Component and Milestone Report

Project Acronym: OpenUp!

Grant Agreement No: 270890

Project Title: Opening up the Natural History Heritage for Europeana

---

C2.2.1 and MS01: Harvesting and Transformation component prototype (contributes to deliverable 02)

Revision: 2a (Final)

Authors:

Walter Koch (AIT)

Editor:

Jutta Stockklauser (AIT)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
01	13.5.2011	Walter Koch	AIT	Described components installed with the help of: Odo Benda (AIT) Gerda Koch (AIT) Bernd Sproger (AIT)
02	27.5.2011	Jutta Stockklauser	AIT	Final editing
02a	31.5.2011	Project Coordination	BGBM	Minor final edits.

Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Distribution

Recipient	Date	Version	Accepted YES/NO
TMG	Continuous	Website	YES
Work Package Leader	27.5.2011	V2	YES
Project Coordinator	27.5.2011	V2(a)	YES
Publication on Website	31.5.2011	V2a (PDF)	n/a

Table of Contents

Overview

Implementing and setting up the component. The OpenUp! harvesting component will be set up to harvest data on request from the data repositories of the content providers and ingest the data to the OpenUp! OAI provider. The data will be transformed according to the requirements of the various harvesting routines of the target repositories (EUROPEANA, BHL etc.).

Chosen Approach

The work carried out to deliver the "Europeana - Natural History Aggregator" follows an "Evolutionary Development Approach"[1]. This fits to the type of the OpenUp! project which - being a "Best Practice Project" - can be considered not as an IT Development Project but as an "Integration Project" based on available and tested components.

Initial project activities related to this work can be divided into three groups:

- Process Analysis
- Data Analysis
- Analysis of the technical framework (Nodes and Components).

Until Project Milestone (MS01: Harvesting and Transformation component prototype) a first process & data analysis has been executed and candidates (existing components) for the system implementation have been identified. A demonstration system has been set up on a test server (see below) which contains a prototype implementation of all components. First ABCD to ESE transformations have been done and ESE compliant test records have been loaded onto an OAI data provider. ESE records can be viewed using an existing aggregation platform from other EU funded projects.

1 http://en.wikipedia.org/wiki/Software_prototyping - Evolutionary_systems_development

Process Analysis

Based on the existing (Spring 2011) situation of the GBIF[2] and BioCASE[3] initiatives there have been identified roles dedicated to human or technical actors:

1. Human actors:
   For simplification reasons there have been identified two roles: **Data Provider** and **OpenUp! Manager.**
2. Helper Applications (non-Human Actors) :
   (IT based) tools and activities helping to execute necessary tasks have been identified in course of the analysis of the technical framework: 1) **Harvesting tools** (GBIF-HIT), 2) **Transformation tools** (Pentaho Kettle), 3) **Data Management tools** (SQL-Server), and 4) **Europeana data provision** (OAI-Data Provider).
3. Control System for the management of the required processes.

Using BPMN 2.0[4] notation and the Intalio Designer[5] the situation has been modelled and documented:

2 http://www.gbif.org/

3 http://www.biocase.org/

4 http://www.bpmn.org/

5 http://www.intalio.com/bpms

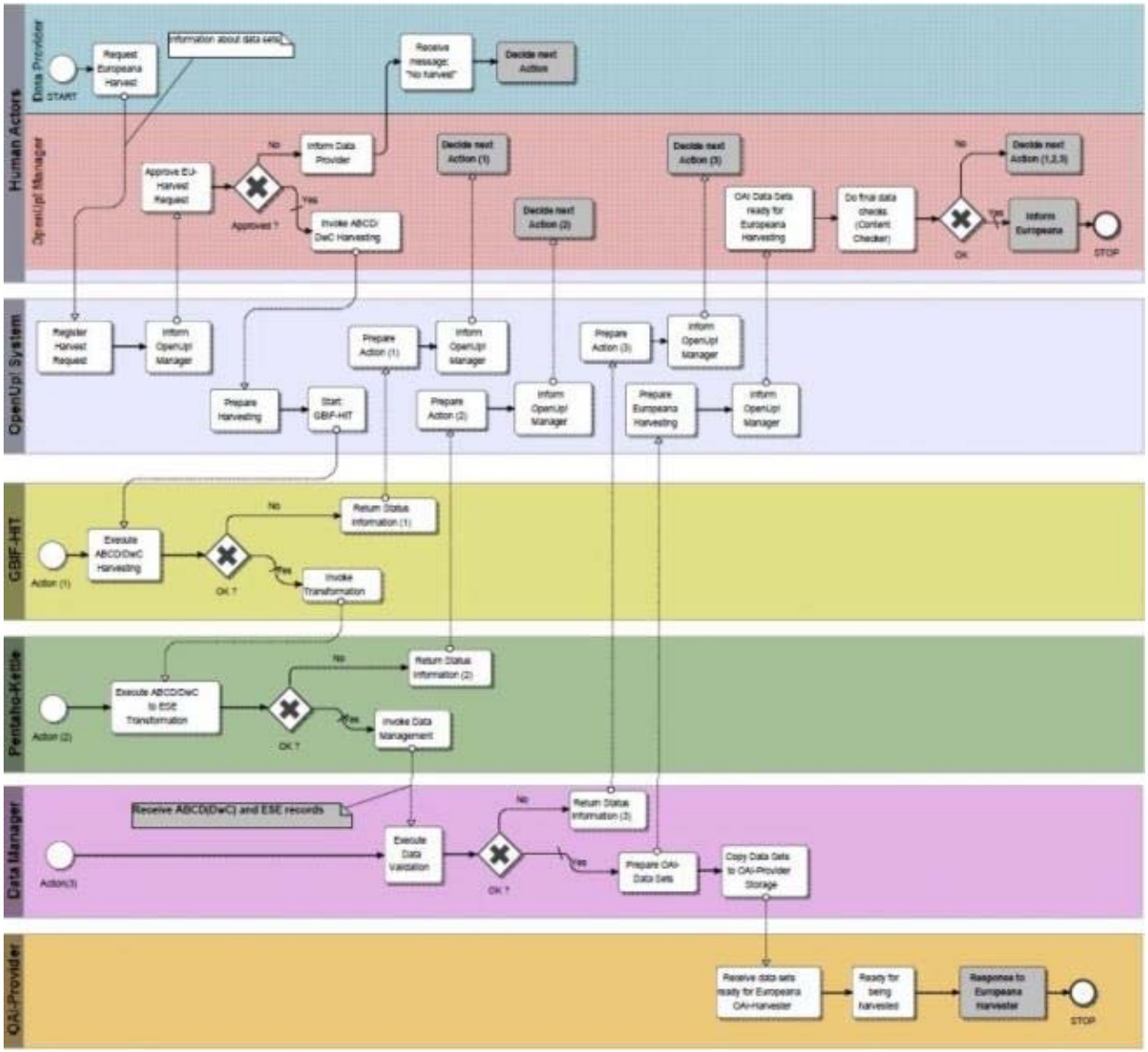


Figure 1: OpenUp! Process Model (high level)

Click here to see an enlarged PDF: BPMN-Diagram[6].

The model provides an end to end (Data Provider -> Europeana) workflow which could be made executable using a native BPMN or BPEL[7] engine. The challenge consists in the preparation of data sets to be harvested by the Europeana Harvester.

6 http://open-up.eu/sites/open-up.eu/files/u22/Intalio_bpm_01b.pdf

7 http://en.wikipedia.org/wiki/Business_Process_Execution_Language

Nodes and Components (Analysis of the technical framework)

Figure 2 in the Task Briefs Overview[8] includes a first iteration of a UML deployment diagram (par. 3.7) showing the nodes and components to be implemented when setting up the "Europeana - Natural History Aggregator". A further refinement was elaborated after the installation and test of the GBIF-HIT aggregator at AIT's test and staging servers (http://test111.ait.co.at:8080/hit/[9] username: admin; password: ait111):

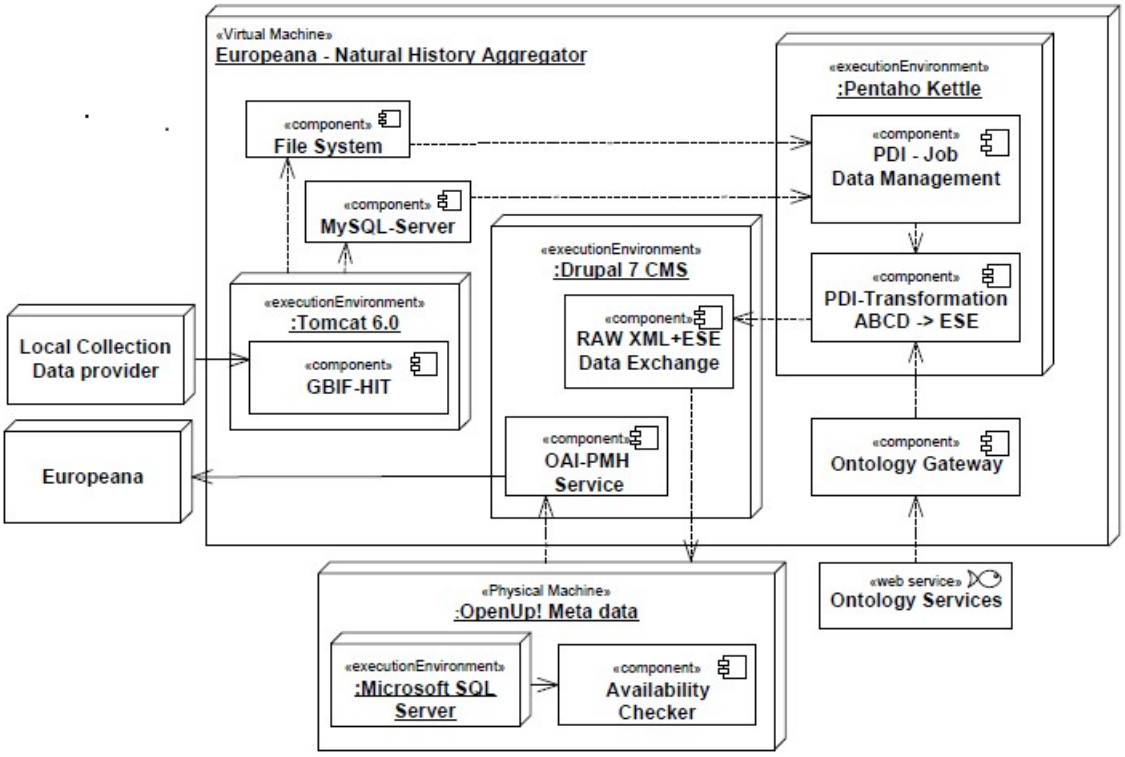


Figure 2: Deployment of the "Natural History Aggregators" components

The test server (test111) is set up as VMWare Image including a Ubuntu 10.04 Linux System.

