The CLARIN Language Resource Switchboard

Claus Zinn
Seminar für Sprachwissenschaft
Universität Tübingen

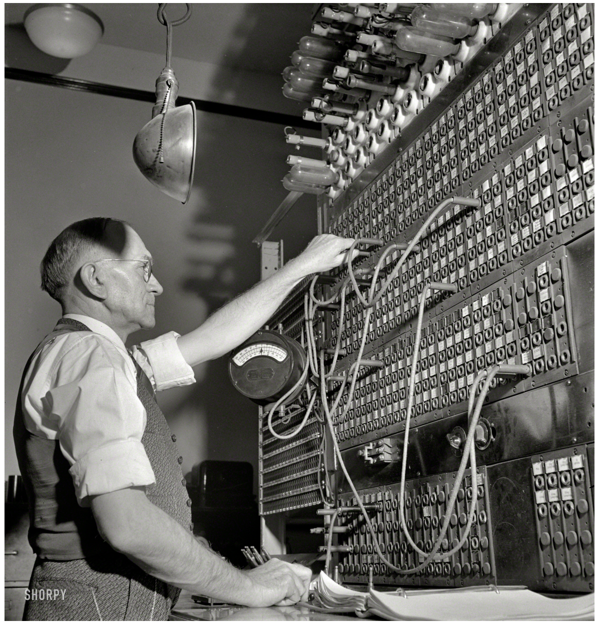
CLARIN 2016 Annual Conference
October 26-28, 2016, Aix-en-Provence, France

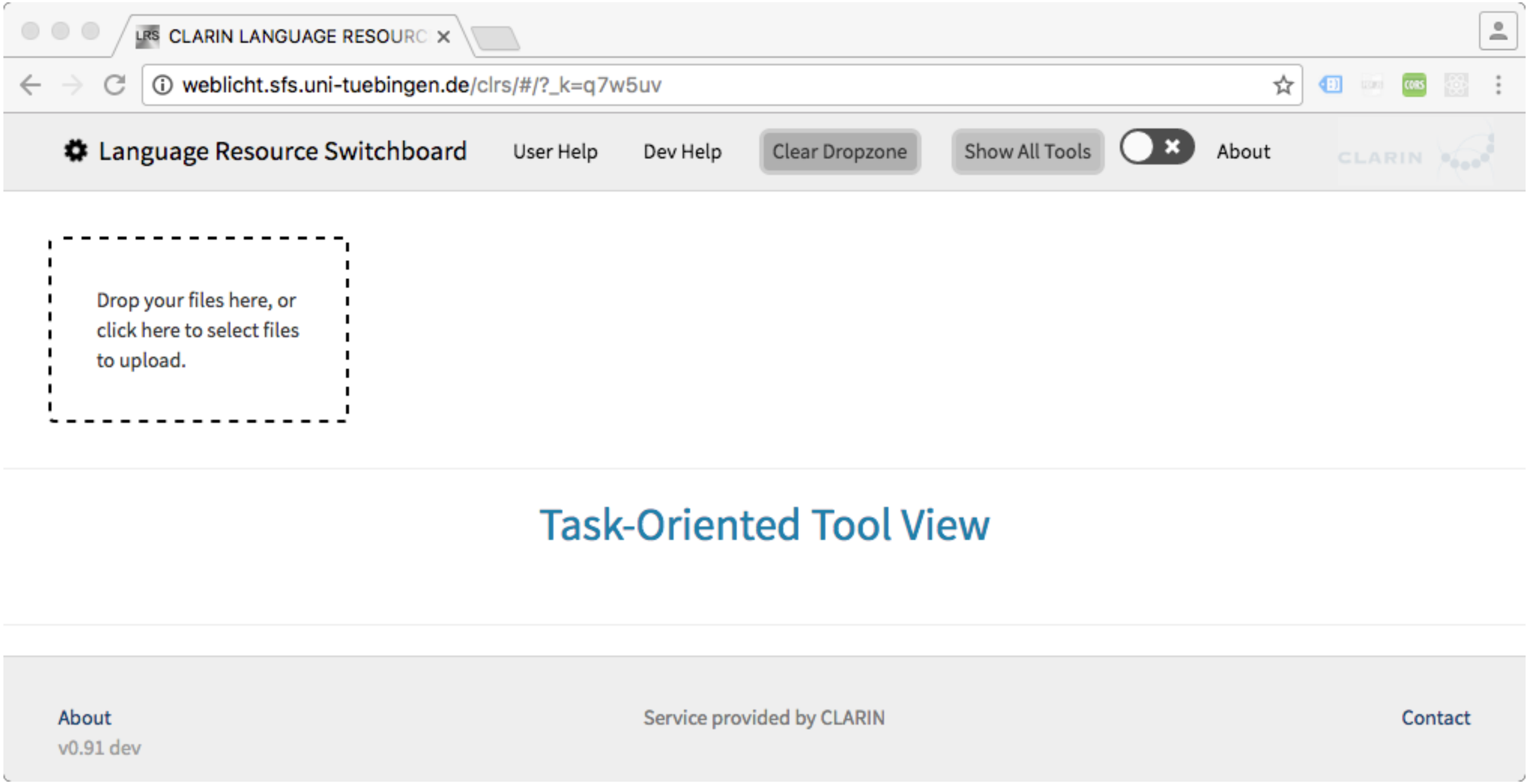
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

CLARIN ERIC
Common Language Resources and Technology Infrastructure

Motivation

- improve user experience for VLO (FCS, VCR) users
- users experience a gap between the resources they find, and the processing workflows they would like to call
- for each resource found, LRS suggests applicable web-based tools (and web services) that can process the resource
- users can invoke applicable tool by simple click
- relevant information forwarded to tool in question
- switchboard has achieved its task, it does no more




CLARIN LANGUAGE RESOURC
weblicht.sfs.uni-tuebingen.de/clrs/#/?_k=q7w5uv
Language Resource Switchboard
User Help
Dev Help
Clear Dropzone
Show All Tools
About
CLARIN

Drop your files here, or click here to select files to upload.

Task-Oriented Tool View

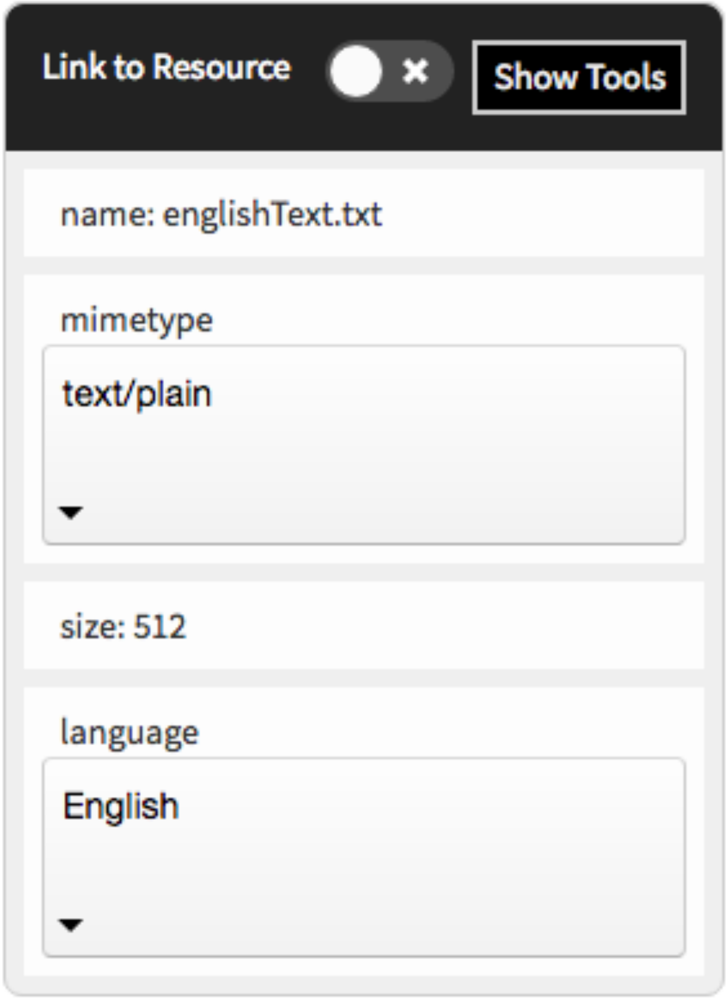
About
v0.91 dev
Service provided by CLARIN
Contact

