State of the technical infrastructure

Dieter Van Uytvanck

Technical Director CLARIN ERIC

dieter@clarin.eu

CLARIN Annual Conference 2018

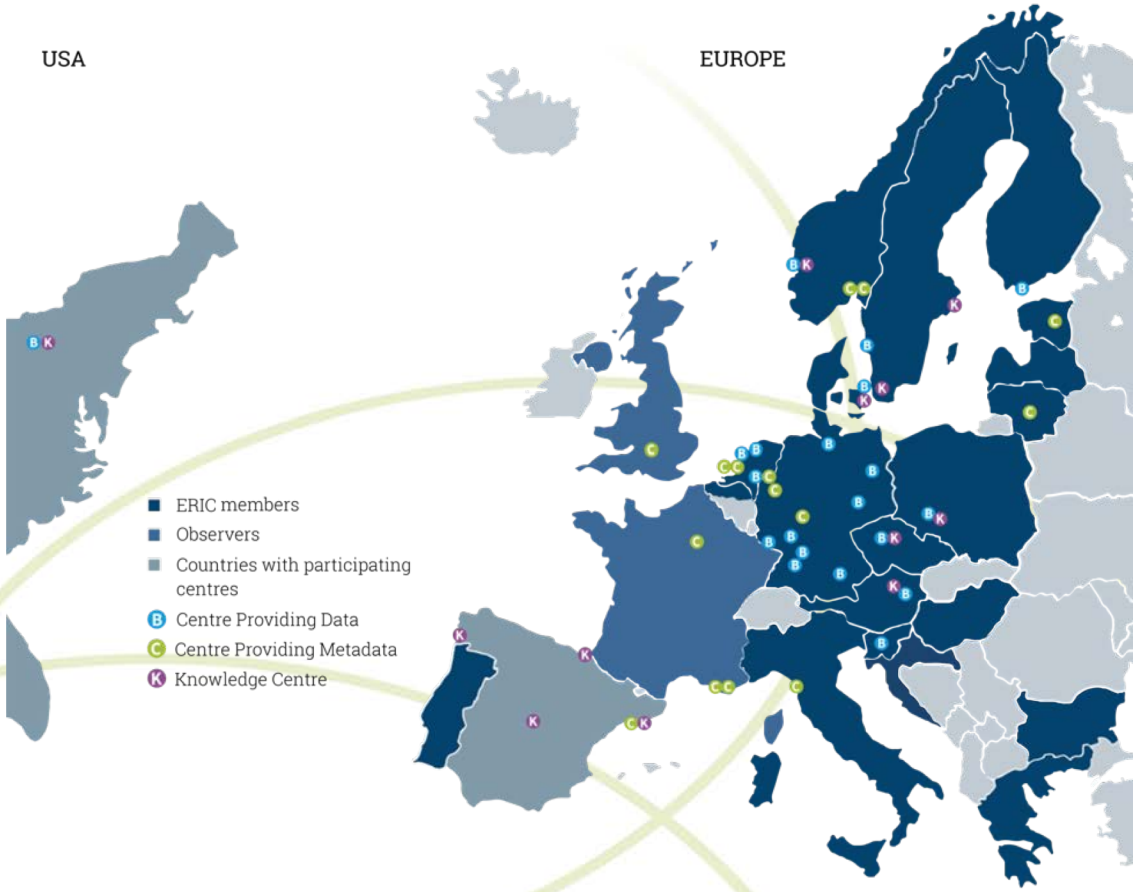
Pisa

10 October 2018

An overview of the technical infrastructure highlights

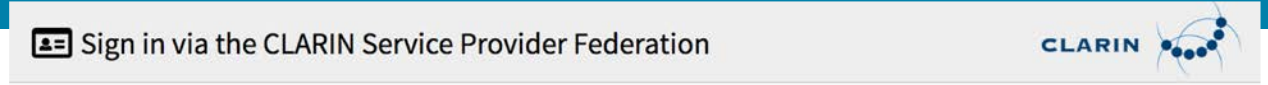
Short round-up 2018: Centres

- New B-centres:
  - Språkbanken (SE)
  - ILC4CLARIN (IT)
- Re-assessment of the ARCHE centre (AT), 10 more re-assessments pending (awaiting CoreTrustSeal finalization)
- in total 21 certified B-centres, 47 registered centres



