

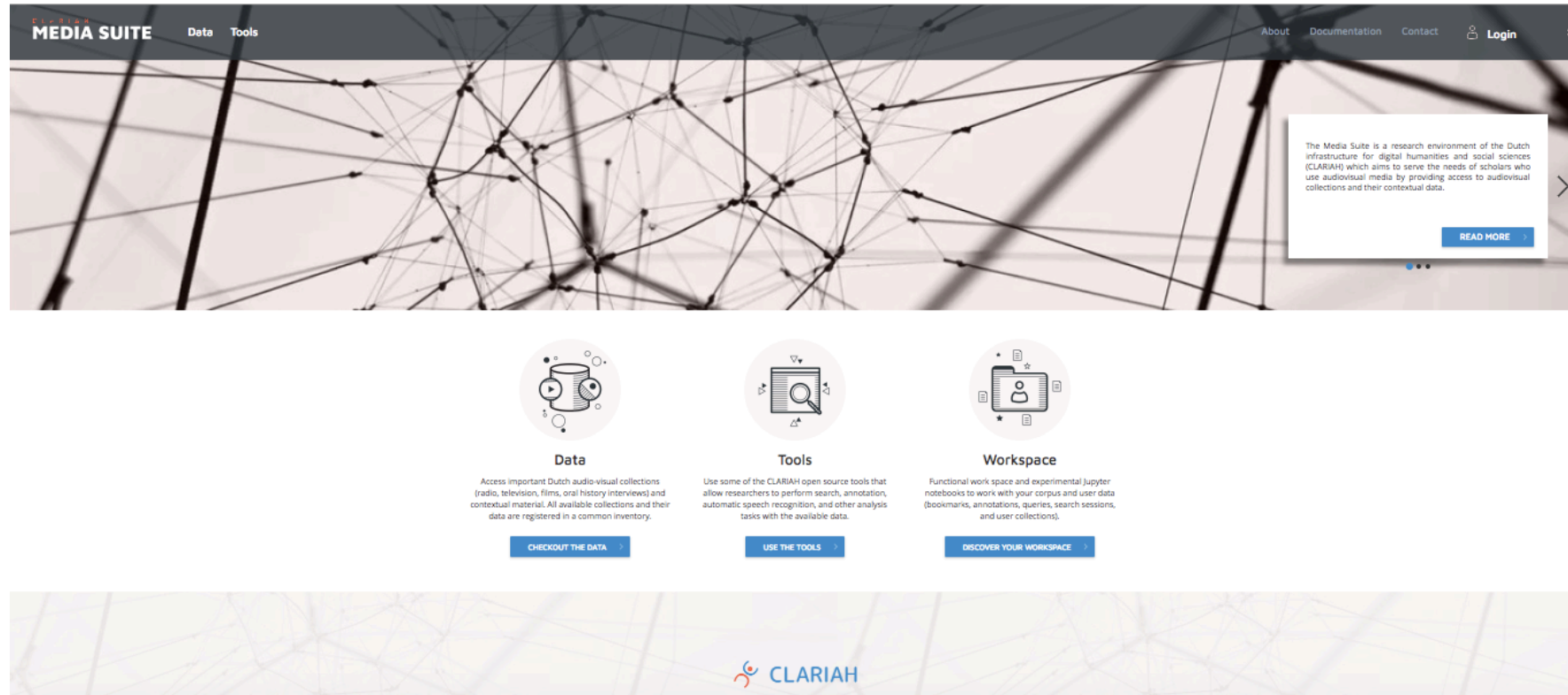
Provenance in WP5's Media Suite

CLARIAH Provenance workshop

DANS

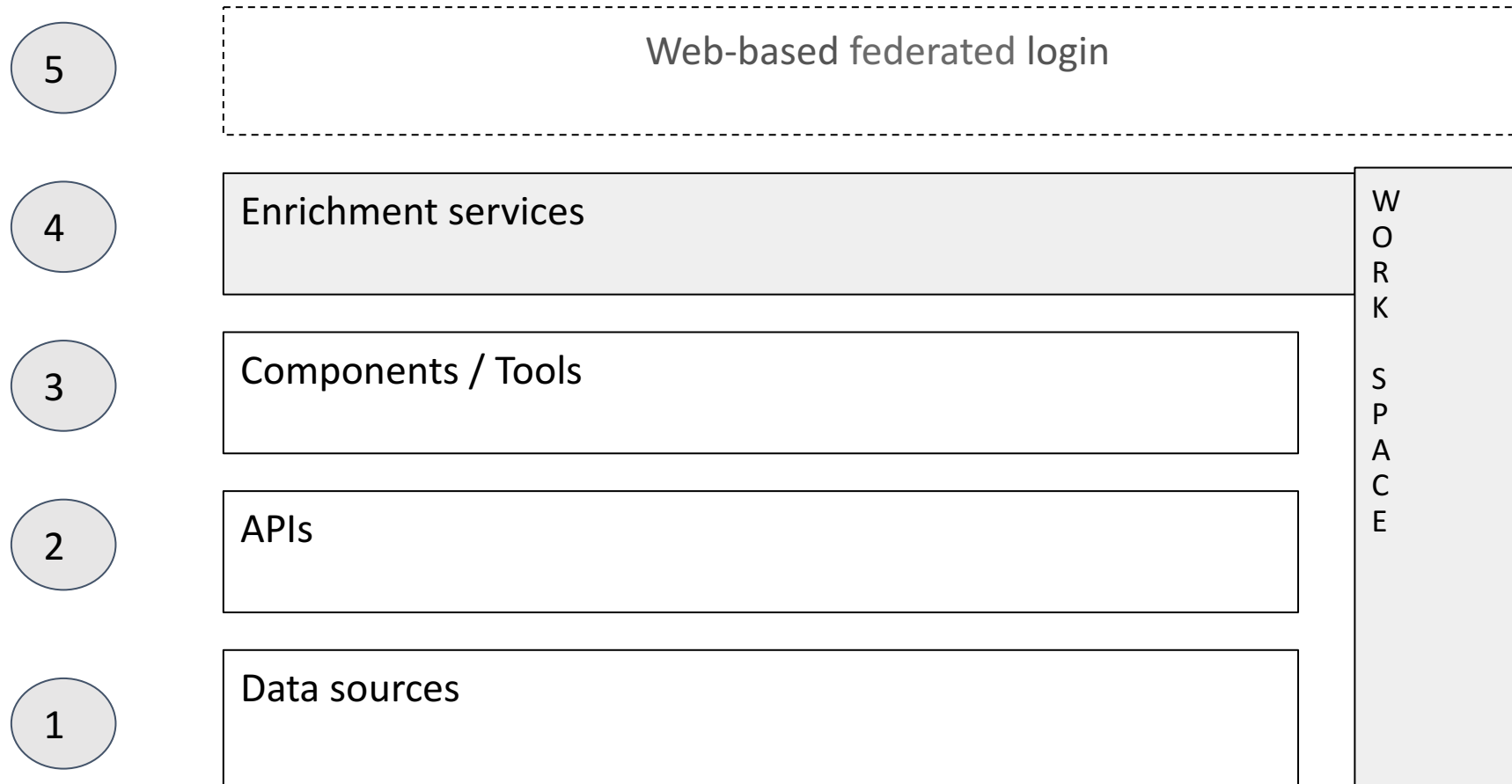
September 4, 2018

The CLARIAH Media Suite: <http://mediasuite.clariah.nl/>



- A high-level tool (user-friendly research environment) built in a modular/sustainable approach
- Integrates data and tools for both novice users and advanced users (via Jupyter notebooks)
- We are not yet actively implementing provenance solutions based on “formal” models/solutions, but rather respecting the principles of provenance information (i.e., “source critique” in scholarly terms): transparency and traceability (location of the sources, and of changes made to the sources on the process of providing access).

Media Suite's building blocks



Source: Ordelman, R., Martínez Ortíz, C., Melgar Estrada, L., Koolen, M., Blom, J., Melder, W., ... Noordegraaf, J. (2018). Challenges in Enabling Mixed Media Scholarly Research with Multi-media Data in a Sustainable Infrastructure – DH2018. Presented at the Digital Humanities 2018, Mexico. Retrieved from <https://dh2018.adho.org/en/challenges-in-enabling-mixed-media-scholarly-research-with-multi-media-data-in-a-sustainable-infrastructure/>

Challenges in respecting provenance and integrating provenance information

1. Data sources:

- CKAN registration
 - Human-readable descriptions of collections/data and the content providers
 - Human-readable descriptions and conversion files explaining data processes

The screenshot shows a CKAN dataset page for 'Radio Collection'. The left sidebar contains metadata for the organization 'Nederlands Instituut voor Beeld en Geluid (Various creators)', including social media links and a license. The main content area features a description of the radio collection, a list of data resources with 'Explore' buttons, a filter bar with 'audio', 'clariah_media_es_in...', and 'radio' options, and an 'Additional Info' table.

