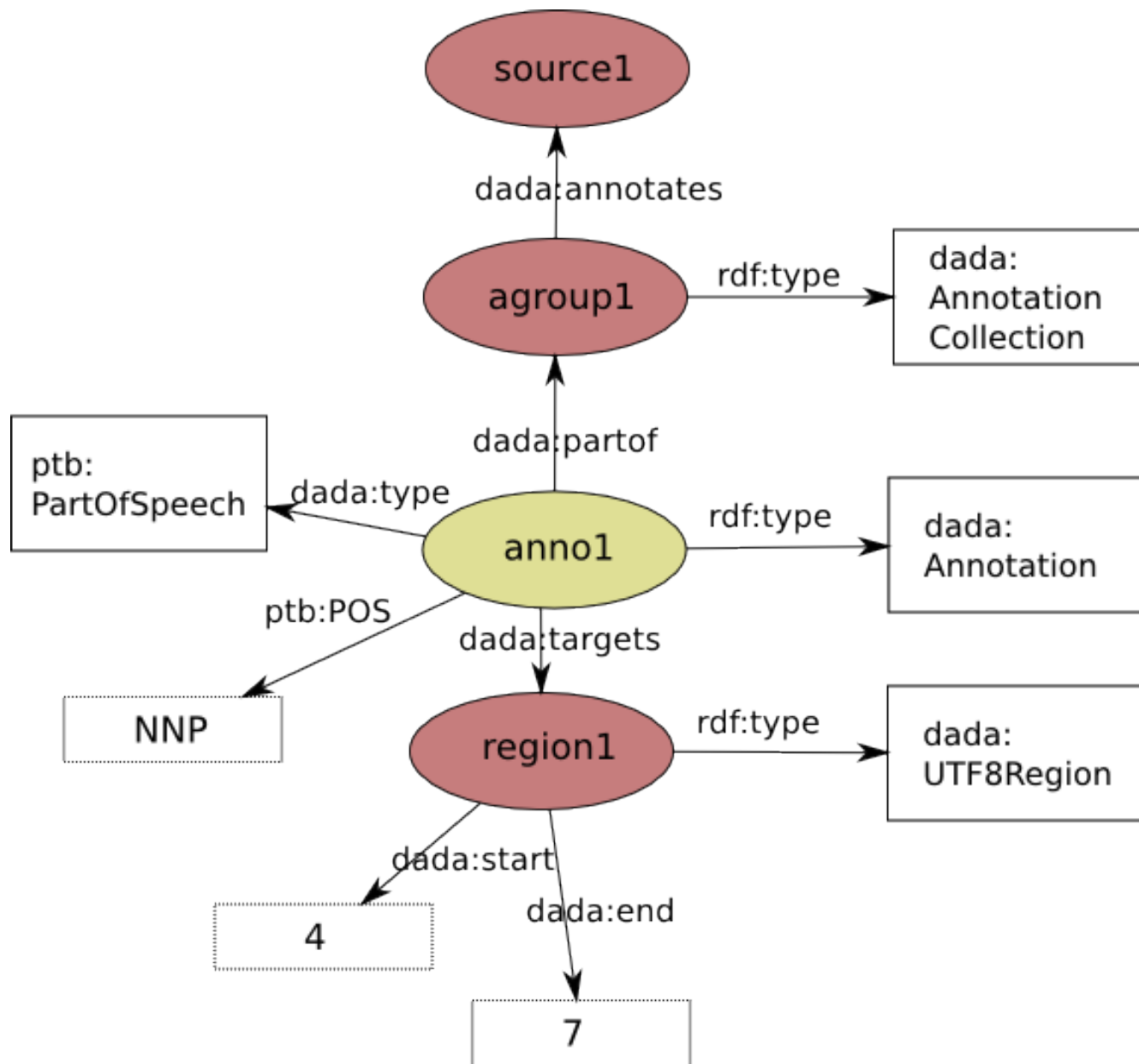


History

Unnamed history

618.1 MB

8: Psysound Plot	👁	🗑
7: Psysound Plot	👁	🗑
6: Psysound Plot	👁	🗑
5: Psysound Data Objects (SLM)	👁	🗑
4: Psysound Data Objects (FFT)	👁	🗑
3: Psysound Data Objects (Hilbert)	👁	🗑
2: Psysound Fileset	👁	🗑
1: S1223n.wav	👁	🗑



Workflow Engine (Galaxy)