Short round-up 2018: Federated login

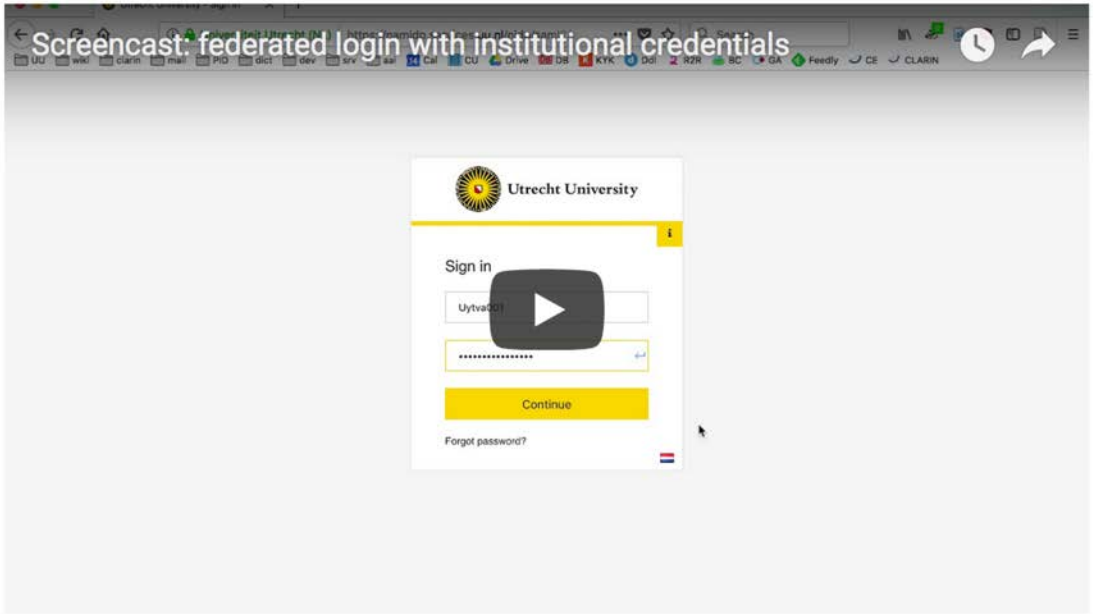
- New country in Service Provider Federation: Croatia
- Total number of organisations that can login: **1800**
- On average **2017 logins per month** via the central discovery service
- Experimental launch of a knowledge base for federated login support: **help.clarin.eu**
- New version of CLARIN discovery service :
  - follows the standard CLARIN styleguide.
  - It looks prettier and is easier to use on mobile platforms.


Sign in via the CLARIN Service Provider Federation
CLARIN
Select your home organisation below. This is usually the organisation where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in. If you cannot find your organisation in the list below, please select the clarin.eu website account and use your CLARIN website credentials. If you don't have such credentials you can register an account here. For questions please contact spf@clarin.eu.
Previously chosen home organisation
Utrecht University
Netherlands
Home organisation list
pisa
All countries
IRCCS San Raffaele Pisana - Rome
Italy
SR
University of Pisa
Italy

How can I login with my institutional credentials?

This screencast demonstrates how to login to a web application using federated login via your own institutional Identity Provider.