Followers
0

Organization

BEELD EN GELUID

Nederlands Instituut voor Beeld en Geluid (Various creators)

The Netherlands Institute for Sound and Vision is a cultural-historical organization of national interest. It collects, preserves and opens the audiovisual heritage for as many... [read more](#)

Social

- Google+
- Twitter
- Facebook

License

Other (Not Open)

Radio Collection

De radiocollectie is een van de oudste collecties van Beeld en Geluid. Van de allereerste uitzending van de Hilversumsche Draadloze Omroep (HDO) uit 1923 bestaat helaas geen opname want apparatuur om de live-uitzendingen op te nemen bestond gewoonweg nog niet. ([Lees meer](#))

Data and Resources

- Website Beeld en Geluid**
De radiocollectie is een van de oudste collecties van Beeld en Geluid. Van de... [Explore](#)
- Clariah Media Suite Elasticsearch index**
[Explore](#)
- Spraakherkenningsresultaten Radio 1**
Automatically generated speech recognition transcripts for radio broadcasts... [Explore](#)

audio clariah_media_es_in... radio

Additional Info

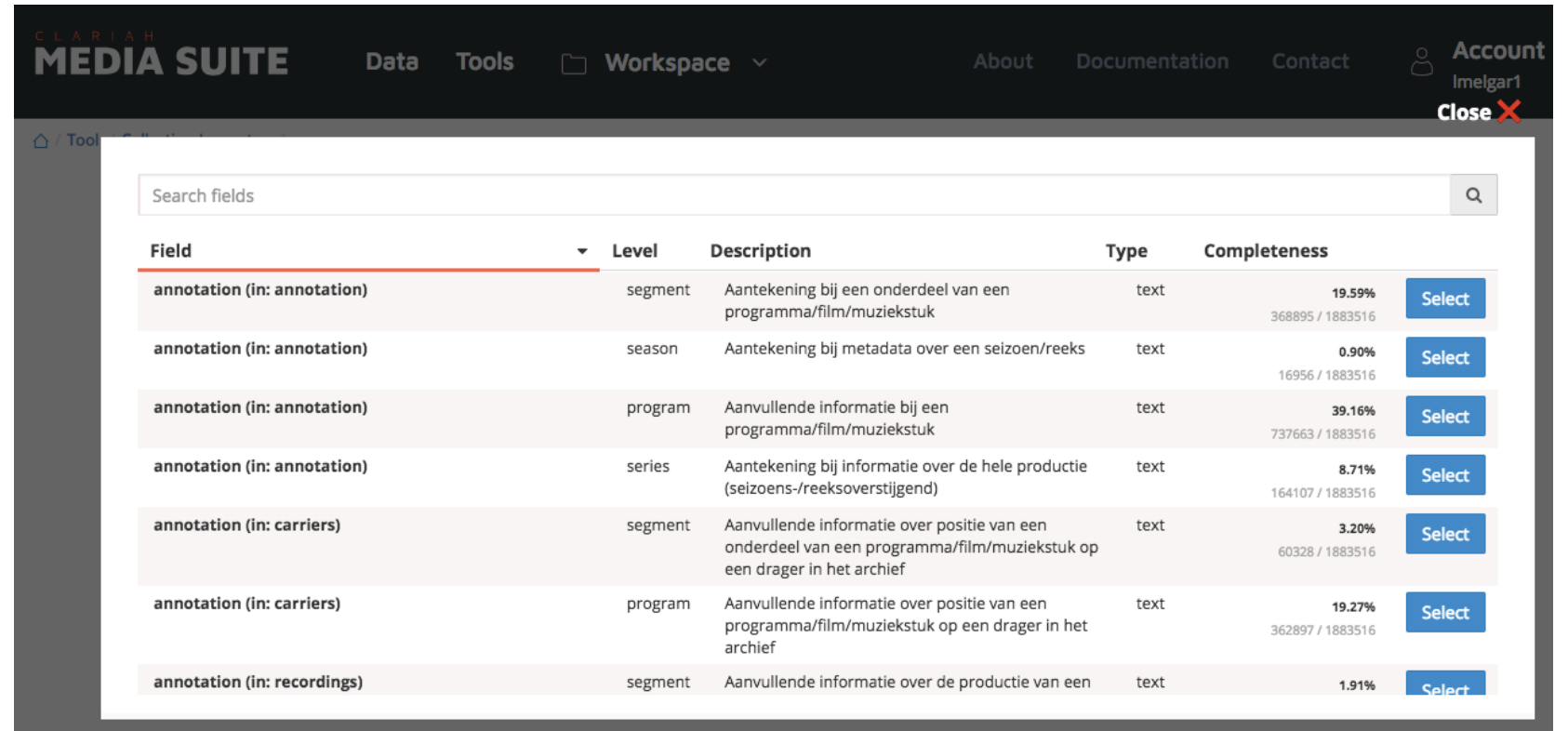
Field	Value
Last Updated	January 29, 2018, 2:50 PM (UTC+01:00)
Created	February 22, 2017, 3:25 PM (UTC+01:00)
Data Disclaimer	This dataset has undergone processing before it was uploaded to this register. Examples of possible processing operations are: filter, transform, enrich, clean, interpret, combine or reconcile.
Operations applied	derive, filter.
Processing details	Derive: from http://52.89.136.166/dataset/nisv-catalogue-aggr Filter: Subset van de Beeld en Geluid catalogus gefilterd op basis van het veld: 'bga:series.bg:catalog.raw' met de waarde: 'Geluidsregistraties/Publieke omroep/Programma's'.