Switchboard

Standalone Environment

http://weblicht.sfs.uni-tuebingen.de/clrs

Placing a file into the drop-zone



name: englishText.txt

mimetype

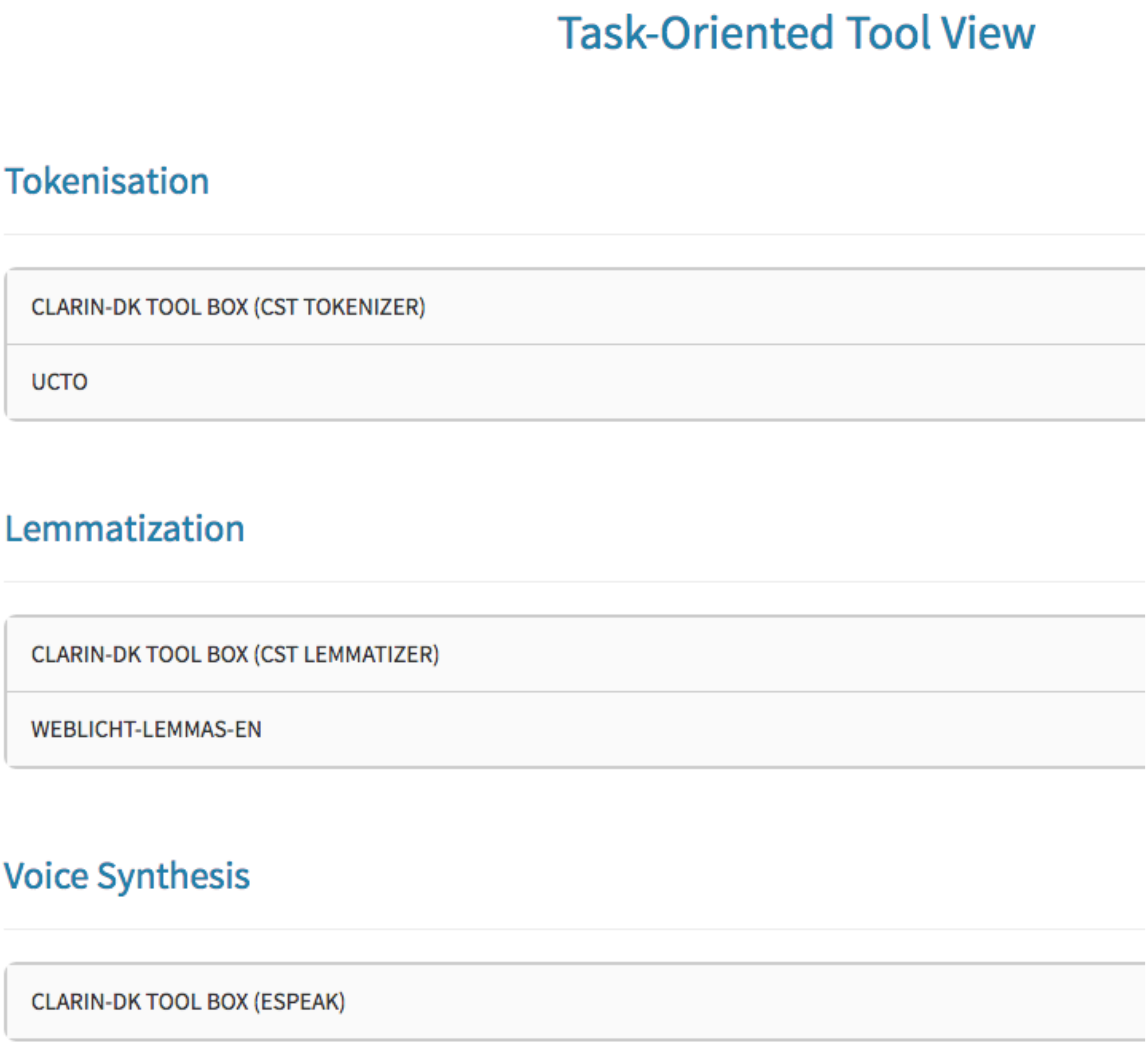
text/plain

size: 512

language

English

- file uploaded on file upload server (RZG)
- identification of mime-type and language, but this may fail
- users should review information before clicking on ‘Show Tools’
- set switch to include web services