Written by Dieter Van Uytvanck
Updated over a week ago



Short round-up 2018: VLO

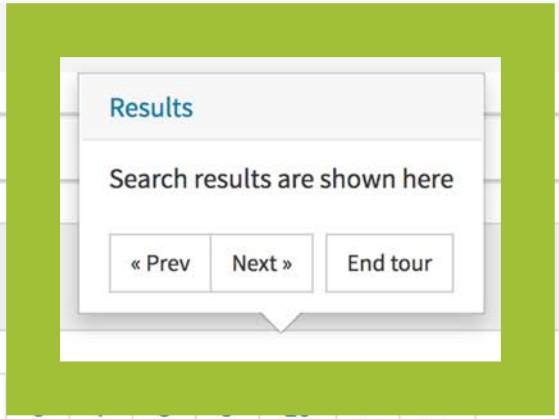
- Three releases brought
  - cleaner facets
  - visualisation of licenses and
  - a **guided tour** [inspired by @PhilosTEI]
- Automatic collapsing of similar result already available for testing at **alpha-vlo.clarin.eu**
  - Reduces number of initially visible records from 850k to 93k

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or continue to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

See all records | Take a quick tour



Results per page: 10

<< < 1 2 3 4 5

Arabic Speech Corpus

(Part of Oxford Text Archive)

The resource is a **speech corpus**, with digital audio files, text transcripts, and files containing time stamps of the phoneme boundaries. 1813 .wav files containing spoken utterances. 1813 .lab files containing text utterances. 1813 .TextGrid files containing the phoneme labels with time stamps of the boundaries where …

IFA speech corpus

(Part of LRT + Open Submissions Data & Tools)

Spoken **corpus** containing **speech** of 4 male and 4 female speakers. 50,000 words segmented at phoneme level

Duplicate folding in alpha-vlo.clarin.eu

SamtaleBank Steensig Corpus

(Part of TalkBank)

⊞ Samtalebanken er det danske MOVIN netværks talesprogskorpus og består af audio- og videooptagelser med tilhørende transs…

See this record and its resources at archive.mpi.nl

The search results include 29 (near) duplicate(s) of this record.

Click here to see the (near) duplicate results

Duplicate folding in alpha-vlo.clarin.eu

SamtaleBank Steensig Corpus

(Part of TalkBank)

Samtalebanken er det danske MOVIN netværks talesprogskorpus og består af audio- og videooptagelser med tilhørende transskription. Transskriptionerne følger konversationsanalytiske principper og er kendetegnet ved en høj granularitet især i forhold til timing og overlap af deltagernes talebidrag. Samtalebanken is compi…

See this record and its resources at archive.mpi.nl

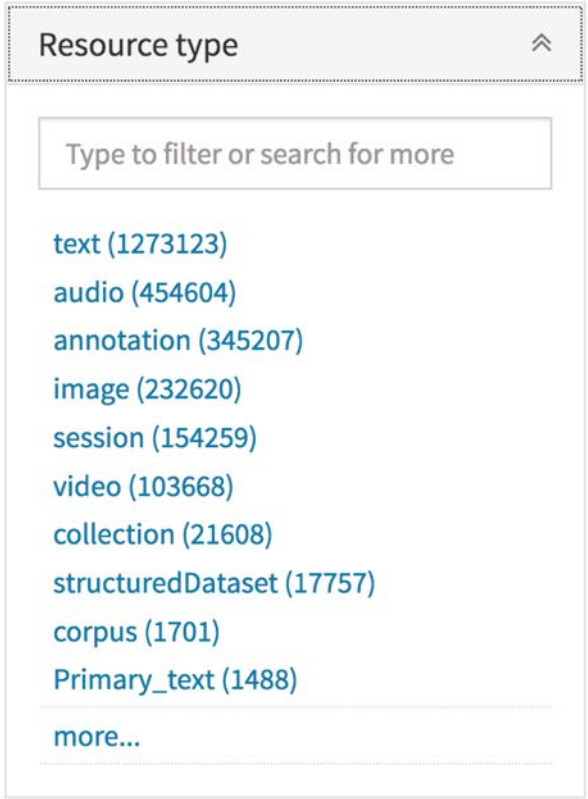
The search results include 29 (near) duplicate(s) of this record.

- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus
- SamtaleBank Steensig Corpus

<< < 1 2 3 > >>

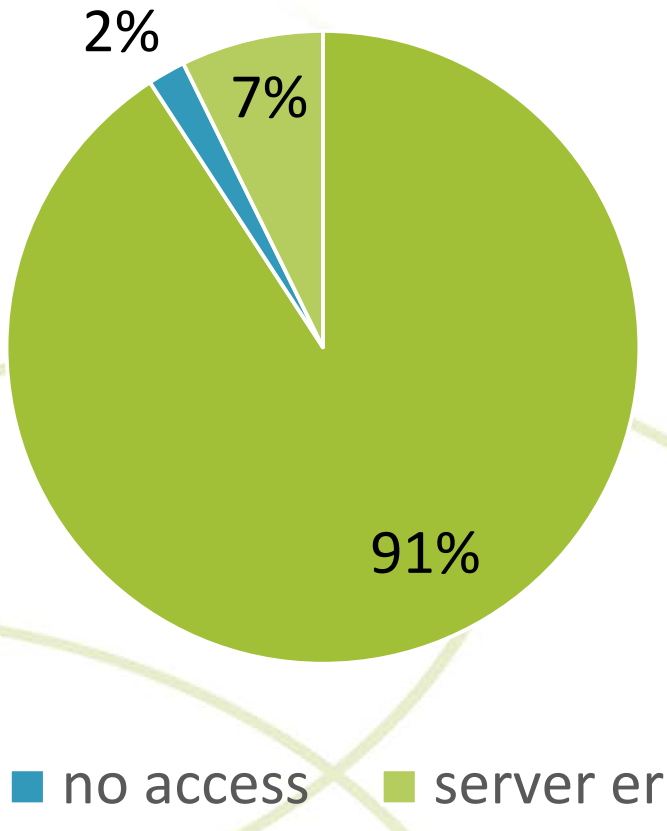
Short round-up 2018: Curation (1)

- Large metadata curation efforts lead to a better experience for the **Resource Type** facet:
  - **coverage** increased from 37% to **95% of all records**.
  - The **number of distinct values** went down from 346 to **53**.



