Fedora in the HZSK Infrastructure

Fedora Users Group Meeting, Hamburg, 07.12.2017

Daniel Jettka, daniel.jettka@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, HZSK
Universität Hamburg/Germany

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

HZSK | RESOURCES | TOOLS | SUPPORT

Login

Home » HZSK Repository

Digital Repository for Linguistic Resources and Tools

The digital repository of the Hamburger Zentrum für Sprachkorpora stores and dissem
access and license restrictions, privacy policy, hosting requirements

Filter

Corpus Type	License type	Modality	Language	Keyword
general corpus (24)	restricted (20)	spoken (23)	German (19)	EXMARaLDA (24)
learner corpus (1)	academic (5)	written (3)	Spanish (9)	L2 data (10)
reference corpus (1)	public (2)	(1)	Portuguese (4)	L1 data (9)
treebank (1)			Swedish (4)	adult bilingualism (8)
			Turkish (4)	successive bilingualism (8)
			...	...

Hits: 27

EXMARaLDA Demo corpus

A selection of short audio and video recordings in various languages to be used for instruction or demo

Language: German, English, French, Spanish, Turkish, Polish, Vietnamese, Swedish, Norwegian, Italian, R

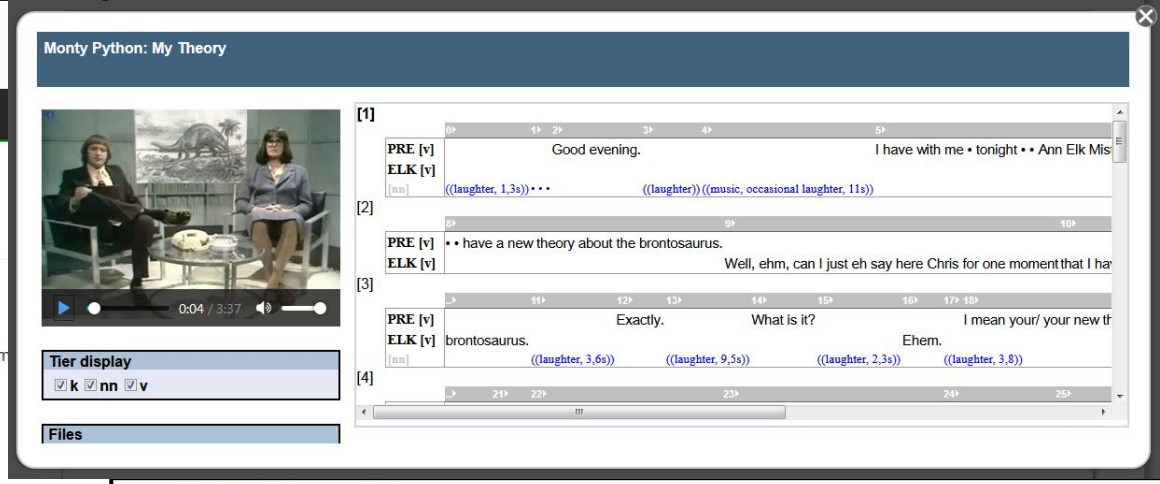
License: HZSK-PUB (public)

