Beyond Islandora
How to misuse Islandora and get away with it

Islandoracon 2017

Daniel Jettka, daniel.jettka@uni-hamburg.de
Hamburg Centre for Language Corpora, HZSK
University of Hamburg/Germany
Evaluate if and how Islandora is misused in the digital ecosystem of the Hamburg Centre for Language Corpora (HZSK)
Evaluate if and how Islandora is misused in the digital ecosystem of the Hamburg Centre for Language Corpora (HZSK)

Spoiler alert: seems like we are getting away with it
Demo
HZSK infrastructure: Islandora

**Internal web interfaces**
- Islandora/Fedora extensions and adaptations
- Webservices for visualizations/export e.g. transcriptions of spoken language
- Web-based annotation

**External web interfaces**
- Tübingen aNnotated Data Retrieval Application (TüNDRA)
- WEBLIGHT: Automatic annotation of texts

**Desktop applications**
- EXMARaLDA: transcription, annotation, administration, query, analysis
- additional transcription and annotation tools
  - ELAN
  - Folklor
  - Praat
  - FLEX etc.

**Central services**
- Metadata harvesting
- Content Search
- Single Sign-on
Customizable XSLT transformation

XML source created in module

```php
// create XML input for XSLT transformation
$xml = '<?xml version="1.0" encoding="UTF-8"?>
<results>
    <result>
        <title>
            '.$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>
</results>';
```
### Generic Metadata XSLT module

**Customizable XSLT transformation**

#### Configuration for HZSK Repository - hzsk:config

<table>
<thead>
<tr>
<th>Label</th>
<th>Type</th>
<th>Mime type</th>
<th>Size</th>
<th>Versions</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>general-functions-variables.xsl</td>
<td>Managed</td>
<td>text/xml</td>
<td>37.09 KiB</td>
<td>17</td>
<td>replace, download, delete</td>
</tr>
<tr>
<td>config-params.xml</td>
<td>Managed</td>
<td>text/xml</td>
<td>13.76 KiB</td>
<td>26</td>
<td>replace, download, delete</td>
</tr>
<tr>
<td>corpus-default.xsl</td>
<td>Managed</td>
<td>text/xml</td>
<td>27.34 KiB</td>
<td>54</td>
<td>replace, download, delete</td>
</tr>
</tbody>
</table>

#### EXMARaLDA Demo corpus - spoken-corpus:demo

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

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
- personal preference for XSLT, heavy use of modularization (overwriting defaults)
### Use or misuse?

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Type</th>
<th>Mime type</th>
<th>Size</th>
<th>Versions</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>general-functions-variables.xsl</code></td>
<td>XSLT containing general stuff (to be imported by other XSLT)</td>
<td>Managed</td>
<td>text/xml</td>
<td>37.09 KiB</td>
<td>17</td>
<td>replace, download, delete</td>
</tr>
<tr>
<td><code>config-params.xml</code></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>corpus-default.xsl</code></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For other entries:
- **XSL**: XSLT stylesheet for collection display
  - Inline XML
  - Type: Managed
  - Mime type: text/xml
  - Size: 219 B
  - Versions: 139
  - Operations: replace, download, delete

- **CMDI**: CMDI metadata for spoken corpus collection
  - Type: Managed
  - Mime type: text/xml
  - Size: 31.72 KiB
  - Versioned: No
  - Operations: replace, download, delete

- **HTML-DESC**: Description of EXMARaLDA Demo Corpus (HTML)
  - Type: Inline XML
  - Mime type: text/html
  - Size: 1.25 KiB
  - Versions: 16
  - Operations: replace, download, delete

- **HTML-VIEW**: Content overview for EXMARaLDA Demo Corpus (HTML)
  - Type: Inline XML
  - Mime type: text/html
  - Size: 173.08 KiB
  - Versions: 28
  - Operations: replace, download, delete

Some content pre-processed
HZSK infrastructure: Single Sign-on

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 aNNnotated Data Retrieval Application (TüNDRA)
- WebLIGHT
- Automatic annotation of texts

Desktop applications
- LDA: transcription, annotation, administration, analysis
- tools
  - ELAN
  - Folker
  - Praat
  - FLEX etc.