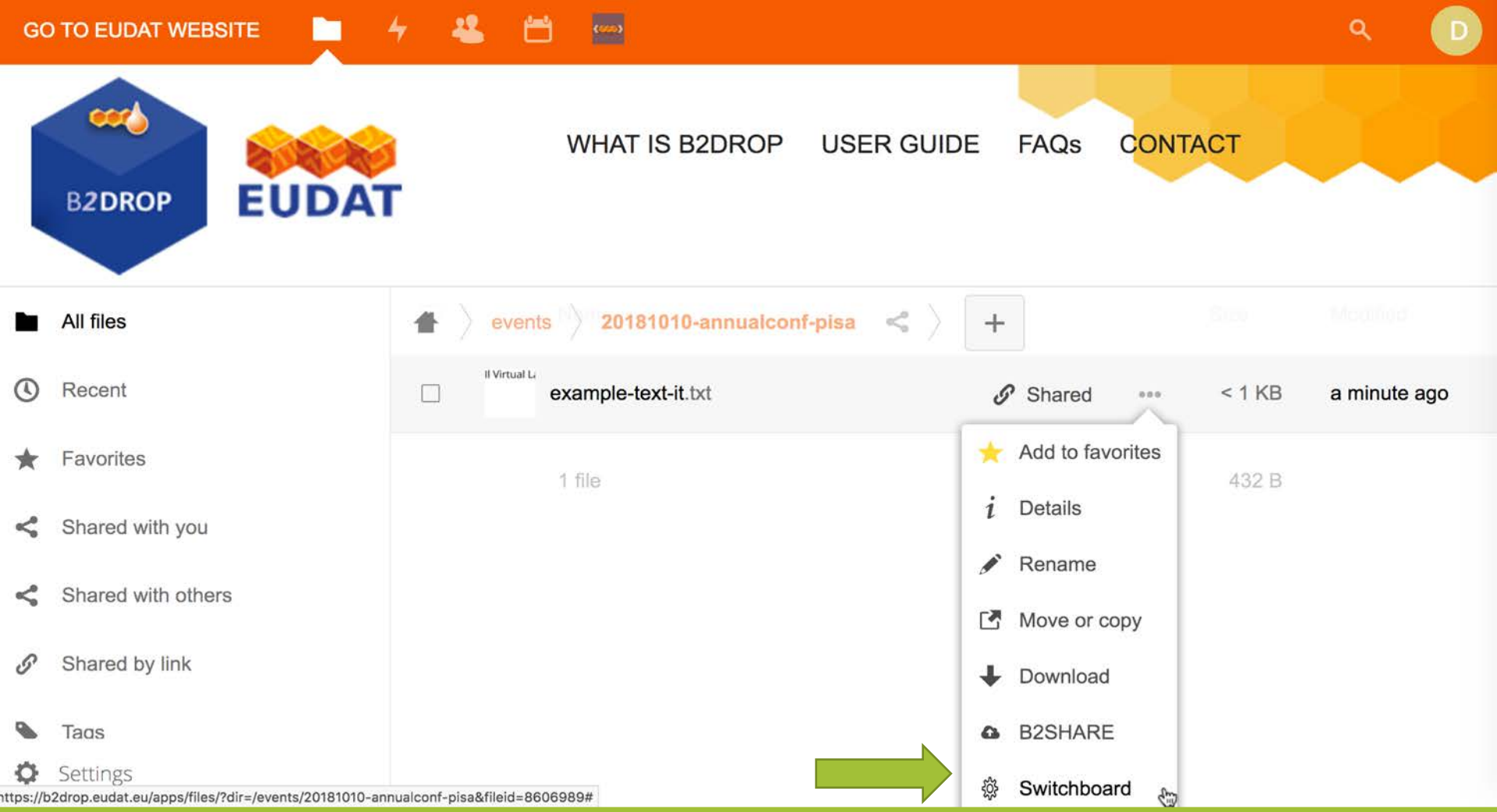
Short round-up 2018: Curation (2)