Data Analysis

Data (e.g. ABCD records) generated by the GBIF-HIT[10] component are stored in the filesystem (/opt/hit) and in different tables of a connected MySQL data base.

An internal Search & Retrieval System (ZEBRA[11]) has been set up to analyse original ABCD-records coming from the contributing partners. This internally used system is using the Z39.50

8 http://open-up.eu/sites/open-up.eu/files/OpenUp_TN_TaskBriefs_Overview-v02.pdf

9 http://test111.ait.co.at:8080/hit/

10 http://code.google.com/p/gbif-indexingtoolkit/

11 http://www.indexdata.com/zebra

protocol.[12] To make use of existing developments a mapping from ABCD Elements to the Bib-1 Attribute[13] set of a Z39.50 application profile has been done. The mapping from "response"s (received from the HIT Harvester) to the Bib-1 Use Attributes[14] can be found here[15].
Please note that the mapping was only done to have some access point for the data analysis work. The internal system which allows execution of RPN Queries[16] can be accessed here[17]. An RPN Query formulating a search for a geographic term and a date might look like: "@and @attr 1=32 1862 @attr 1=58 sachsen"[18]. Note that queries using boolean logic are evaluated against the whole ABCD record and not within a single unit. Index look up (Scan) is also possible (click here)[19]. A look up (scan) for a geographic term could look like: "@attr 1=58 sachsen"[20] (only one token possible).

Besides these tools other available tools provided by BioCASe and GBIF are used for data analysis (original ABCD/DwC records) reasons: Simple access to ABCD providers[21] or GBIF Search (occurrences)[22] . The last one shows records on a "per unit basis" which can be used as starting point for the definition of an ESE Record (see: Occurrence Details)[23].

There seem to be two alternatives for the ESE Record creation: 1) to put all relevant data (e.g. collection/metadata into the "unit/ESE" record, or 2) produce unit/ESE records which are interlinked to a single and separate "collection"(metadata) record. The first approach seems to be in line with the GBIF Occurrence presentations, the latter case was the idea behind for the transformation process as described below.

To have a reasonable amount of test data available for designing and implementing the data migration (mapping process) some collections from the German provider BGBM have been harvested (Overview):

12 http://en.wikipedia.org/wiki/Z39.50

13 http://www.loc.gov/z3950/agency/defns/bib1.html

14 http://open-up.eu/sites/open-up.eu/files/u22/bib1.att_.txt

15 http://open-up.eu/sites/open-up.eu/files/u22/biocase_response.abs_.txt

16 http://www.indexdata.com/zebra/doc/querymodel-rpn.html

17 http://test111.ait.co.at/zebra/

18 http://test111.ait.co.at/zebra/?host[]=127.0.0.1%3A9999%2FDefault&query=%40and+%40attr+1%3D32+1862++%40attr+1%3D58+sachsen&action=Search&display=raw

19 http://test111.ait.co.at/zebra/scan.php

20 http://test111.ait.co.at/zebra/scan.php?host[]=127.0.0.1%3A9999%2FDefault&query=%40attr+1%3D58+sachsen&action=Search

21 http://search.biocase.org/simple-ui/

22 http://data.gbif.org/occurrences/

23 http://data.gbif.org/occurrences/273130/

harvested_count	name aufsteigend	provider_name	country
1.426	03666bc0 - Bridel Herbar	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
1.036	0bc58c10 - AlgaTerra	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
3.677	349ffa50 - Christopher Walker collection of Glomer...	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
196	3d03cfcd - Naturhistorisches Museum Mainz, Botanic...	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
6.191	3d03cfcd - Naturhistorisches Museum Mainz, Zoologi...	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
6	4d310580 - Desmidiaceae Rostock	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
125.444	57c36090 - Lichen Herbarium Berlin	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
10.316	5aac5e90 - Desmidiaceae Engels	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
46.311	63f549f0 - Herbarium Willing	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
31.384	67cb61c0 - Staatliches Museum für Naturkunde Stutt...	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
1.711	95e58910 - BoGART	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
2.171	9a2061a0 - Flora exsiccata Bavarica	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
9	a1521e10 - BoBO - Botanic Garden Berlin BDI Observ...	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
2.866	a37305c0 - Pollichia Pilzherbar	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
121.559	a42fc9a0 - Herbarium Berolinense	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
1.359	aa037af0 - GEO Biodiversity Day	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
7.587	c49f9b80 - Algaterra Types	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
46.298	c666fbc0 - Herbarium Senckenbergianum Görlitz (GLM...	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
23	d141a3a0 - AlgaTerraMovies	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
448	e0393fe0 - Harpactus Shuckard	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
616	e3826de0 - Desmidiaceae Mollenhauer	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
1.490	e6416f50 - PonTaurus collection	Botanic Garden and Botanical Museum Berlin-Dahlem	Germany
412.124			

Figure 3: Harvested Data Sets (statistics)

To support data analysis a sample data base (pontaurus: ER-diagram) coming along with the BioCASe Provider Software[24] was quite helpful. The provider software can be accessed at: http://test111.ait.co.at/biocase/[25] (test data source: ait111-pontaurus). In course of the project it is foreseen to set up "reference data source" which include valid datasets for the ESE transformation of different "flavours" of ABCD and DwC records.