Central services
- Metadata harvesting
- Content Search
- Single Sign-on
Authentication of users

Background
HZSK stores data with different access levels (PUB, ACA, RES) → users have to be identified and in most cases have to apply for access to individual collections

Authentication methods

Effect
- No storage of passwords; if IdPs release enough user information we do not have to deal with it → less maintenance (communication cycles)
- Accounts do not expire, roles do however → Authorization
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
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: OAI provider

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

Central services
- Metadata harvesting

External web interfaces
- Tübingen aNNoted Data Retrieval Application (TüNDRA)
- W3BLIGHT - Automatic annotation of texts

Desktop applications
- EXMARaLDA: transcription, annotation, administration, query, analysis
- additional transcription and annotation tools
- ELAN, Folklore, Praat, FLEX etc.

OAI Provider

Content Search
- CLARIN
- DFN (Deutsches Forschungsnetzwerk)

Single Sign-on
Background
Metadata has to be harvested by catalogues (e.g. CLARIN Virtual Language Observatory)

Implementation
● jOAI (Java web application for OAI-PMH 2.0)
● Metadata fetched from datastreams by service (using RISearch/SPARQL) and stored redundantly in file system
● (Re-)Indexing managed manually in jOAI GUI
Why implemented like this?

- after Fedora update OAI Provider Service 1.2.2 was not compatible anymore
- fast workaround was needed
- jOAI: few dependencies (Tomcat app); unlike Islandora OAI: Solr, GSearch
Background
Metadata has to be harvested by catalogues (e.g. CLARIN Virtual Language Observatory)

Implementation
- jOAI (Java web application for OAI-PMH 2.0)
- Metadata fetched from datastreams (using RISearch/SPARQL) and stored redundantly in file system
- (Re-)Indexing managed manually in jOAI GUI
HZSK infrastructure: Metadata Search with Solr

Internal web interfaces
- Islandora/Fedora extensions and adaptations
- Webservices for visualizations/export e.g. transcriptions of spoken language

External web interfaces
- Tübingen aNNoted 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
- Content Search
- Single Sign-on

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
Why implemented like this?

- Struggled with GSearch, Islandora Solr Search module, Solr, and their dependencies for loooong time
- standalone Solr worked
Background
Facet search needed so that users can find relevant resources in repository

Implementation
- Java webservice: indexing metadata with Solr 5.0.0 (using RISearch to find datastreams)
- 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

Use or misuse?
Lack of time

Lack of money

Lack of knowledge
Evaluation: workforce
Evaluation

- Lack of time
- Lack of money
- Personal factors
- Lack of knowledge
## Evaluation: personal experience

<table>
<thead>
<tr>
<th>LDP</th>
<th>METS</th>
<th>FOXML</th>
<th>OAI</th>
<th>XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCR</td>
<td>UUID</td>
<td>MADS</td>
<td>JWT</td>
<td>PHP</td>
</tr>
<tr>
<td>XSLT</td>
<td>JSON-LD</td>
<td>OMG</td>
<td>XACML</td>
<td>LAME</td>
</tr>
<tr>
<td>RELS-EXT</td>
<td>RDF</td>
<td>FITS</td>
<td>API</td>
<td>MODS</td>
</tr>
<tr>
<td>PREMIS</td>
<td>PCDM</td>
<td>GUI</td>
<td>Fedora</td>
<td>TechMD</td>
</tr>
</tbody>
</table>
Evaluation

Lack of time

Lack of money

Suboptimal documentation

Module dependencies

Personal factors

Lack of knowledge
Evaluation

Lack of time
Lack of knowledge
Suboptimal documentation
Module dependencies
Lack of money
Personal factors

Module dependencies

Personal factors
tackle Islandora
tackle yourself
tackle institution

tackle yourself
tackle institution
tackle Islandora

tackle Islandora
tackle yourself
tackle institution
Questions, comments?