Task-Oriented Tool View

Tokenisation

- CLARIN-DK TOOL BOX (CST TOKENIZER)
- UCTO

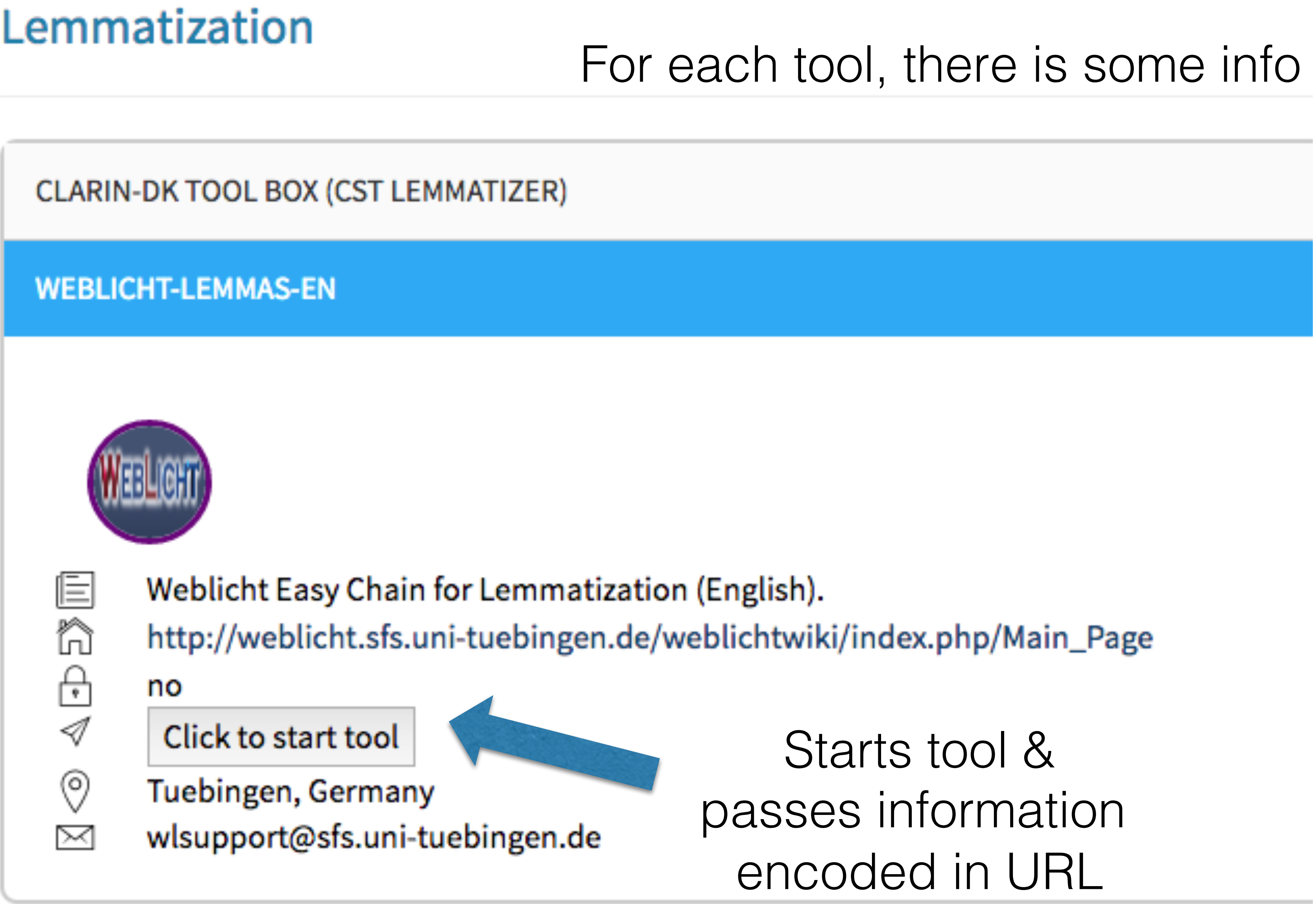
Lemmatization

- CLARIN-DK TOOL BOX (CST LEMMATIZER)
- WEBLICHT-LEMMAS-EN

Voice Synthesis

- CLARIN-DK TOOL BOX (ESPEAK)

For each tool, there is some info



Lemmatization

CLARIN-DK TOOL BOX (CST LEMMATIZER)

WEBLICHT-LEMMAS-EN

Weblicht Easy Chain for Lemmatization (English).

http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

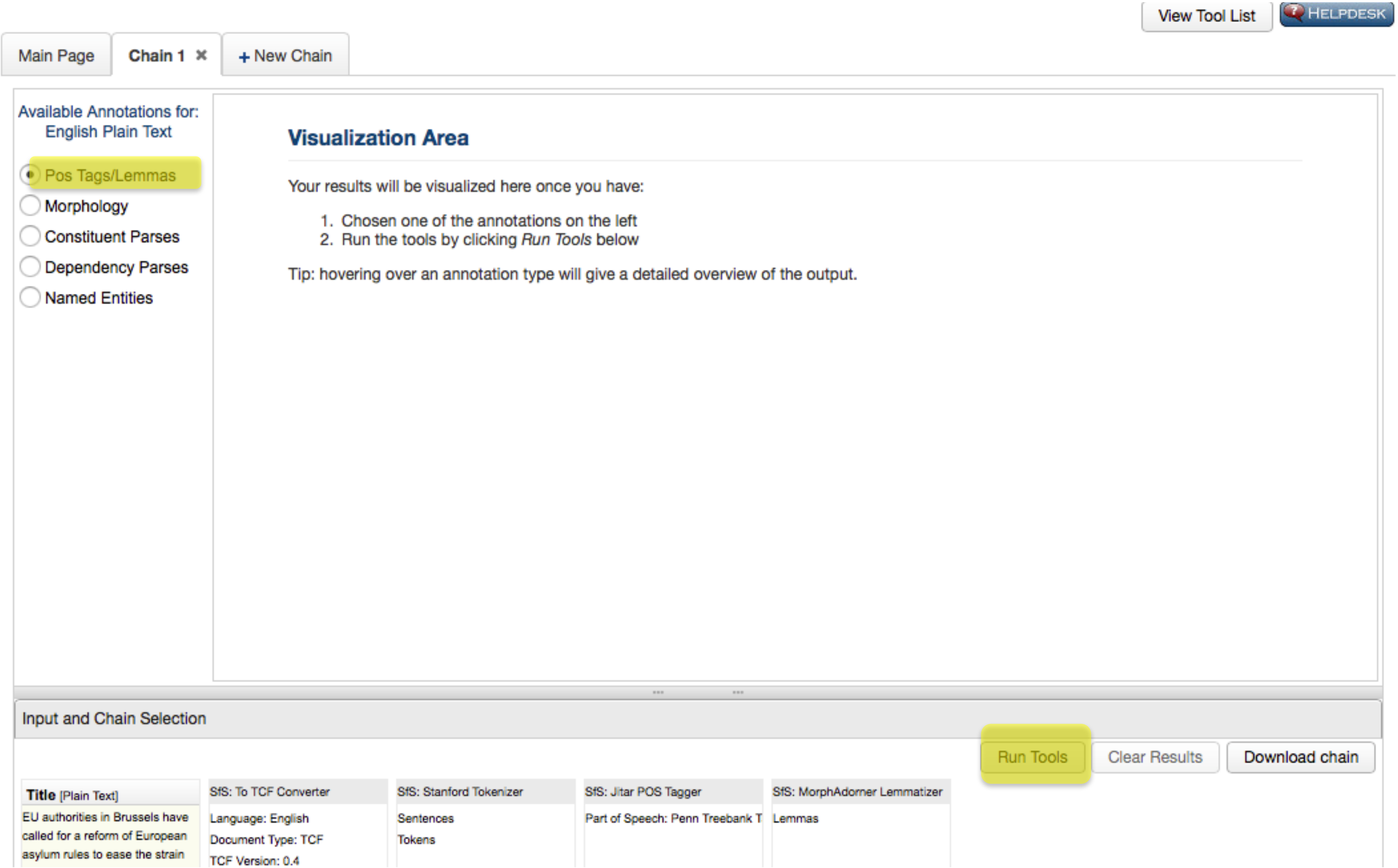
no

Click to start tool

Tuebingen, Germany

wlsupport@sfs.uni-tuebingen.de

Starts tool &
passes information
encoded in URL



Visualization Area

Your results will be visualized here once you have:

1. Chosen one of the annotations on the left
2. Run the tools by clicking *Run Tools* below

Tip: hovering over an annotation type will give a detailed overview of the output.