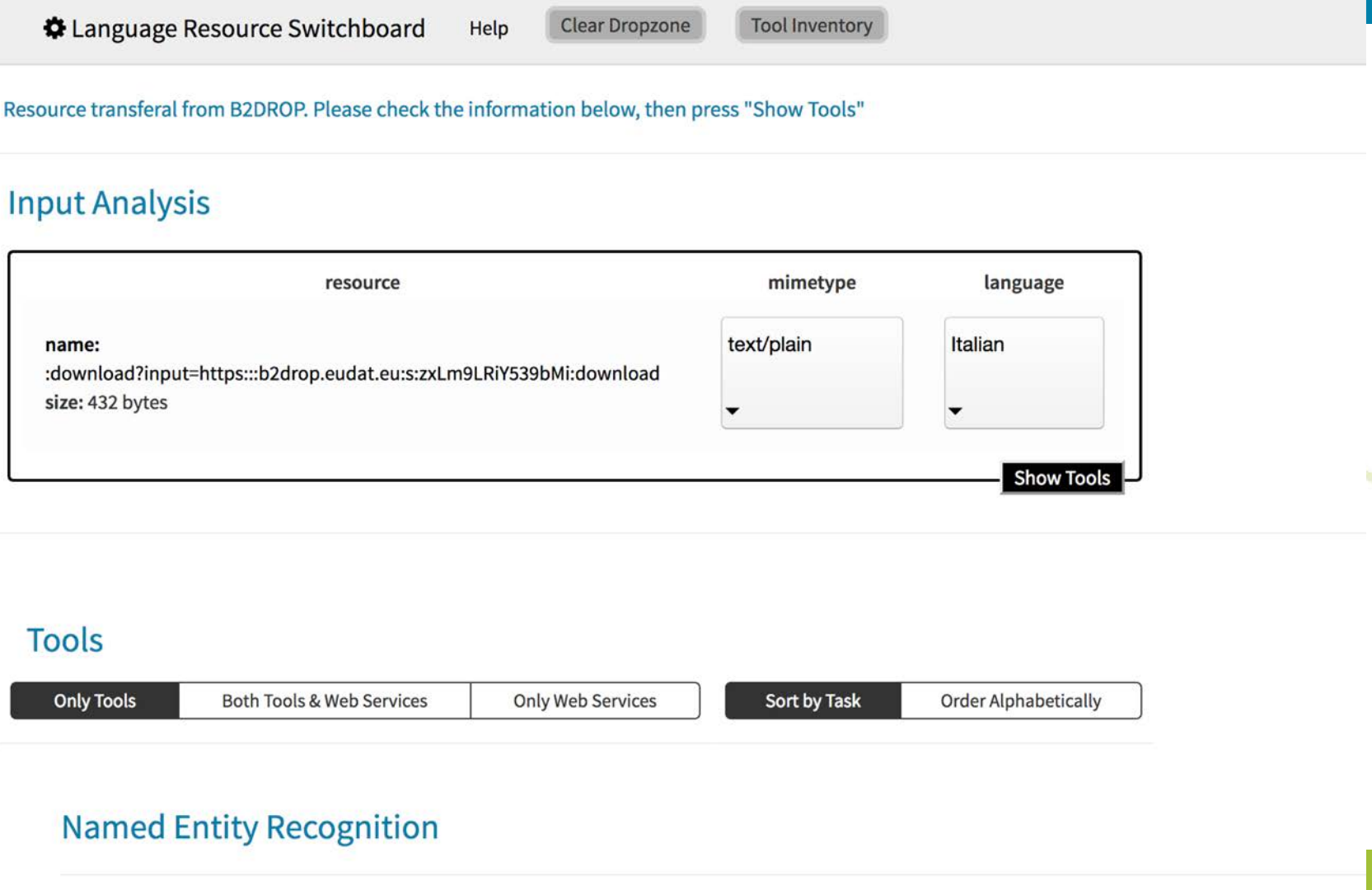
- First statistics by the “kings of curation” at ACDH (Matej, Can, Wolfgang)
- URL checks on **3.9 million URLs** from the VLO



Short round-up 2018: Switchboard

- Transitioning from a beta service into a stable production version: **switchboard.clarin.eu**
- Addition of many tools, e.g. UDPipe
- Now also integrated into the EUDAT cloud storage platform (B2DROP)


GO TO EUDAT WEBSITE
D
B2DROP
EUDAT
WHAT IS B2DROP
USER GUIDE
FAQs
CONTACT
All files
Recent
Favorites
Shared with you
Shared with others
Shared by link
Tags
Settings
events
20181010-annualconf-pisa
example-text-it.txt
Shared
< 1 KB
a minute ago
1 file
432 B
Add to favorites
Details
Rename
Move or copy
Download
B2SHARE
Switchboard
https://b2drop.eudat.eu/apps/files/?dir=/events/20181010-annualconf-pisa&fileid=8606989#


Language Resource Switchboard
Help
Clear Dropzone
Tool Inventory
Resource transferal from B2DROP. Please check the information below, then press "Show Tools"
Input Analysis
resource
mimetype
language
name:
:download?input=https:::b2drop.eudat.eu:s:zxLm9LRiY539bMi:download
size: 432 bytes
text/plain
Italian
Show Tools
Tools
Only Tools
Both Tools & Web Services
Only Web Services
Sort by Task
Order Alphabetically
Named Entity Recognition

