DwB Discovery Portal
A New CESSDA Portal for European Research Data Discovery

John Shepherdson - UKDA
Pascal Heus - Metadata Technology
Ørnulf Risnes - NSD
Overview

• Supports DwB goal “equal and easy access to official microdata for the European Research Area”

  ➢ provides “more coherent system for resource discovery of official statistics”

  ➢ demonstrates ability to ingest metadata from multiple sources, via multiple protocols
Scope of the portal work in WP12

- Content-wise
  - Metadata from NSIs + Archives

- Technical
  - Build prototype/beta
  - Sound, future proof methods, architecture, components
  - Standards-based
  - Extensible
  - Easy to hand over to ‘sustainability’ body
Functional aspects of the portal

• Research data discovery (obviously)

• Provider portal, QA

• Platform for additional services
Metadata ingest

- Metadata gets...
  - harvested
  - made ready for QA
  - transformed into canonical model
  - indexed
  - exposed
Canonical metadata model

• Harmonisation
  ➢ DDI-C, DDI-L, MISSY, CIMES, etc.

• Builds on DISCO
  ➢ DDI discovery RDF
## Metadata ingest dependencies

<table>
<thead>
<tr>
<th>Step</th>
<th>Source/standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Harvesting</td>
<td>specific</td>
</tr>
<tr>
<td>2) Produce harvesting report</td>
<td>specific</td>
</tr>
<tr>
<td>3) Conversion to Raw-RDF</td>
<td>agnostic</td>
</tr>
<tr>
<td>4) Produce conversion report</td>
<td>agnostic</td>
</tr>
<tr>
<td>5) Harmonization</td>
<td>agnostic</td>
</tr>
<tr>
<td>6) Produce harmonization report</td>
<td>agnostic</td>
</tr>
<tr>
<td>7) Loading</td>
<td>agnostic</td>
</tr>
<tr>
<td>8) Produce loading report</td>
<td>agnostic</td>
</tr>
<tr>
<td>9) Indexing</td>
<td>agnostic</td>
</tr>
<tr>
<td>10) Produce indexing report</td>
<td>agnostic</td>
</tr>
<tr>
<td>11+) Discovery, other downstream processes</td>
<td>agnostic</td>
</tr>
</tbody>
</table>
Platform for services

• DwB search portal is just a front-end application

• Machine-actionable interfaces for most functions (REST)
Search Portal (alpha)

- Powered by Solr
- Facets
  - producer, geography, date, data type
  - ...

Filter By

- Software
- Producer
- AuthEnty
- FundAg
- Distibtr
- TopClas
- GeogCover
- GeogUnit
- Universe
- DataKind
- TimeMeth
Search Portal (alpha)

- Suggestions / autocomplete
Search Portal (alpha)

- ‘Did you mean?’ functionality

```
houshold
GO
```

Did you mean?

`household`
Sprint in Colchester, May 2014

- Use Jenkins CI tool to:
  - Harvest Nesstar metadata
    - any public instance
  - Load DDI XML in to BaseX
  - Convert DDI XML to raw DwB-RDF
  - Harmonize DwB-RDF
    - Simple
Sprint in Colchester, May 2014

- Integrated Jenkins with Git to
  - Build Nesstarvester and BasexSync tools automatically
  - Update harmonization scripts automatically
- Identified mechanism to detect metadata language
  - So can check language tag is correct
- Produced Solr schema
# Jenkins Dashboard