Title [Plain Text]
EU authorities in Brussels have called for a reform of European asylum rules to ease the strain

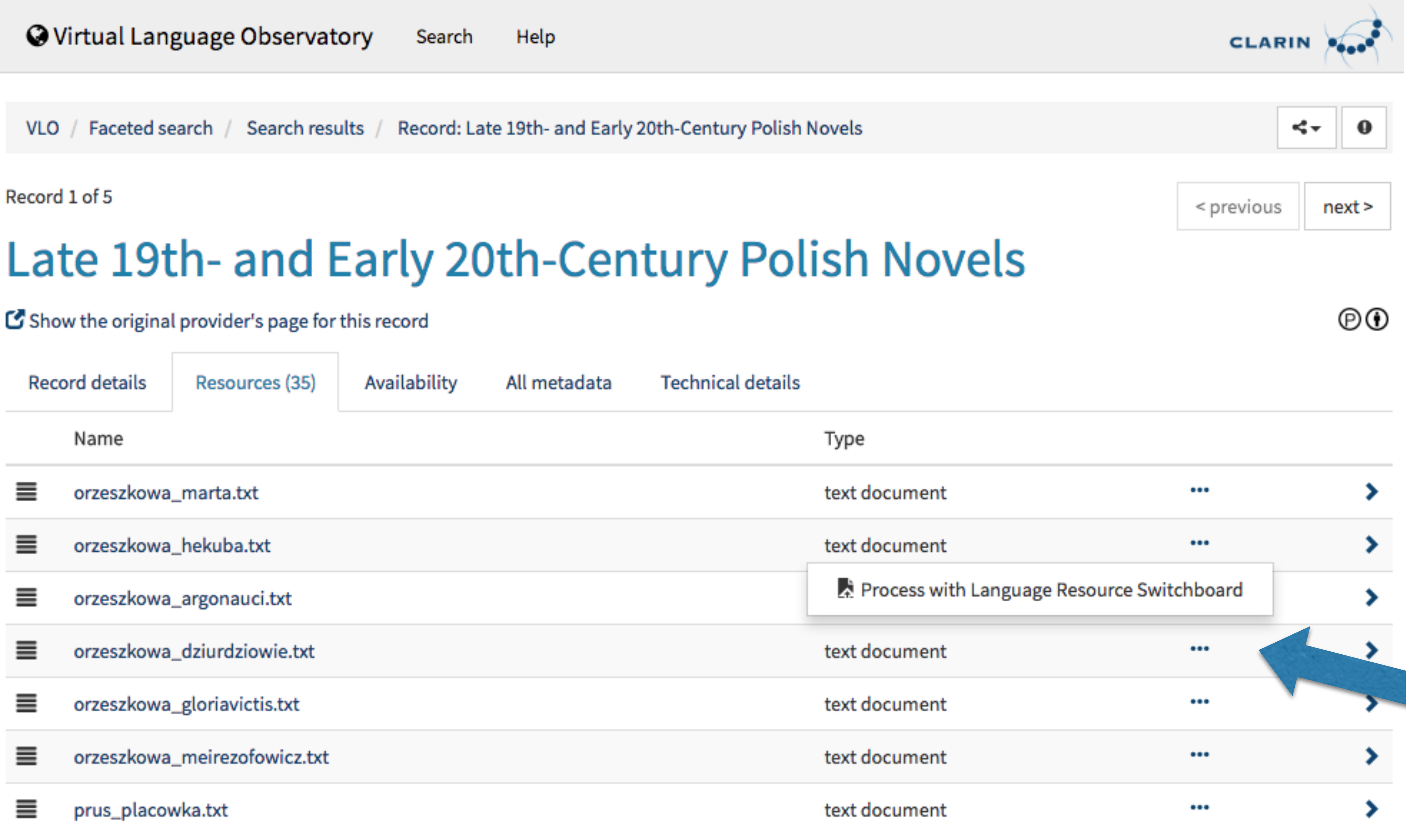
SfS: To TCF Converter
Language: English
Document Type: TCF
TCF Version: 0.4

SfS: Stanford Tokenizer
Sentences
Tokens

SfS: Jitar POS Tagger
Part of Speech: Penn Treebank T

SfS: MorphAdorner Lemmatizer
Lemmas

Weblicht called at the intended entry point.



Virtual Language Observatory Search Help CLARIN

VLO / Faceted search / Search results / Record: Late 19th- and Early 20th-Century Polish Novels

Record 1 of 5

< previous next >

Late 19th- and Early 20th-Century Polish Novels

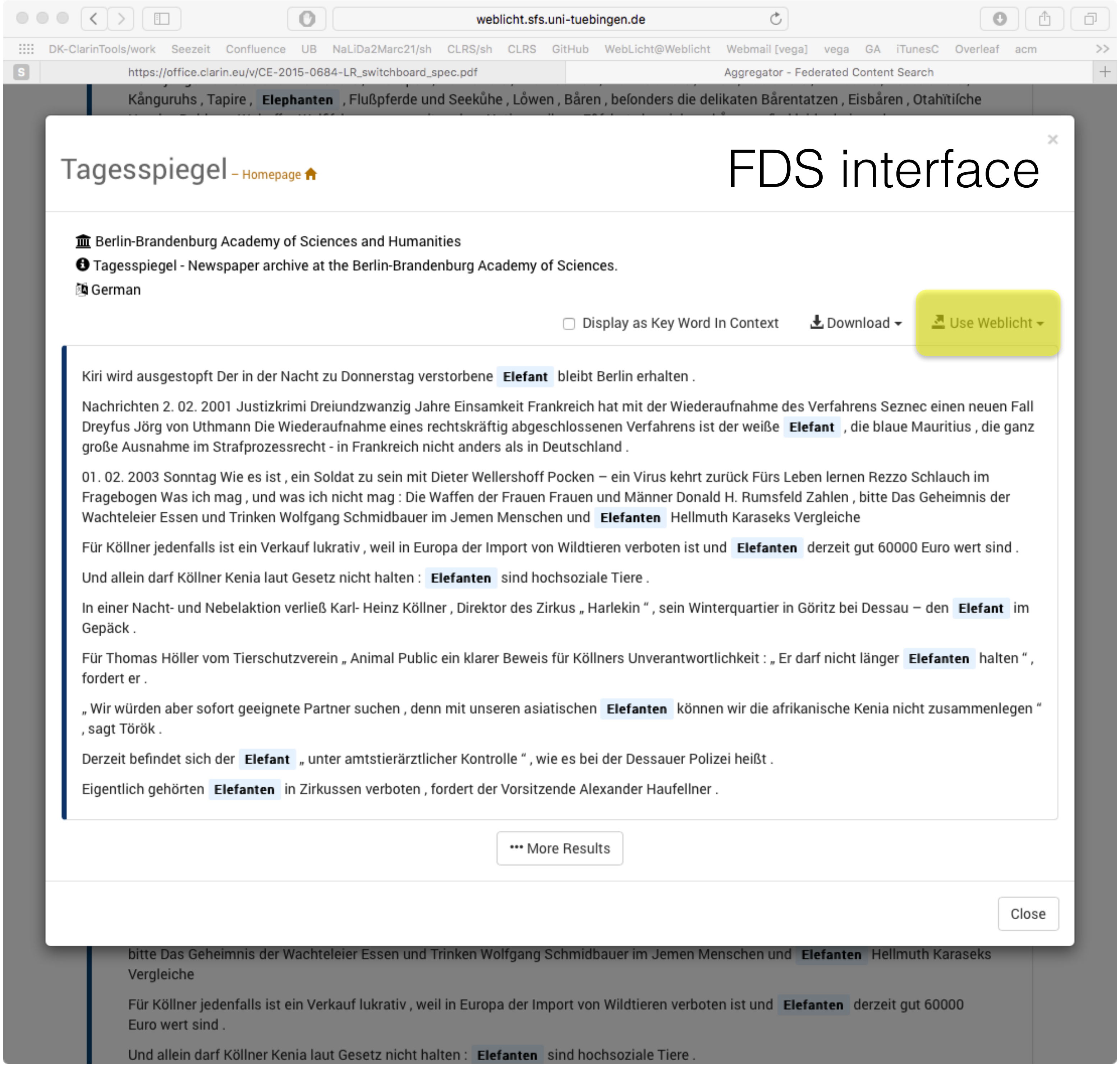
Show the original provider's page for this record