Hamburg Dependency Treebank

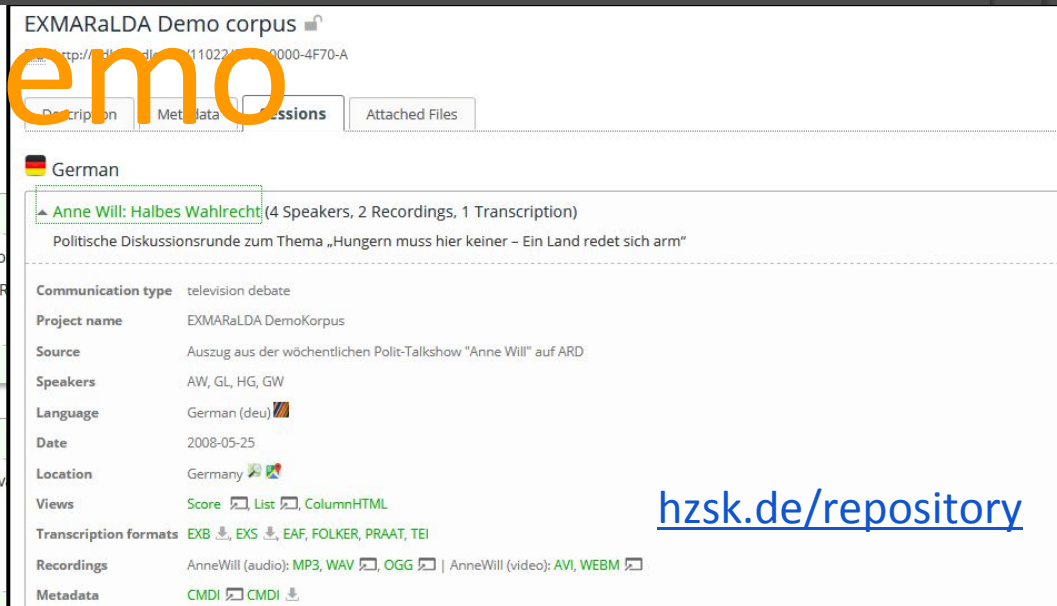
The Hamburg Dependency Treebank is to our knowledge the largest dependency treebank currently av
have not been transformed from phrase structures.

Language: German

License: HZSK-ACA (Text) / CC-by-sa-4.0 (Annotation) (academic)