24 http://www.biocase.org/products/provider_software/

25 http://test111.ait.co.at/biocase/

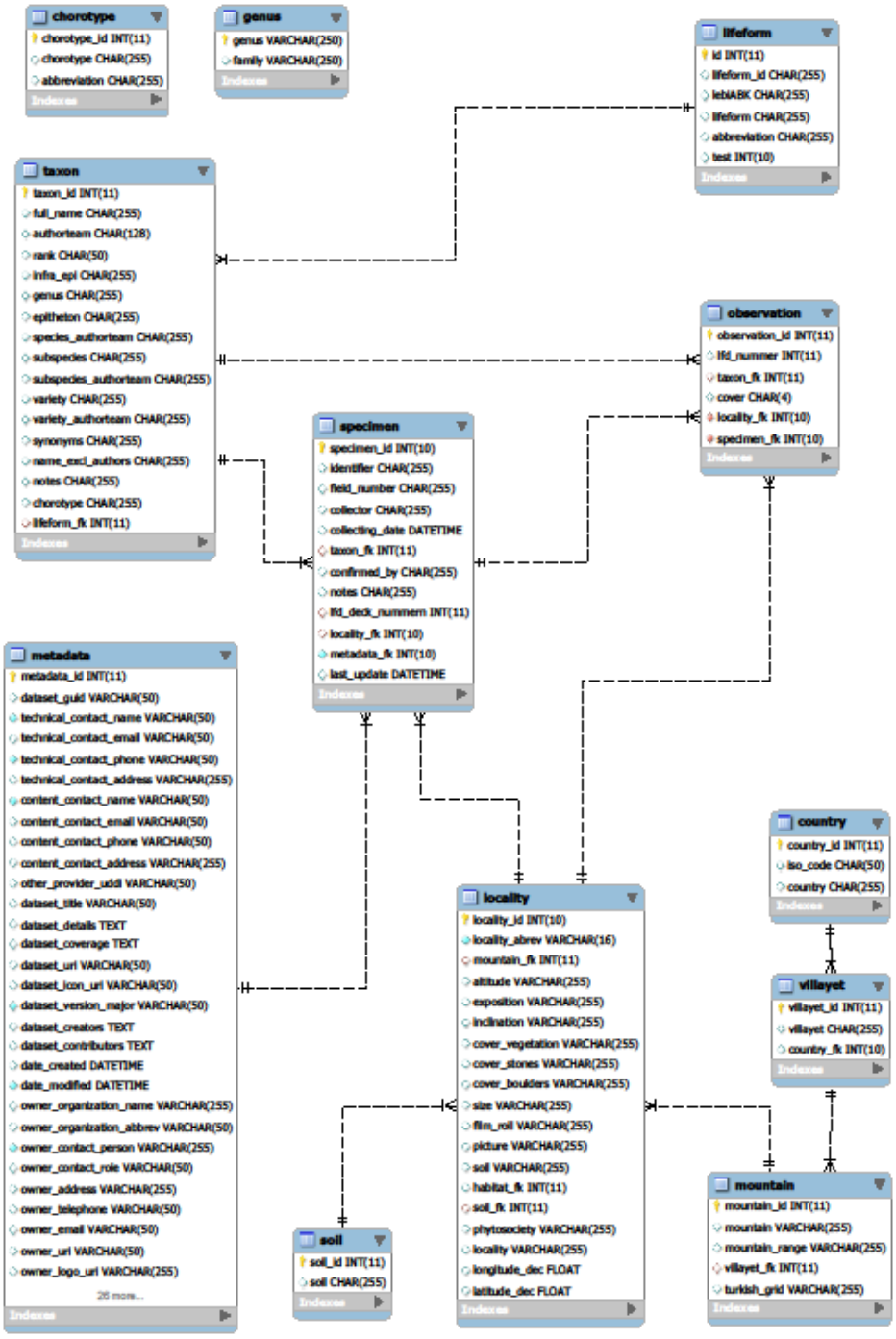


Figure 4: ER Diagram of a BioCASe test data set

For preparing (PDI-Job) ABCD-XML records for transformation (PDI-Transformation) and input into the RAW-XML component the Pentaho Data Integration[26] environment is used. In a first step to

26 http://www.pentaho.com/products/data_integration/

demonstrate the usefulness and applicability of this BI (Business Intelligence) tool the ABCD records have been split up into records each of them presenting the information contained in the "abcd:Units" element. These "unit-records" will be interlinked with a record describing the "collection/dataset" (abcd:Metadata) to which the units belong.

The migration process is split up into "jobs" (Wiley-Pentaho_Solutions_Figure_9-1) invoking "transformations" (Wiley-Pentaho_Solutions_Figure_9-2).