<table>
<thead>
<tr>
<th>S</th>
<th>W</th>
<th>Name</th>
<th>Last Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>☀️</td>
<td>☀️</td>
<td><strong>Build BaseXSs</strong>ync via Maven</td>
<td>21 days - #1</td>
</tr>
<tr>
<td>🔴</td>
<td>⛅️</td>
<td><strong>Build Nesstarvester via Maven</strong></td>
<td>21 days - #28</td>
</tr>
<tr>
<td>☀️</td>
<td>☀️</td>
<td><strong>Config files sync</strong></td>
<td>13 days - #3</td>
</tr>
<tr>
<td>☀️</td>
<td>☀️</td>
<td><strong>RunCaelum</strong></td>
<td>15 days - #20</td>
</tr>
<tr>
<td>☀️</td>
<td>☀️</td>
<td><strong>RunDatabaseSync</strong></td>
<td>15 days - #32</td>
</tr>
<tr>
<td>☀️</td>
<td>☀️</td>
<td><strong>RunHarvester</strong></td>
<td>15 days - #101</td>
</tr>
<tr>
<td>⛅️</td>
<td>⚡️</td>
<td><strong>RunNesstarvesterBatch</strong></td>
<td>N/A</td>
</tr>
</tbody>
</table>
Jenkins Job Details

Project RunHarvester

Run the harvester against a data source using certain harvester type (both specific)

Workspace

Recent Changes

Downstream Projects

RunDatabaseSync

Permalinks

- Last build (#101), 15 days ago
- Last stable build (#101), 15 days ago
- Last successful build (#101), 15 days ago
Jenkins Job Details

Build #100 (20-May-2014 12:11)

No changes.

Started by user John Shepherdson

Revision: 898c6d024512e6e6a04733db22d48f44774cf4f7
  origin/master
Metadata harmonization

- **Standard level** - sources based on the various metadata standards
- **Version level** - within a standard, the use of different versions (e.g. DDI 1.2.2, 2.5, 3.x)
- **Template/flavour level** - the use of elements of the standard for different purposes; presence/absence of optional elements
  - driven by institutional practices, templates, or software tooling
Typical Console Output (Captured)

```
set-harvester-profile:
  [echo] set up path to harvester properties file

update-harvester-profile:

fail-if-file-not-found:
  [echo] Checking file exists /var/db/config/source/uk.ekda.teaching.nv.properties
  [echo] Updating existing Nesstarvester property file path details
  [propertyfile] Updating property file: /var/db/config/source/uk.ekda.teaching.nv.properties

run-harvester:
  [echo] Running Nesstarvester process

run-jar-with-props:
  [echo] run-jar-with-props called with Jar file = ../../lib/nesstarvester.jar
  [echo] properties file = /var/db/config/source/uk.ekda.teaching.nv.properties
  [echo] and class name = org.openmetadata.harvester.Harvester

fail-if-file-not-found:
  [echo] Checking file exists ../../lib/nesstarvester.jar

fail-if-file-not-found:
  [echo] Checking file exists /var/db/config/source/uk.ekda.teaching.nv.properties

[java]
[java]
[Java]
[Java]
[Java]
Output folder is /var/db/provider/uk.ekda/teaching/ddi
[Java] Starting DI harvest on server: ukda_research
[Java]
[Java] Setting up server configuration for : http://nesstar.ukdataservice.ac.uk:80
[Java] Server configured.
[Java]
[Java] Incremental Update in Progress...
[Java]
[Java] Incremental Update Completed.
[Java]
[Java] Harvest of ukda_research completed.
[Java]
[Java] All harvest operations have been completed.
[Java]
[Java] Thank you for using the OpenMetadata Nesstarvester.
[Java]

log-activity:
BUILD SUCCESSFUL
Total time: 12 seconds
Email was triggered for: Success
Sending email to: email.rinsenhd.ub.no@yeshopessen.ac.uk pascel.ekus@metadatatecnology.com
Warning: you have no plugin providing access control for builds, so falling back to legacy behavior
Trigering a new build of RunDatabaseSync #31
Finished: SUCCESS
```
What next?

- Perform provider/format specific transformations
- Apply DwB specific adjustments (identifiers, system metadata, etc.)
- Apply DwB harmonizers (map metadata in to DwB standard facets/CV etc.)
- Load harmonized DwB-RDF in to Virtuoso RDF database
- Index DwB-RDF with Solr
What next?

• Producing various ingestion / QA reports
• Propagate deletes for survey that have been dropped
• Synchronize various metadata files to repository
  ➢ For ‘before and after’ comparisons/provider feedback
Any Questions?