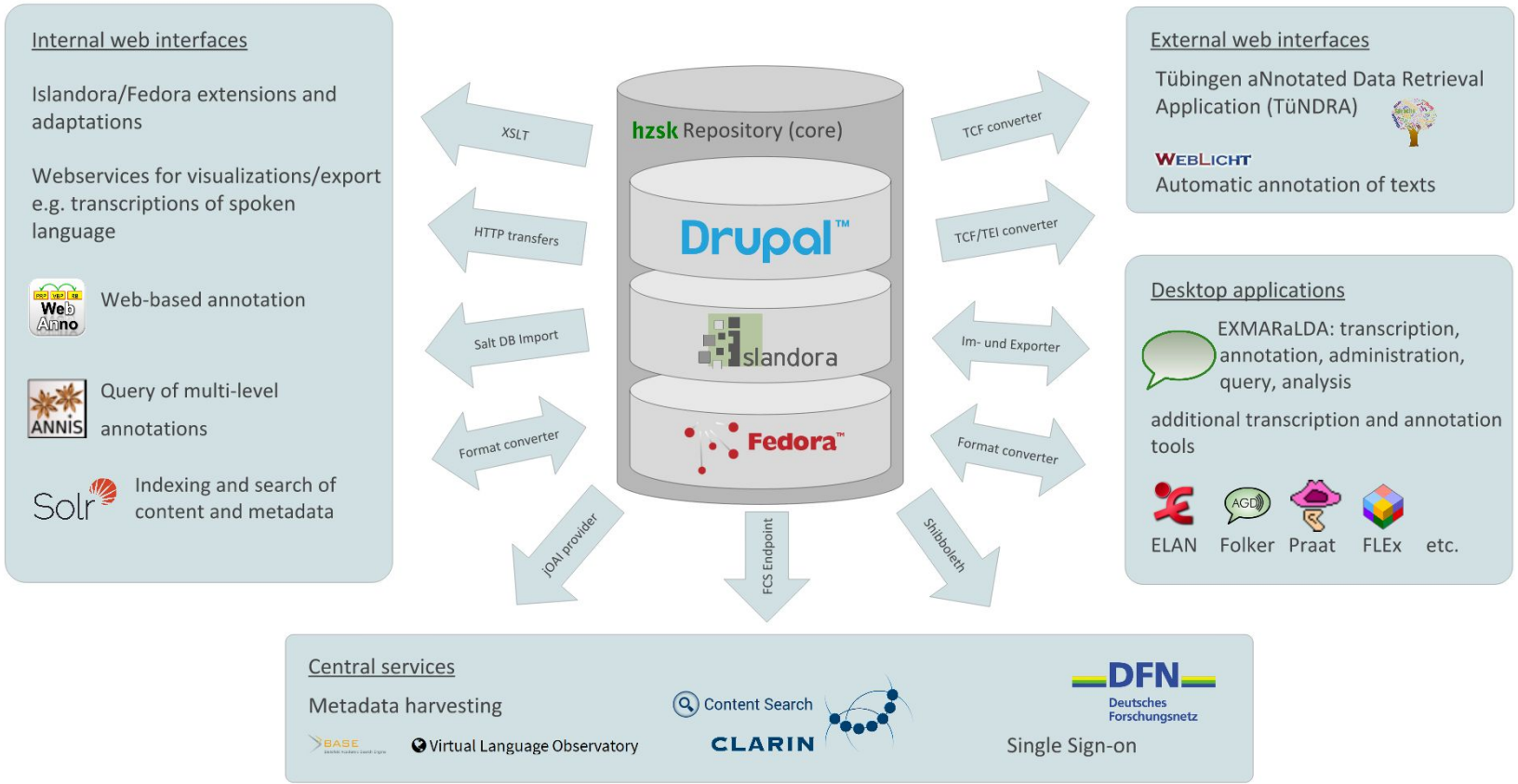


Demo



hzsk.de/repository

HZSK infrastructure



HZSK infrastructure

hzsk

Internal web interfaces

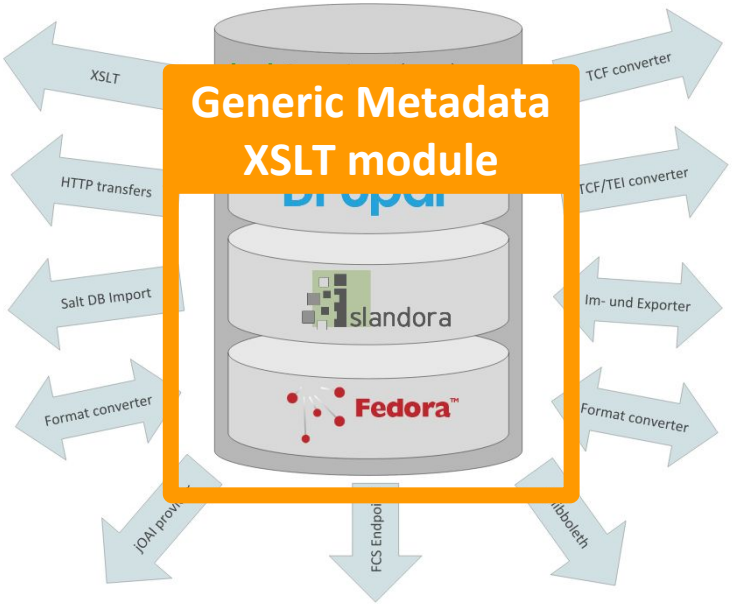
Islandora/Fedora extensions and adaptations

Webservices for visualizations/export e.g. transcriptions of spoken language

Web-based annotation

Query of multi-level annotations

Indexing and search of content and metadata



External web interfaces

Tübingen aNnotated Data Retrieval Application (TüNDRA)

WEBLICHT
Automatic annotation of texts

Desktop applications

EXMARaLDA: transcription, annotation, administration, query, analysis

additional transcription and annotation tools

ELAN Folker Praat FLEx etc.

Central services

Metadata harvesting

BASE

Virtual Language Observatory

Content Search

CLARIN

DFN Deutsches Forschungsnetz

Single Sign-on

Generic Metadata XSLT module

Customizable XSLT transformation

XML source created in module

```
// create XML input for XSLT transformation
$xml = '<results
  for="'.$fedoraIdentifier.'" roles="'.$drupalRoles.'"
  user="'.$drupalUserUID.'" lang="'.$drupalLanguage.'"
  dssid="'.$drupalUserSecureSessionID.'">'."\n";
foreach($associated_objects_array as $associated_object){
  $xml .= '
  <result>
    <title>
      '.$associated_object['dc_array']['dc:title']['value'].'
    </title>
    <description>
      '.$associated_object['dc_array']['dc:description']['value'].'
    </description>
    <identifier>
      '.$associated_object['dc_array']['dc:identifier']['value'].'
    </identifier>
  </result>';
}
$xml .= '</results>';
```