Record details | Resources (35) | Availability | All metadata | Technical details

Name	Type
orzeszkowa_marta.txt	text document
orzeszkowa_hekuba.txt	text document
orzeszkowa_argonauci.txt	
orzeszkowa_dziurdziowie.txt	text document
orzeszkowa_gloriavictis.txt	text document
orzeszkowa_meirezofowicz.txt	text document
prus_placowka.txt	text document

Process with Language Resource Switchboard

FDS interface



Kånguruhs , Tapire , **Elephanten** , Flußpferde und Seekühe , Löwen , Bären , beſonders die delikaten Bårentatzen , Eisbären , Otahitiſche

Tagesspiegel - Homepage

Berlin-Brandenburg Academy of Sciences and Humanities

Tagesspiegel - Newspaper archive at the Berlin-Brandenburg Academy of Sciences.

German

Kiri wird ausgestopft Der in der Nacht zu Donnerstag verstorbene **Elefant** bleibt Berlin erhalten .

Nachrichten 2. 02. 2001 Justizkrimi Dreiundzwanzig Jahre Einsamkeit Frankreich hat mit der Wiederaufnahme des Verfahrens Seznec einen neuen Fall Dreyfus Jörg von Uthmann Die Wiederaufnahme eines rechtskräftig abgeschlossenen Verfahrens ist der weiße **Elefant** , die blaue Mauritius , die ganz große Ausnahme im Strafprozessrecht - in Frankreich nicht anders als in Deutschland .

01. 02. 2003 Sonntag Wie es ist , ein Soldat zu sein mit Dieter Wellershoff Pocken – ein Virus kehrt zurück Fürs Leben lernen Rezzo Schlauch im Fragebogen Was ich mag , und was ich nicht mag : Die Waffen der Frauen Frauen und Männer Donald H. Rumsfeld Zahlen , bitte Das Geheimnis der Wachteleier Essen und Trinken Wolfgang Schmidbauer im Jemen Menschen und **Elefanten** Hellmuth Karaseks Vergleiche

Für Köllner jedenfalls ist ein Verkauf lukrativ , weil in Europa der Import von Wildtieren verboten ist und **Elefanten** derzeit gut 60000 Euro wert sind .

Und allein darf Köllner Kenia laut Gesetz nicht halten : **Elefanten** sind hochsoziale Tiere .

In einer Nacht- und Nebelaktion verließ Karl- Heinz Köllner , Direktor des Zirkus „ Harlekin “ , sein Winterquartier in Göritz bei Dessau – den **Elefant** im Gepäck .

Für Thomas Höller vom Tierschutzverein „ Animal Public ein klarer Beweis für Köllners Unverantwortlichkeit : „ Er darf nicht länger **Elefanten** halten “ , fordert er .

„ Wir würden aber sofort geeignete Partner suchen , denn mit unseren asiatischen **Elefanten** können wir die afrikanische Kenia nicht zusammenlegen “ , sagt Török .

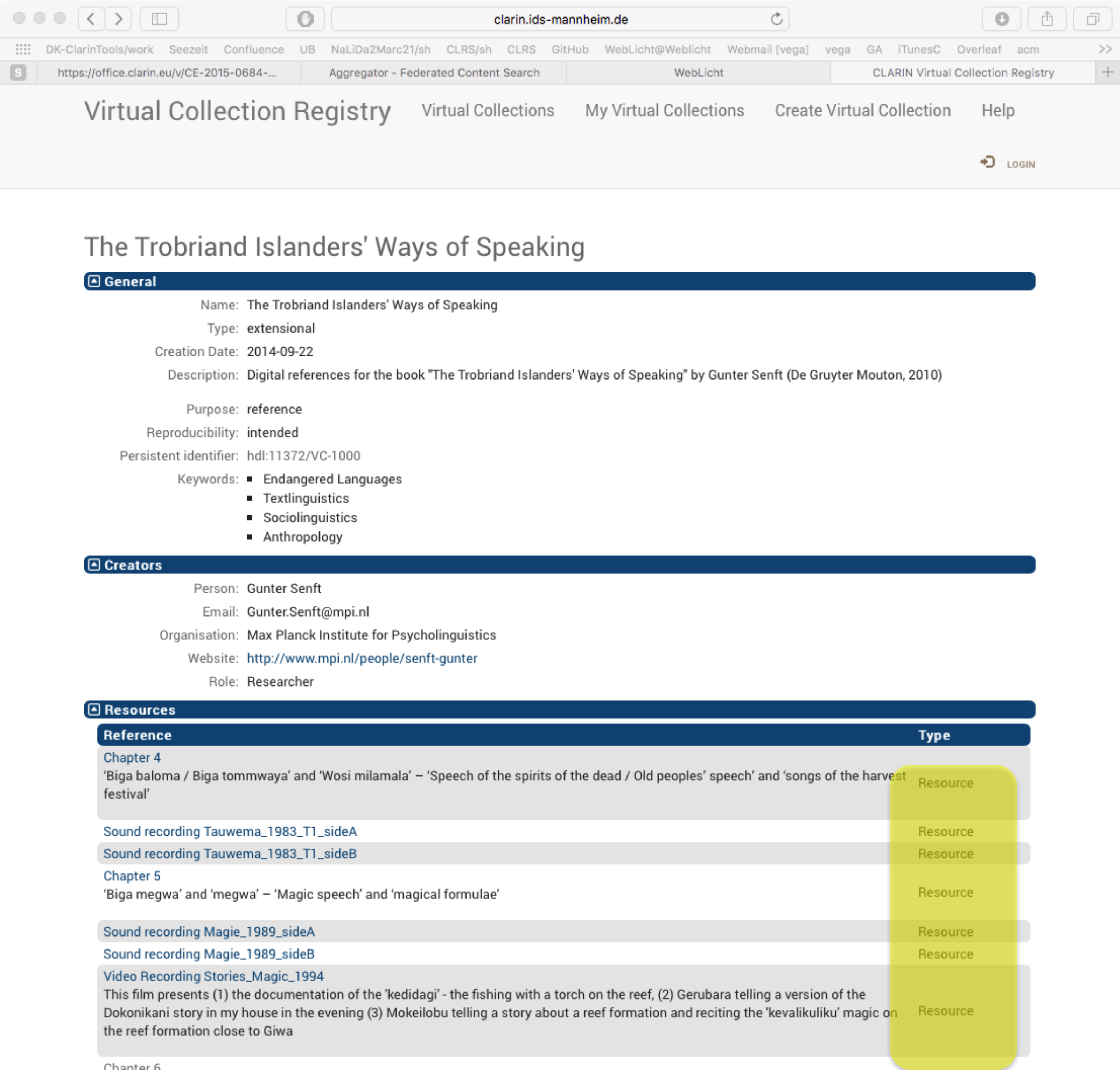
Derzeit befindet sich der **Elefant** „ unter amtstierärztlicher Kontrolle “ , wie es bei der Dessauer Polizei heißt .