Pre-configured tools

Web based workflows

Data interface

Data Repository (Hydra/Fedora + Sesame Triple Store)

Web discovery service

Repository

Index(es)

API

Item List Manager

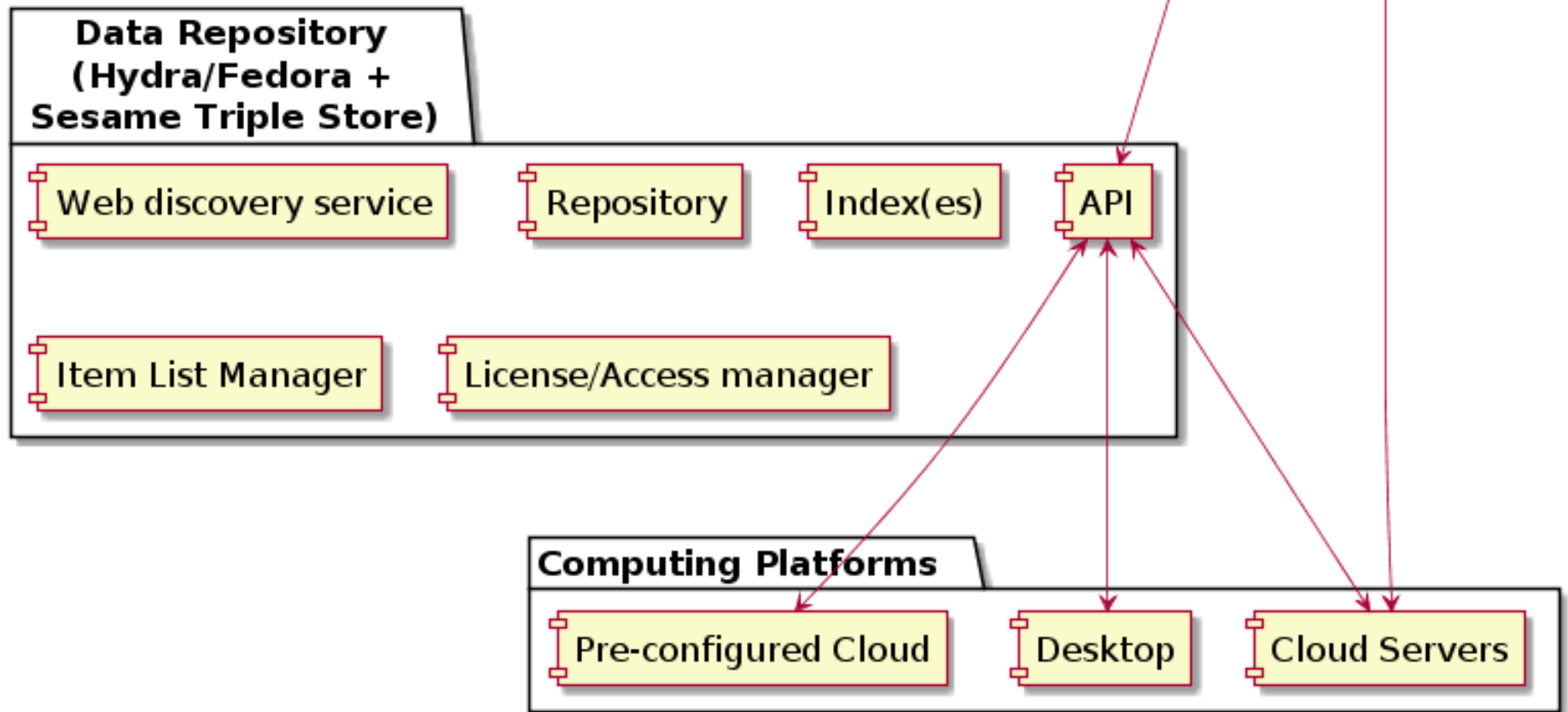
License/Access manager

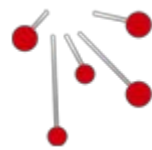
Computing Platforms

Pre-configured Cloud

Desktop

Cloud Servers





FedoraCommons™

Reuse



blacklight



Get a head

on your repository.

Multi-Purpose Repository Solutions
Flexible User Interfaces
Durable Digital Asset Management

powered by
 **Galaxy**



Reproducible Research







User feedback

"I really liked using the system and the instructions were very easy to use and the system easy to navigate. [...] This platform would be very useful for my research."

--Tester