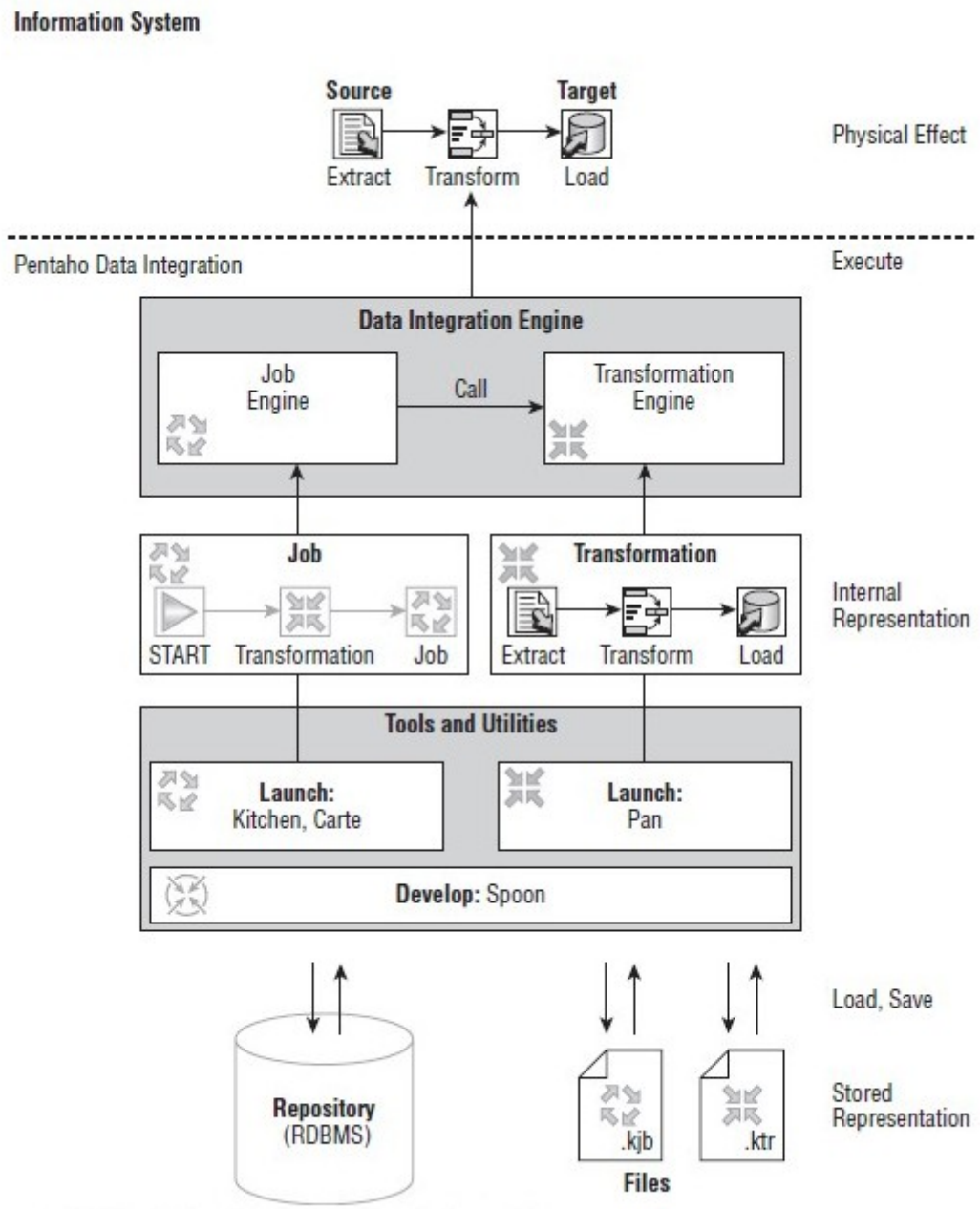


Figure 9-1: Pentaho data integration tools and components

Figure 5: Pentaho Data Integration (PDI) tools and components

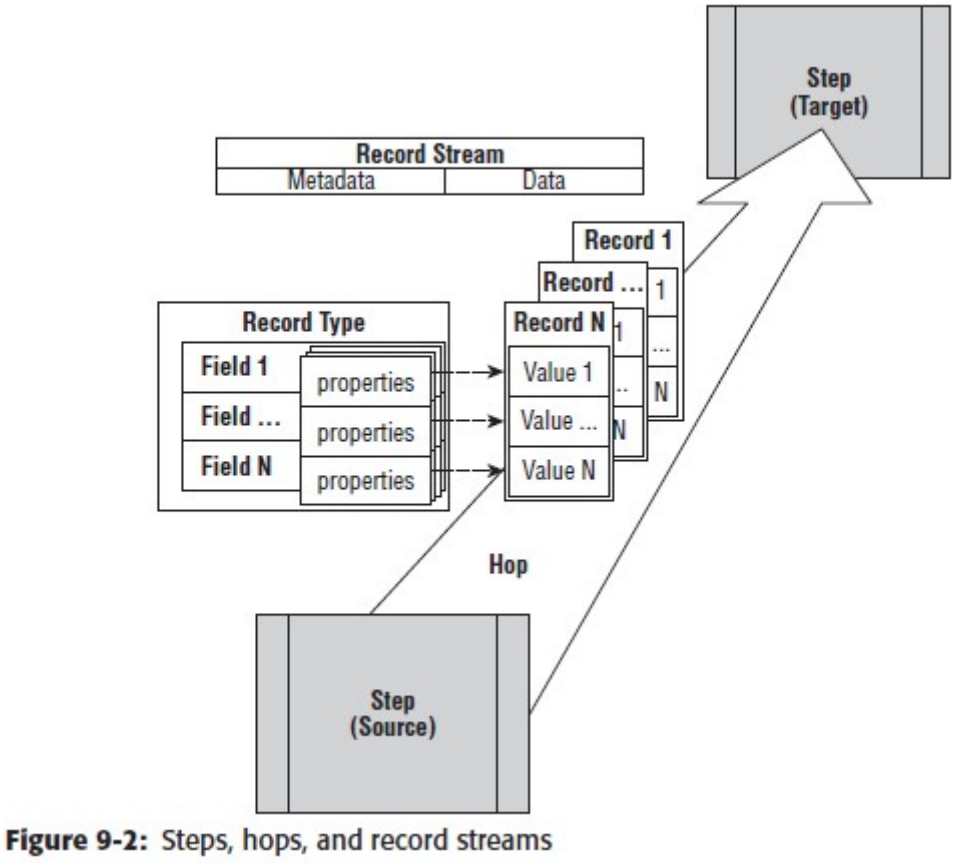


Figure 6: PDI - Steps, hops, and record streams

The figures are taken out from Pentaho® Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL®[27]. Examples how the ABCD -> ESE Migration Process looks like can be seen from the stubs below:

27 http://open-up.eu/sites/open-up.eu/files/u22/Wiley-Pentaho_Solutions_Cover.jpg

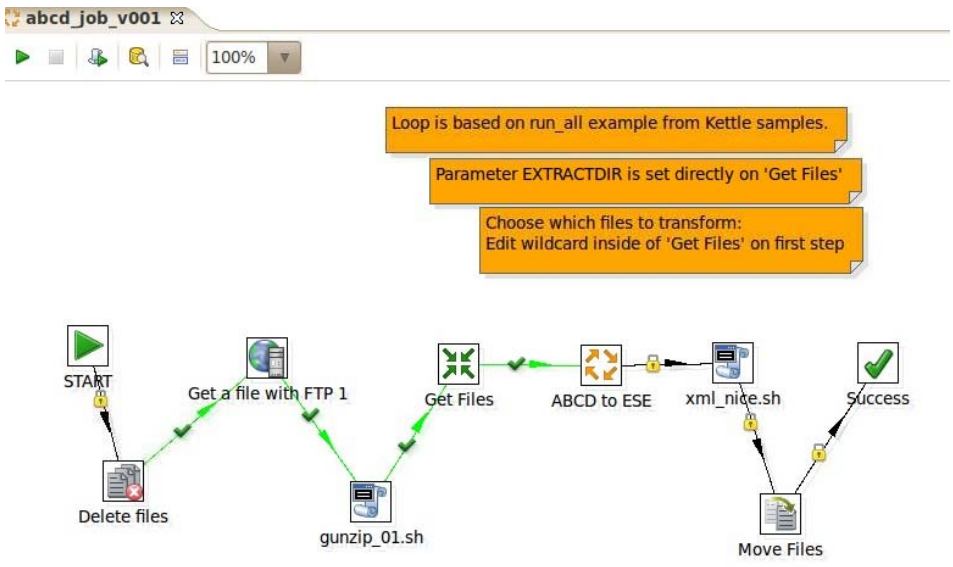


Figure 7: ABCD to ESE migration – PDI – Job (example)

Transformation of selected metadata elements:

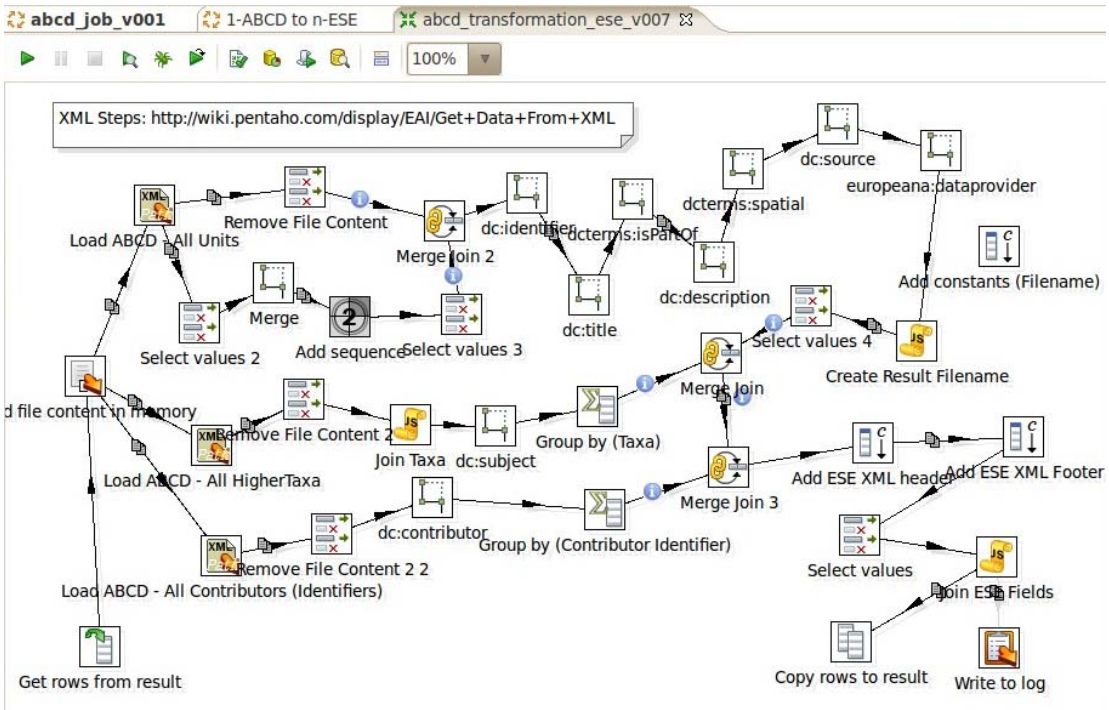


Figure 8: ABCD to ESE migration – PDI – Transformation (example)