Eigentlich gehörten **Elefanten** in Zirkussen verboten , fordert der Vorsitzende Alexander Haufellner .

bitte Das Geheimnis der Wachteleier Essen und Trinken Wolfgang Schmidbauer im Jemen Menschen und **Elefanten** Hellmuth Karaseks Vergleiche

Für Köllner jedenfalls ist ein Verkauf lukrativ , weil in Europa der Import von Wildtieren verboten ist und **Elefanten** derzeit gut 60000 Euro wert sind .

Und allein darf Köllner Kenia laut Gesetz nicht halten : **Elefanten** sind hochsoziale Tiere .



The Trobriand Islanders' Ways of Speaking

General

Name: The Trobriand Islanders' Ways of Speaking

Type: extensional

Creation Date: 2014-09-22

Description: Digital references for the book "The Trobriand Islanders' Ways of Speaking" by Gunter Senft (De Gruyter Mouton, 2010)

Purpose: reference

Reproducibility: intended

Persistent identifier: hdl:11372/VC-1000

Keywords:
- Endangered Languages
- Textlinguistics
- Sociolinguistics
- Anthropology

Creators

Person: Gunter Senft

Email: Gunter.Senft@mpi.nl

Organisation: Max Planck Institute for Psycholinguistics

Website: http://www.mpi.nl/people/senft-gunter

Role: Researcher

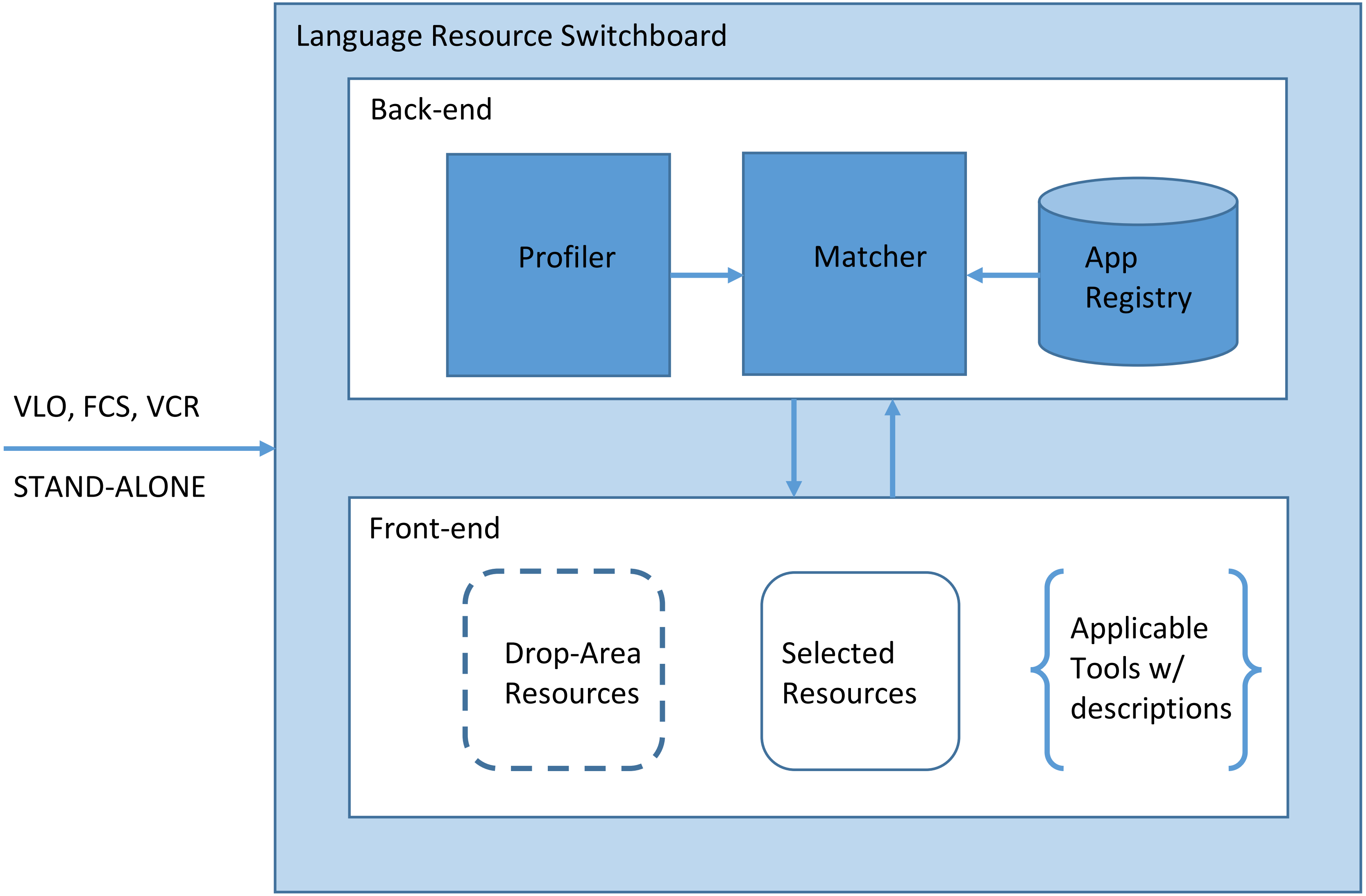
Resources

Reference	Type
Chapter 4 'Biga baloma / Biga tommwaya' and 'Wosi milamala' – 'Speech of the spirits of the dead / Old peoples' speech' and 'songs of the harvest festival'	Resource
Sound recording Tauwema_1983_T1_sideA	Resource
Sound recording Tauwema_1983_T1_sideB	Resource
Chapter 5 'Biga megwa' and 'megwa' – 'Magic speech' and 'magical formulae'	Resource
Sound recording Magie_1989_sideA	Resource
Sound recording Magie_1989_sideB	Resource
Video Recording Stories_Magic_1994 This film presents (1) the documentation of the 'kedidagi' - the fishing with a torch on the reef, (2) Gerubara telling a version of the Dokonikani story in my house in the evening (3) Mokeilobu telling a story about a reef formation and reciting the 'kevalikuliku' magic on the reef formation close to Giwa	Resource
Chapter 6	