Generic Metadata XSLT module

Customizable XSLT transformation

Configuration for HZSK Repository - hzsk:config

Label	Type	Mime type	Size	Versions	Operations
general-functions-variables.xsl — XSLT containing general stuff (to be imported by other XSLT)	Managed	text/xml	37.09 KiB	17	replace download delete
config-params.xml — Configuration file for display of digital objects	Managed	text/xml	13.76 KiB	26	replace download delete
corpus-default.xsl — Default XSLT stylesheet for corpora	Managed	text/xml	27.34 KiB	54	replace download delete

EXMARaLDA Demo corpus - spoken-corpus:demo

ID	Label	Type	Mime type	Size	Versions	Operations
XSL	XSLT stylesheet for collection display	Inline XML	text/xml	219 B	139	replace download delete
CMDI	CMDI metadata for spoken corpus collection	Managed	text/xml	31.72 KiB	Not Versioned	replace download delete
HTML-DESC	Description of EXMARaLDA Demo Corpus (HTML)	Inline XML	text/html	1.25 KiB	16	replace download delete
HTML-VIEW	Content overview for EXMARaLDA Demo Corpus (HTML)	Inline XML	text/html	173.08 KiB	28	replace download delete

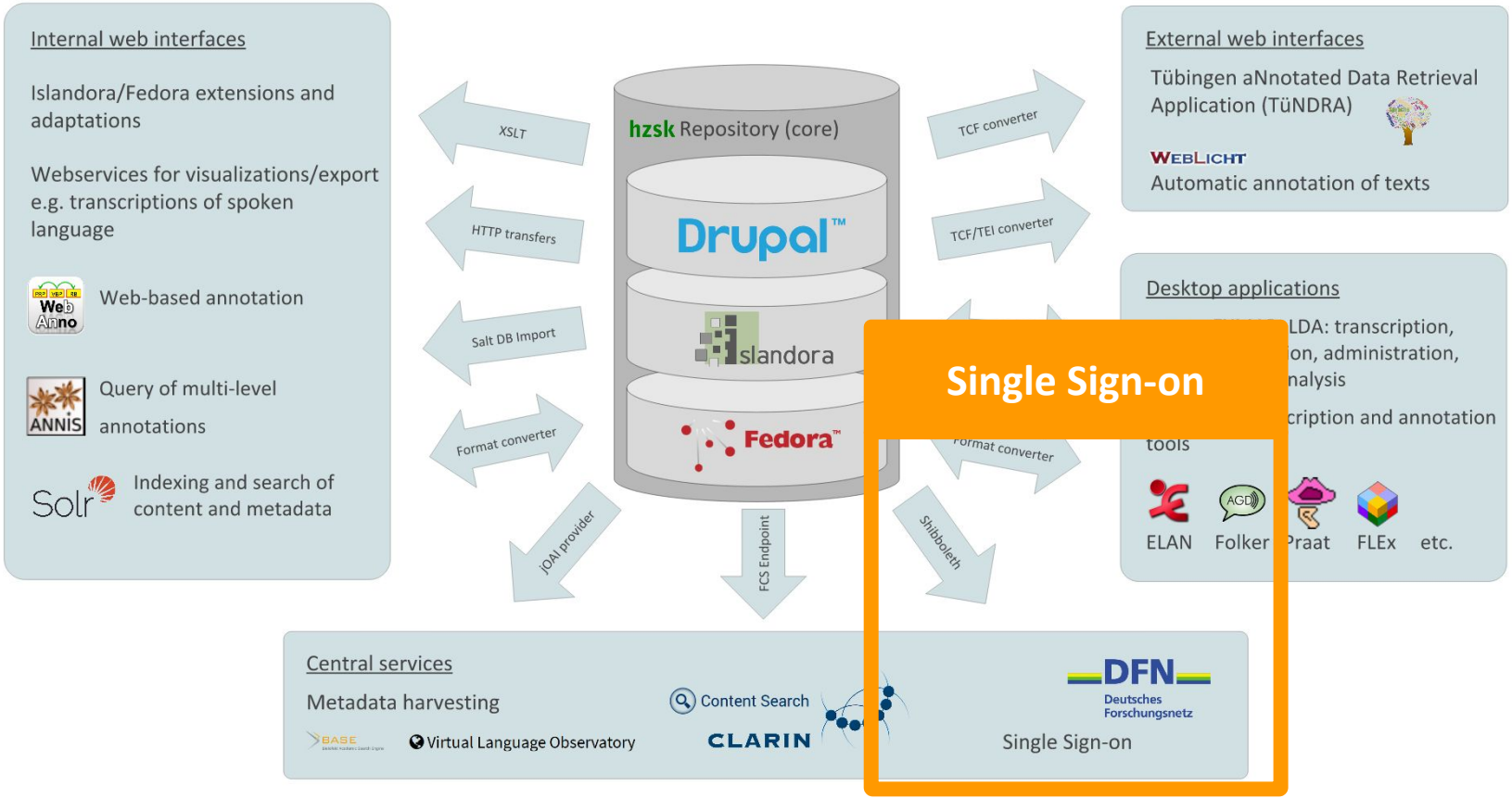
Some content pre-processed

Why implemented like this?

- similar collections but specific display requirements (spoken corpora, text corpora, treebanks, other types)
- allow data depositors to provide own styles (custom XSLT)

- influenced by Islandora's COLLECTION_VIEW datastream
- personal preference for XSLT, heavy use of modularization (overwriting defaults)

HZSK infrastructure: Single Sign-on



Authorization of users

Background

- Authentication of users outsourced to IdPs (CLARIN IdP as home of the homeless)
  → Drupal accounts still created in background by Shibboleth module
- Authorization (access to resources) determined by role assignment (and **XACML policies**)