Challenges in respecting provenance and integrating provenance information

1. Data sources

Currently: “user-friendly” way of respecting provenance information:

- Respecting/offering original metadata (by content provider)
- Following the principles of data transparency: visualization of metadata completeness
- Providing metadata dictionaries



The screenshot shows the CLARIAH MEDIA SUITE interface. At the top, there is a navigation bar with 'Data', 'Tools', 'Workspace', 'About', 'Documentation', and 'Contact'. A user account 'Imelgar1' is visible in the top right corner. Below the navigation bar, there is a search bar labeled 'Search fields'. The main content area displays a table with the following columns: Field, Level, Description, Type, and Completeness. Each row represents a different metadata field and includes a 'Select' button.

Field	Level	Description	Type	Completeness
annotation (in: annotation)	segment	Aantekening bij een onderdeel van een programma/film/muziekstuk	text	19.59% 368895 / 1883516
annotation (in: annotation)	season	Aantekening bij metadata over een seizoen/reeks	text	0.90% 16956 / 1883516
annotation (in: annotation)	program	Aanvullende informatie bij een programma/film/muziekstuk	text	39.16% 737663 / 1883516
annotation (in: annotation)	series	Aantekening bij informatie over de hele productie (seizoens-/reeksverstijgend)	text	8.71% 164107 / 1883516
annotation (in: carriers)	segment	Aanvullende informatie over positie van een onderdeel van een programma/film/muziekstuk op een drager in het archief	text	3.20% 60328 / 1883516
annotation (in: carriers)	program	Aanvullende informatie over positie van een programma/film/muziekstuk op een drager in het archief	text	19.27% 362897 / 1883516
annotation (in: recordings)	segment	Aanvullende informatie over de productie van een	text	1.91%

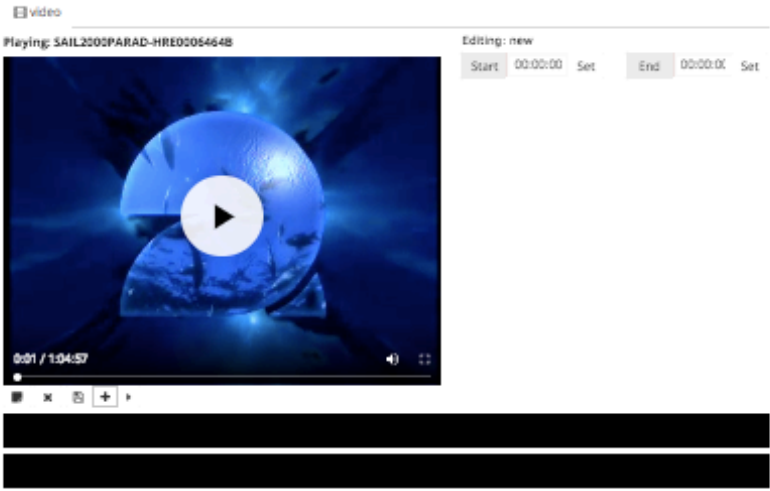
Challenges in respecting provenance and integrating provenance information