Specification

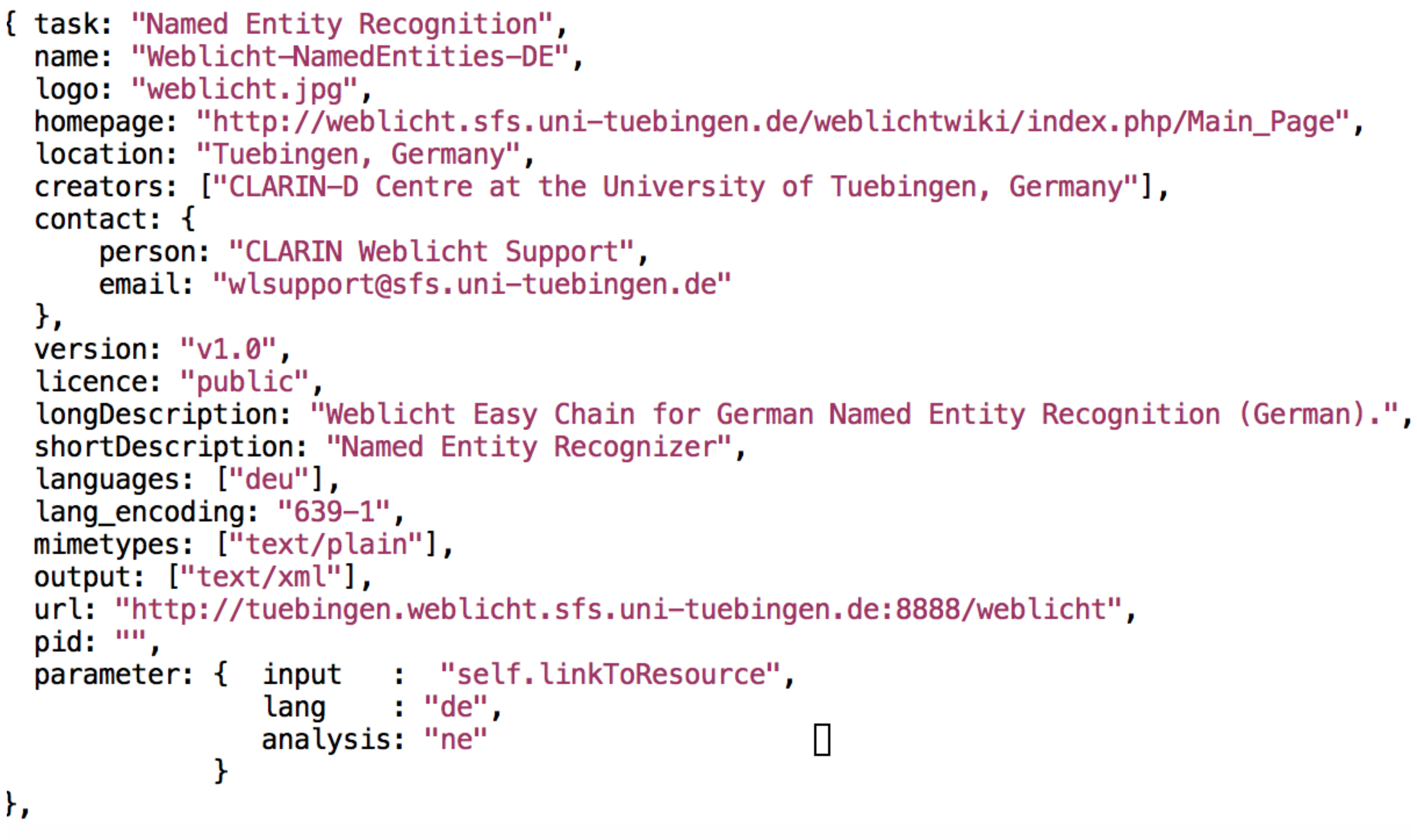
- specification document @ https://office.clarin.eu/v/CE-2015-0684-LR_switchboard_spec.pdf
- Participating tool providers give metadata description of their tool (format below) ➢ feeds into app registry
- LRS has a ➢ profiler that extracts all relevant information from a resource' metadata description (and does more)
- LRS has a ➢ matcher that matches metadata description of resource with metadata description of tool
  - mime type of the resource
  - language of the resource
- LRS has a ➢ front-end that displays applicable tools, ordered by the analyses the tool provide, and where a chosen tool can be invoked

Architecture



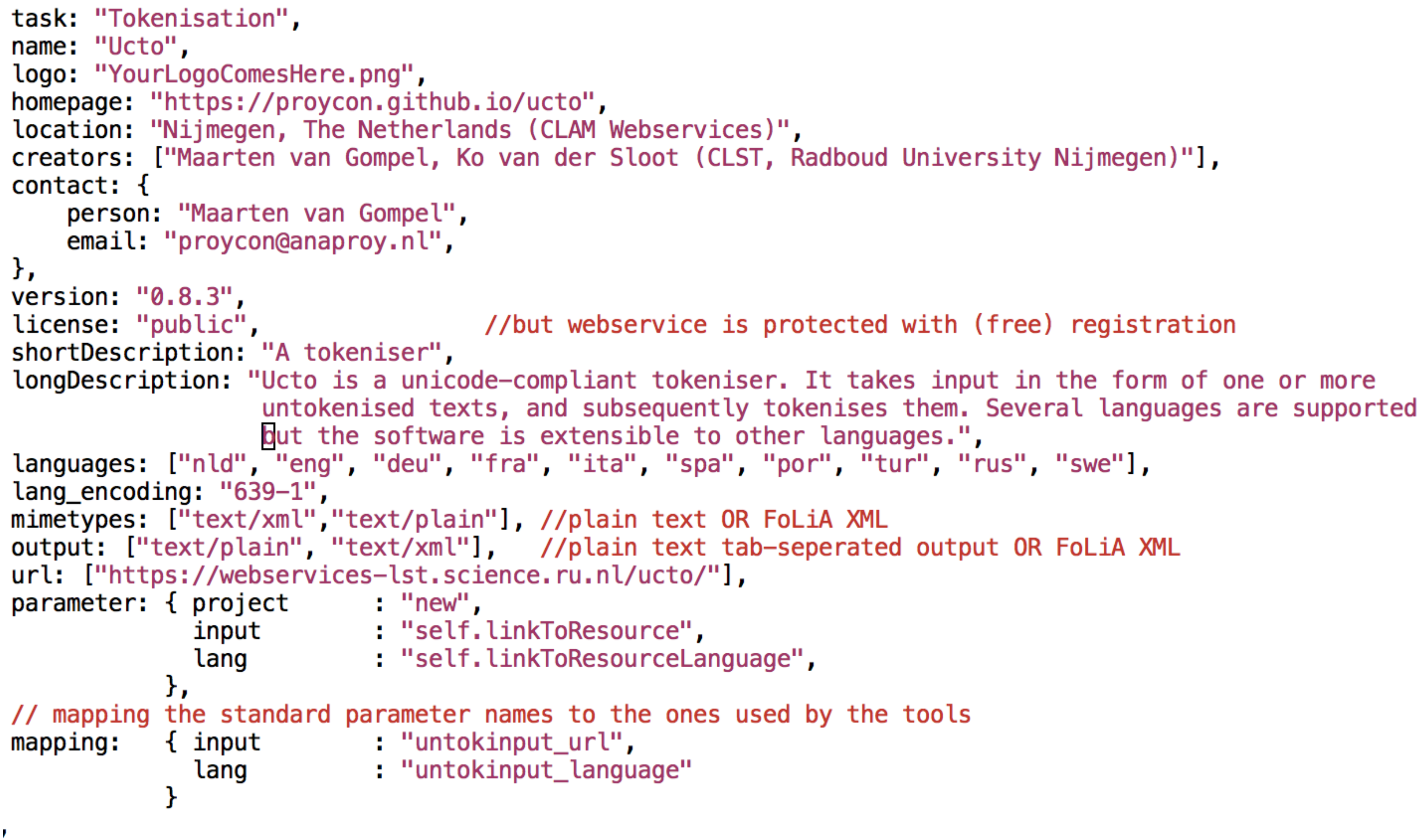
Technology

- Front-end: ReactJS +Altjs container
- Back-end:
  - app registry: JSON structure
  - matcher: simple JS
    - mime type and language main two criteria
  - profiler: supported by LRS users
    - (relevant information taken from resource metadata in VLO)
    - plus mime type detection during dnd (standalone version)
    - plus language detection using TIKA (tika.apache.org)
    - plus user intervention



```
{ task: "Named Entity Recognition",
  name: "Weblicht-NamedEntities-DE",
  logo: "weblicht.jpg",
  homepage: "http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page",
  location: "Tuebingen, Germany",
  creators: ["CLARIN-D Centre at the University of Tuebingen, Germany"],
  contact: {
      person: "CLARIN Weblicht Support",
      email: "wlsupport@sfs.uni-tuebingen.de"
  },
  version: "v1.0",
  licence: "public",
  longDescription: "Weblicht Easy Chain for German Named Entity Recognition (German).",
  shortDescription: "Named Entity Recognizer",
  languages: ["deu"],
  lang_encoding: "639-1",
  mimetypes: ["text/plain"],
  output: ["text/xml"],
  url: "http://tuebingen.weblicht.sfs.uni-tuebingen.de:8888/weblicht",
  pid: "",
  parameter: {  input    :  "self.linkToResource",
                lang     : "de",
                analysis: "ne"                     ▯
             }
},
```