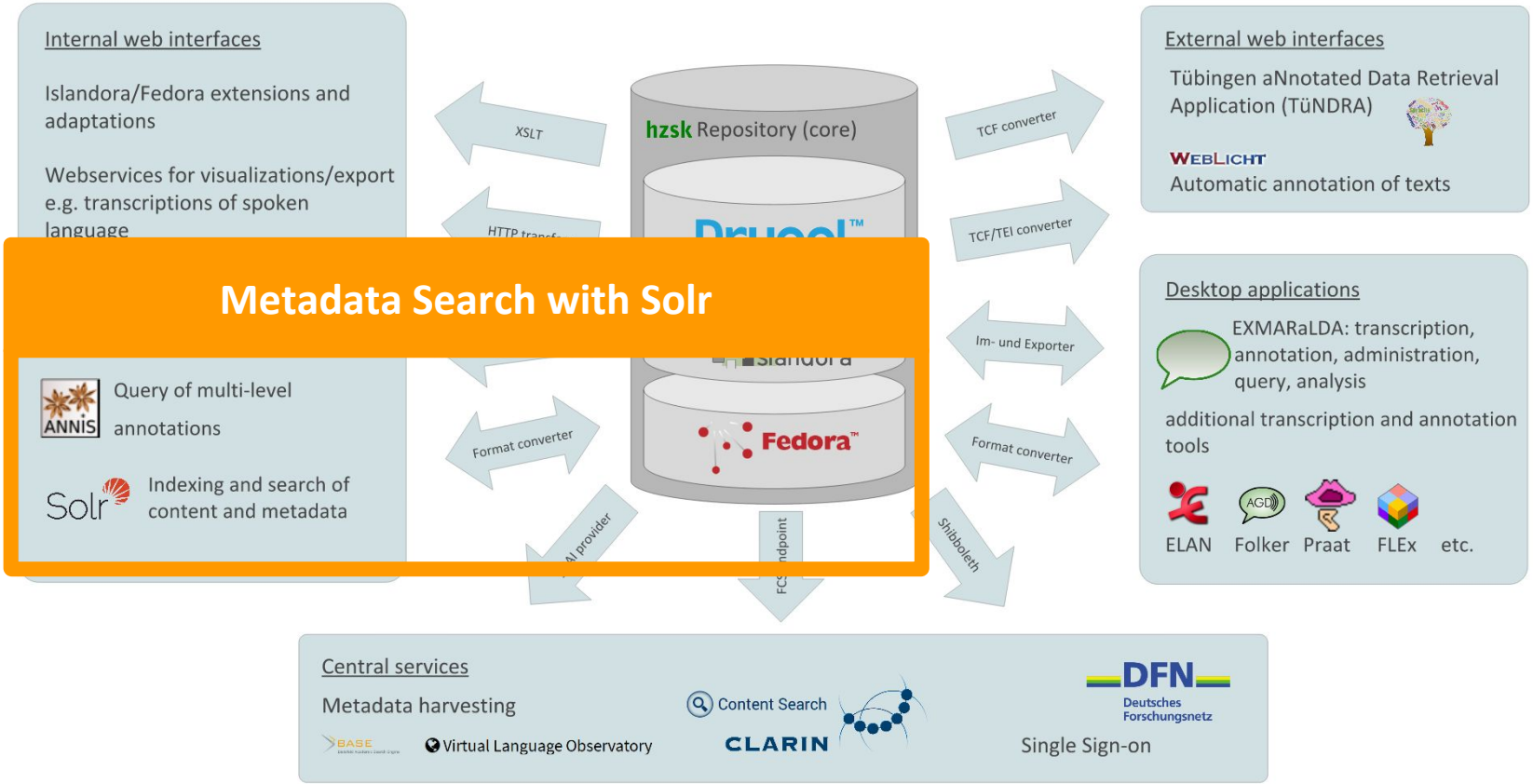
Authorization process

(a) Login via Shibboleth

(b) PUB & ACA resources available to users entitled “academic”
  → RES requires sending access form

(c) Corpus-specific role assigned to user after approval

HZSK infrastructure: Metadata Search with Solr



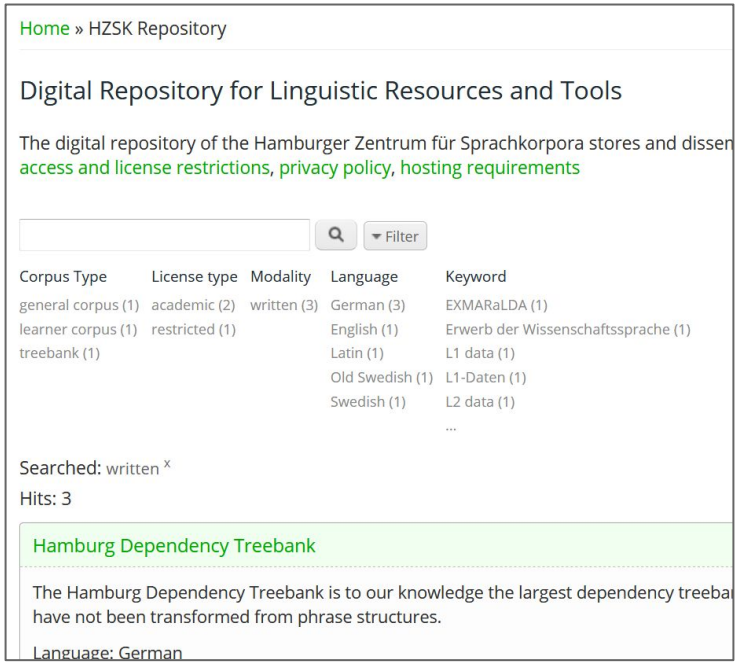
Metadata Search with Solr

Background

Facet search needed so that users can find relevant resources in repository

Implementation

- Java webservice for indexing metadata with Solr 5.0.0 (using RISearch to find datastreams, and SolrJ client)
- Drupal page (redirected to from islandora:root) reads from Solr and displays facet search with formatted results
- three-level sorting: (a) access, (b) display priority in config, (c) a-z



Fedora 4: **Linked Data**

Link metadata, catalogues and content (internally & externally)