Dependency Parsing

UDPIPE

UDPipe is an trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given only annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks.

http://ufal.mff.cuni.cz/udpipe

no

CoNLL-U Format

Click to start tool

Charles University, Prague, Czech Republic

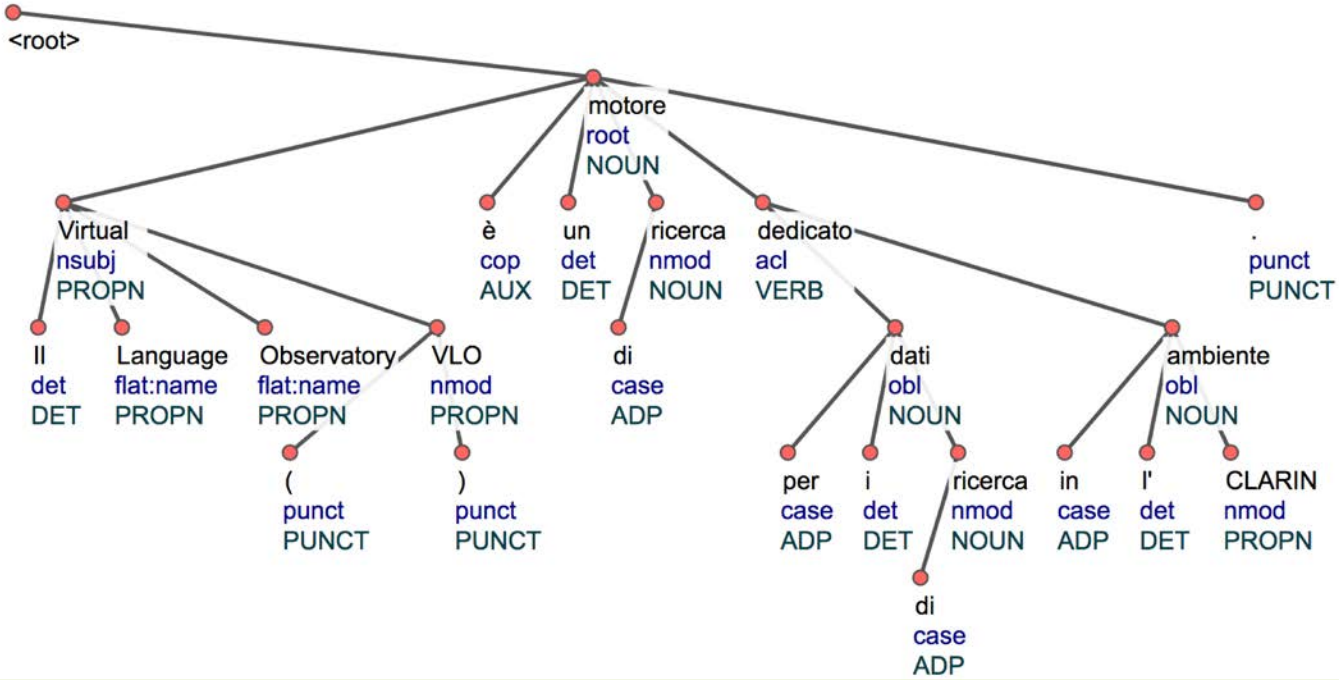
straka@ufal.mff.cuni.cz

A Output Text | Show Table | Show Trees

Save Tree as SVG

Previous 1 2 3 4 Next

Il Virtual Language Observatory (VLO) è un motore di ricerca dedicato per i dati di ricerca in l' ambiente CLARIN .