Metadata Entry



```
task: "Tokenisation",
name: "Ucto",
logo: "YourLogoComesHere.png",
homepage: "https://proycon.github.io/ucto",
location: "Nijmegen, The Netherlands (CLAM Webservices)",
creators: ["Maarten van Gompel, Ko van der Sloot (CLST, Radboud University Nijmegen)"],
contact: {
    person: "Maarten van Gompel",
    email: "proycon@anaproy.nl",
},
version: "0.8.3",
license: "public",                //but webservice is protected with (free) registration
shortDescription: "A tokeniser",
longDescription: "Ucto is a unicode-compliant tokeniser. It takes input in the form of one or more
                  untokenised texts, and subsequently tokenises them. Several languages are supported
                  but the software is extensible to other languages.",
languages: ["nld", "eng", "deu", "fra", "ita", "spa", "por", "tur", "rus", "swe"],
lang_encoding: "639-1",
mimetypes: ["text/xml","text/plain"], //plain text OR FoLiA XML
output: ["text/plain", "text/xml"],   //plain text tab-seperated output OR FoLiA XML
url: ["https://webservices-lst.science.ru.nl/ucto/"],
parameter: { project      : "new",
             input        : "self.linkToResource",
             lang         : "self.linkToResourceLanguage",
           },
// mapping the standard parameter names to the ones used by the tools
mapping:   { input        : "untokinput_url",
             lang         : "untokinput_language"
           }
,
```

Metadata Entry (II)

Calling LRS with parameters

- apart from standalone app (dev vehicle), LRS can be invoked by passing relevant information as URL-encoded parameters

- http://weblicht.sfs.uni-tuebingen.de/clrs/#/vlo/http:%2F%2Fhdl.handle.net%2F10932%2F00-01B8-AF59-4FB9-9201-B%09/text%2Fplain/spa

1. link to resource

2. mime type

3. language

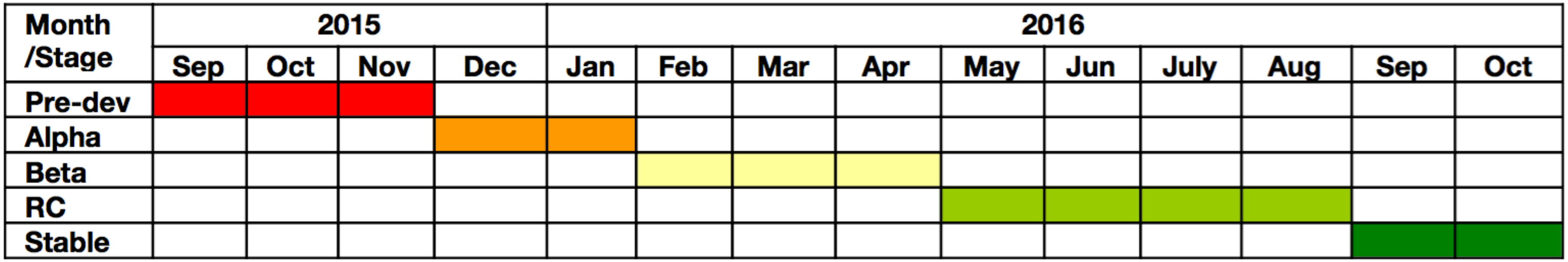
Encoding
for slash

State of the LRS

- 35 tools entered with their metadata
  - 5 tools from CLARIN-DK Tools website
  - 8 tools from WebLicht predefined chains for eng, deu, nld, tur
  - 12 tools from LST Webservice Portal, CSLT, Radboud Univ.
- for a dozen different tasks:
  - tokenization, lemmatization, morphology analysis, POS tagging, constituent parsing, dependency parsing, NER, OCR, voice synthesis, spell-checking, text analytics, machine translation
- a dozen web services (mostly metadata conversion)

Timeline

- next-up: integration with FDS and VCR
- increase participation of more tools
- usability testing
- iterative improvements



Month /Stage	2015				2016									
	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Pre-dev	🟥	🟥	🟥											
Alpha				🟧	🟧									
Beta						🟨	🟨	🟨						
RC									🟩	🟩	🟩	🟩		
Stable													🟩	🟩

Open Issues

- resource metadata given as input to the profiler, not taken for granted but counter-checked. But what metadata exactly?
- handling of large files
  - for checking mime type and language
  - from the tool's perspective and from the usability aspect
- batch mode (multiple file mode)
- controlled vocabulary for tool description
  - *e.g.,* task, license, I/O mime-types (esp. xml sub-types)
- availability of resource (license, authentication)
- VTO mode as part of the LRS and as complement to the VLO

Problematic cases

- https://vlo.clarin.eu/record?4&docId=hdl_58_1839_47_00-39F88B30-54B2-4497-AB6E-2C26BEB71AF8_64_format_61_imdi_64_format_61_cmdi&fq=format:text/plain&fq=languageCode:code:nld&index=1&count=1427

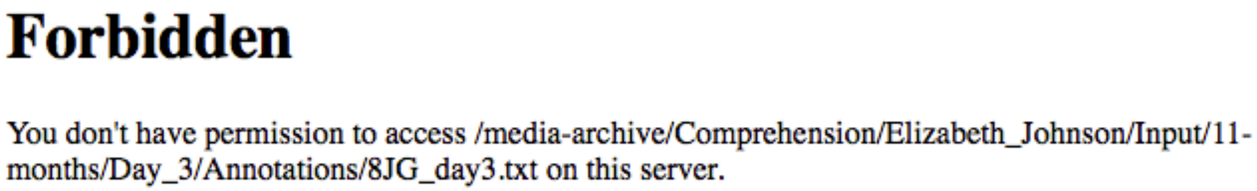
- metadata description misleading

    - zip file does not have all other files

    - incorrect mime-type (xml files rather than plain text)

Problematic cases

- https://corpus1.mpi.nl/media-archive/Comprehension/Elizabeth_Johnson/Input/11-months/Day_3/Annotations/8JG_day3.txt
- after authentification:



Forbidden

You don't have permission to access /media-archive/Comprehension/Elizabeth_Johnson/Input/11-months/Day_3/Annotations/8JG_day3.txt on this server.

Conclusion

- project on course, **1.0 release out (soon)!**
- more input from tool providers appreciated
- LRS needs more tools to be integrated, **so tell others!**
- also promote standalone version, good entry point for your **local** resources