The transformed data (ESE-Record candidates) are loaded into an aggregation platform already in use within different projects (e.g.: BHLE: German "throw away" prototype) and can - internally - be searched and inspected:

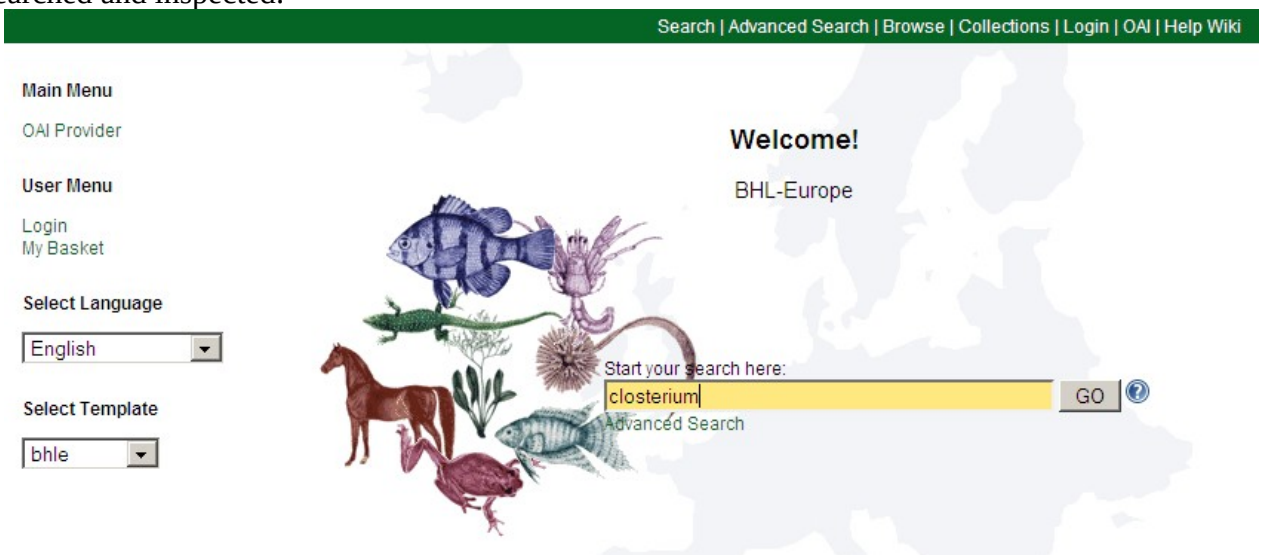


Figure 9: Internal Search & Retrieval system used for inspection of ESE records

A simple query for "closterium"[28] would deliver following result list:

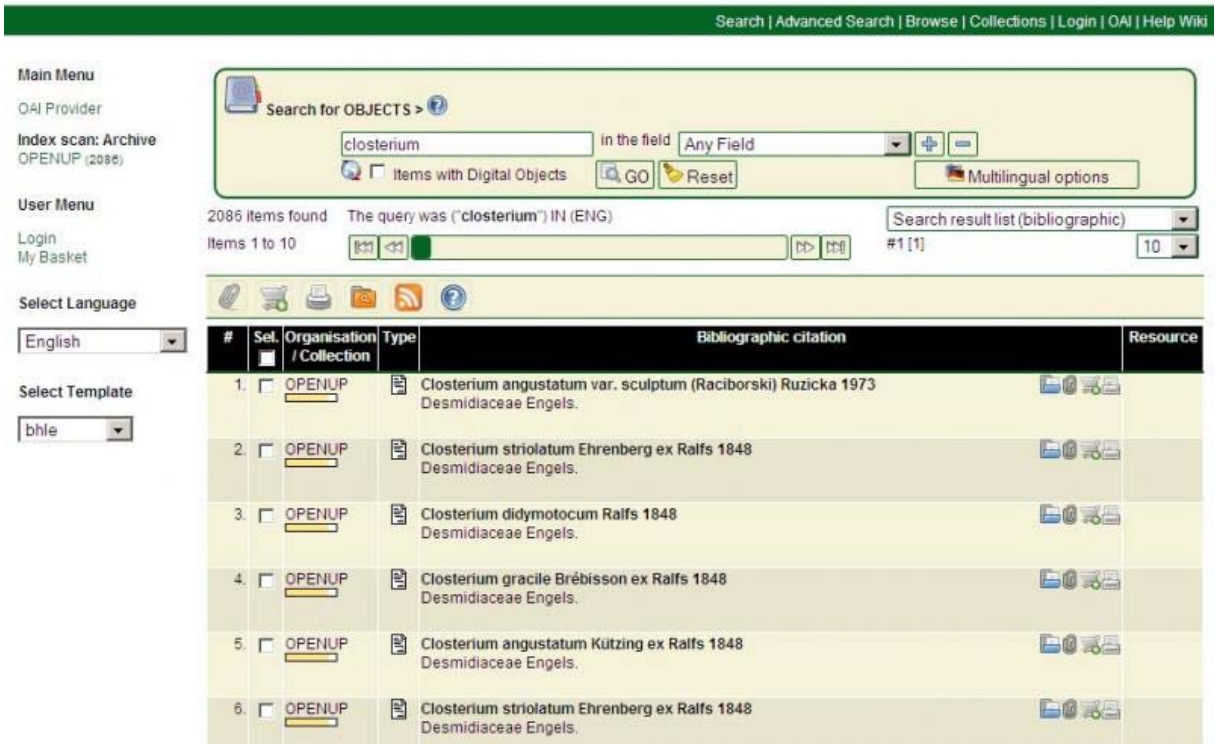


Figure 10: Presentation of Search results

28 http://en.wikipedia.org/wiki/Closterium

The detailed view shows selected abcd-metadata which have been used for this "proof of concept" data migration (please note that the field "Description" describes the "Collection/Data Set" ("is Part Of"):