3. Components/Tools:

Currently: Making easier for users to trace metadata back

- Links to original sources
- Metadata presentation in a transparent way

Future work: encode provenance information formally (depending on CLARIAH common approach)



video

Playing: SAIL2000PARAD-HRE00054648

Editing: new

Start 00:00:00 Set End 00:00:00 Set

0:01 / 1:04:57

Metadata

Saved annotations

ID	3942285@program
Index	nissv-catalogue-aggr-fulb-18-158 (type: program_aggr)
Title	SAIL 2000: PARADE OF SAIL
Date	24-08-2000
Description	Verslag van de Parade of Sail, de grote intocht van zeil-schepen vanaf de sluizen in IJmuiden naar de haven van Amsterdam.
Source	View in catalogue
Broadcast	NOS

All data

```
▼ Object
  ▼ bg:languages: Object
    ▼ bg:language: Object
      bg:use: "uitgangstaal"
      bg:language: "Nederlands"
    bg:sourcecatalog: "BAS.DAG.AANWAS EVENT"
  ▼ bg:recordings: Object
    No properties
  ▼ dcterms:isPartOf: Object
    dc:identifier: "373900#season"
    bg:summary: "Verslag van de Parade of Sail, de
    datestamp: "2016-10-02T11:31:37Z"
  ▼ bg:context: Object
    bg:colour: "kleur"
    dc:identifier: "3942285@program"
    dc:relation: "http://in.bseidengeleid.nl/colle
```

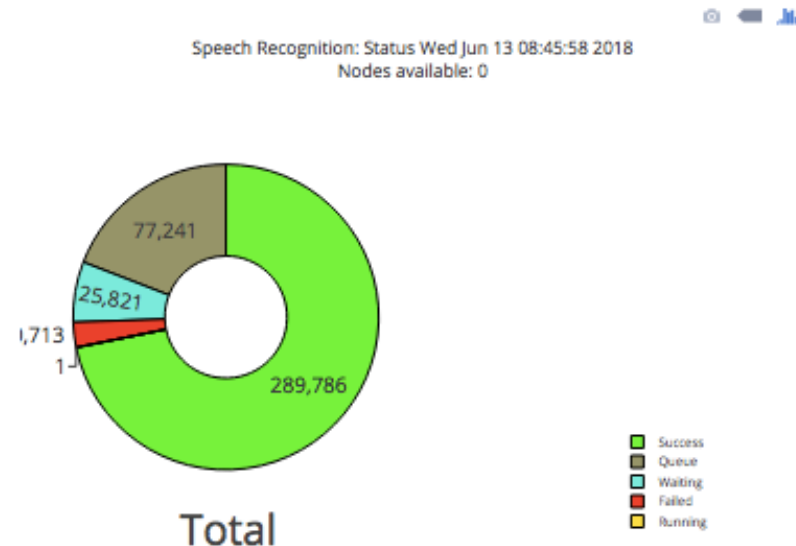
Challenges in respecting provenance and integrating provenance information

4. Enrichment services:

- Currently: we provide live statistics/ASR work on a separate web site,
- Future work: analyze how to integrate provenance information in the ASR metadata



Deze pagina is voor gebruik van Beeld en Geluid personeel, voor monitoring en troubleshooting van de automatische spraakherkenning. De grafieken zijn aanvullend op de grafieken op de 'Spraakherkenning Radio en TV' site. Let op: de informatie kan nog veranderen, omdat ik volop nog aan het uitzoeken ben hoe het in elkaar zit. De grafieken zijn dus mogelijk nog niet betrouwbaar. Graag je feedback geven (neem contact op met Mari Wigham)