Short round-up 2018: Back-end

- Public availability reporting for the central services is in place.
  - See: **status.clarin.eu**
  - The average uptime for the period September 2017 to August 2018 for 9 central services is **99.91%**
- Central reverse proxy for the clarin.eu domain:
  - including a geographically redundant setup to avoid a single-point-of-failure
  - currently used for 38 HTTPS-enabled subdomains

A closer look at the CLARIN Identity Provider

CLARIN IdP - numbers

- Based on account requests from September 2017 to begin October 2018
- 505 requests
  - For 176 of these the source and target country could be determined automatically
  - **62%** of these account requests were **cross-border requests**

CLARIN IdP – stories from students

- PhD in **information extraction** from electronic **medical records**
- **Tokyo**: I would like to use a corpus for my Bachelor Thesis. The thesis focuses on a phenomenon in **Norwegian** where sentence agreement does not follow conventional linguistic patterns.
- **Tanzania**: to use **Swahili** corpus for my Phd Research
- **USA**: I would like to submit my data to one of the corpora titled "**Tangsa, Tai, Singpho** in North East India" under the DoBes Archive in The Language Archive Corpora at MPI.

CLARIN IdP – stories from students

- **Taiwan**: I am a doctoral student and currently in a research of personal attributes and **emotion identification via speech**. I found the Nautilus Speaker Characterization Corpus has great potential for multiple research topics.
- **Colombia**: Soy aspirante al programa de **Antropologia** de la universidad de Antioquia y tengo una fascinación por los estudios lingüísticos.
- **Hongkong**: I am carrying out research on gestures and the clitic 'se' in the oral text of Spanish native speakers. I would like to have access to the **Hamburg Corpus of Argentinean Spanish** (HaCASpa) to compare speakers of different Spanish variations.

CLARIN IdP – stories from citizen scientists

- Citizen scientist **contributing** to a research project at Radboud University, *English in the Dutch language*.
- I am **learning** the Python **programming** language, and since I am also interested in natural language I would like to have access to a large corpus of text. I want to use this to run various analysis. Project Gutenberg would be another option, but those texts are all very outdated...

CLARIN IdP – stories from language learners & teachers

- I want to **improve my language** by looking at how words can be composed
- I would like to use the data of COLT to write a dissertation in the university. The dissertation will be about education, **learning English as a second language in elementary school in Japan**.
- I am studying **literary translation** English-Dutch, so I am interested in correct and usual expressions

CLARIN IdP – stories from policy makers

- I work at the **research department of the municipality** […]. We are currently exploring the possibilities of NLP models. We have a lot of **text data that we don't use at the moment**, because we don't know how to obtain valuable information out of it. […] We make our research results available for the general public.
- **Ministry of education and research**: Test the infrastructure, and also do part of speech tagging and lemmatisation of **government documents**.

CLARIN IdP – even more stories

- **Retired** professor
- I'm working in a private foundation **developing a speech corpus for the Swedish dialects spoken in Finland**: Talko. It's very important for me to get access to similar corpora on other languages.
- Australia - Catholic **Theological** College: I am a graduate student studying various texts on Thomas Aquinas.
- We investigate at the **University Clinic Freiburg** speech intelligibility of hearing impaired patients provided with **cochlear implants** in challenging noise conditions. To create real noise conditions we want to generate a multitalker environment. Therefore we would like to use some of the recorded sentences provided within the Clarin infrastructure.

Conclusions

- There is a broad interest in accessing language resources via CLARIN.
- Many motivations and usage scenarios
- Cross-border resource access is a reality

Acknowledgements

- The assessment committee: Lene, Cyprian, Daan, Jozef, Riccardo, Tomas
- The CLARIN ERIC central developer team: André, Hendrik, Menzo, Twan, Willem
- Developers and contributors from the national consortia: Can, Claus, Leif-Jöran, Matej, Thomas, Wolfgang
- The taskforces: AAI, CMDI & Curation (Susanne, Hanna, Alexander, …), FCS, PIDs
- And all the others who are contributing to the construction of the infrastructure!

Thank you for your attention!

- Questions?