Challenges in respecting provenance and integrating provenance information

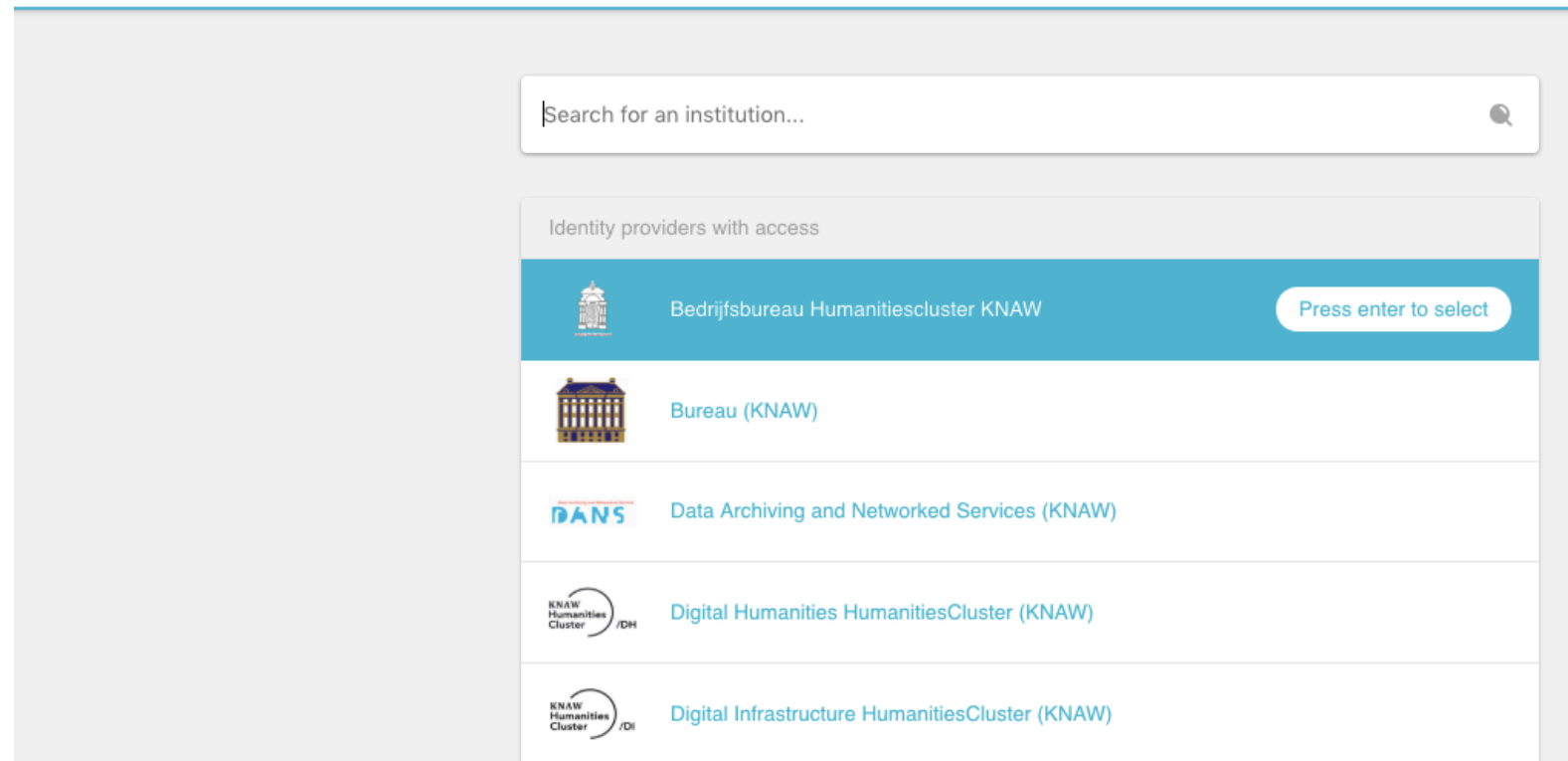
5. Login/Authentication:

- Working together with WP2 on an “authenticatie route”






6. Workspace

- Currently: storing user annotations and saved queries
- Future work: deciding on what data from the users should be stored for keeping provenance information in user annotations (and also for log analysis)

SURFconext - Select an institution to login to the service: [CLARIAH IAA Engine](#) | [CLARIN](#)



The screenshot shows the SURFconext login interface. At the top, there is a search bar with the placeholder text "Search for an institution...". Below the search bar, there is a section titled "Identity providers with access". This section contains a list of four identity providers, each with a logo and a name. The first provider, "Bedrijfsbureau Humanitiescluster KNAW", is highlighted in blue and has a button that says "Press enter to select". The other providers are "Bureau (KNAW)", "Data Archiving and Networked Services (KNAW)", "Digital Humanities HumanitiesCluster (KNAW)", and "Digital Infrastructure HumanitiesCluster (KNAW)".

Identity providers with access	
	Bedrijfsbureau Humanitiescluster KNAW Press enter to select
	Bureau (KNAW)
	Data Archiving and Networked Services (KNAW)
	Digital Humanities HumanitiesCluster (KNAW)
	Digital Infrastructure HumanitiesCluster (KNAW)