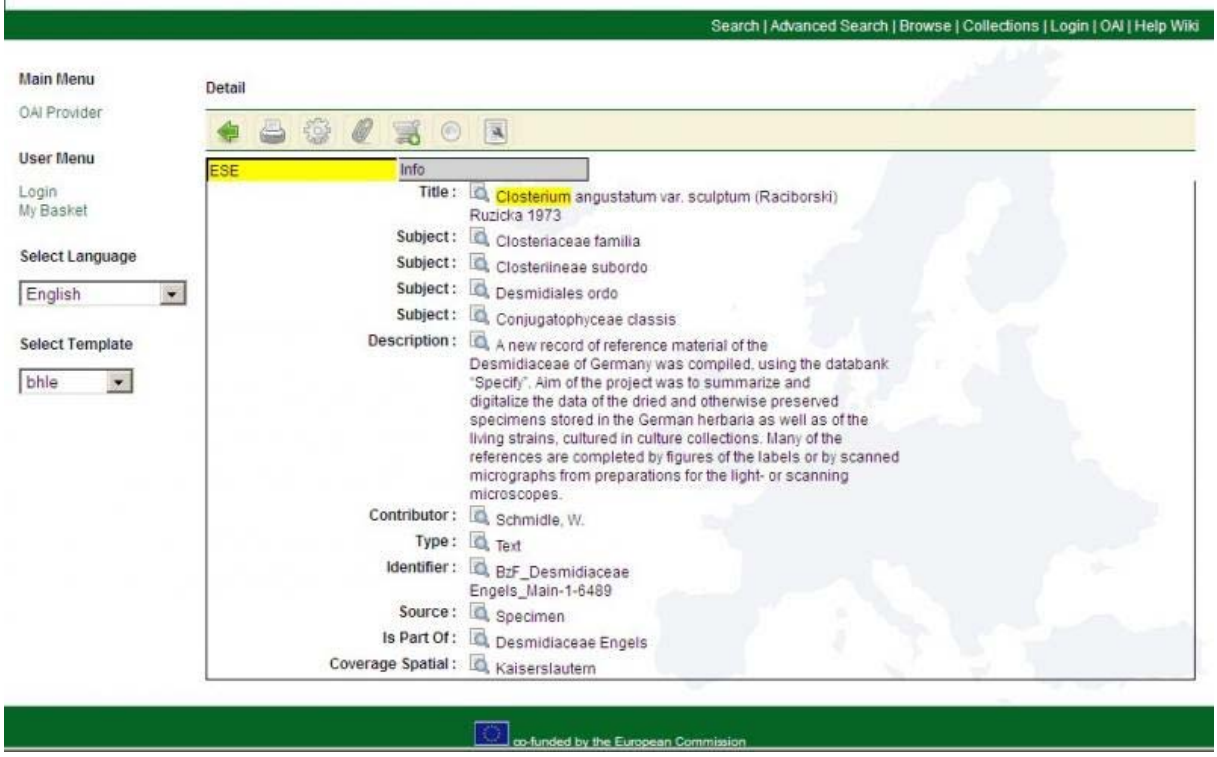


Figure 11: Presentation of metadata elements (detailed view)

The XML-presentation of this records looks like:

```
<section name="raw">
    <europeana:record xmlns:europeana="http://www.europeana.eu" xmlns:dc="http://purl.org/dc/e
    <dc:identifier>BzF_Desmidiaceae
    Engels_Main-1-6489</dc:identifier>
    <dc:title>Closterium angustatum var. sculptum (Raciborski)
    Ruzicka 1973</dc:title>
    <dc:subject>Closteriaceae familia</dc:subject>
    <dc:subject>Closteriineae subordo</dc:subject>
    <dc:subject>Desmidiales ordo</dc:subject>
    <dc:subject>Conjugatophyceae classis</dc:subject>
    <dc:contributor>Schmidle, W.</dc:contributor>
    <dc:description>A new record of reference material of the
    Desmidiaceae of Germany was compiled, using the databank
    "Specify". Aim of the project was to summarize and
    digitalize the data of the dried and otherwise preserved
    specimens stored in the German herbaria as well as of the
    living strains, cultured in culture collections. Many of the
    references are completed by figures of the labels or by scanned
    micrographs from preparations for the light- or scanning
    microscopes.</dc:description>
    <dcterms:spatial>Kaiserslautern</dcterms:spatial>
    <dcterms:isPartOf>Desmidiaceae Engels</dcterms:isPartOf>
    <dc:source>Specimen</dc:source>
```

Figure 12: XML presentation of a result record

There exists also a "look up" facility:

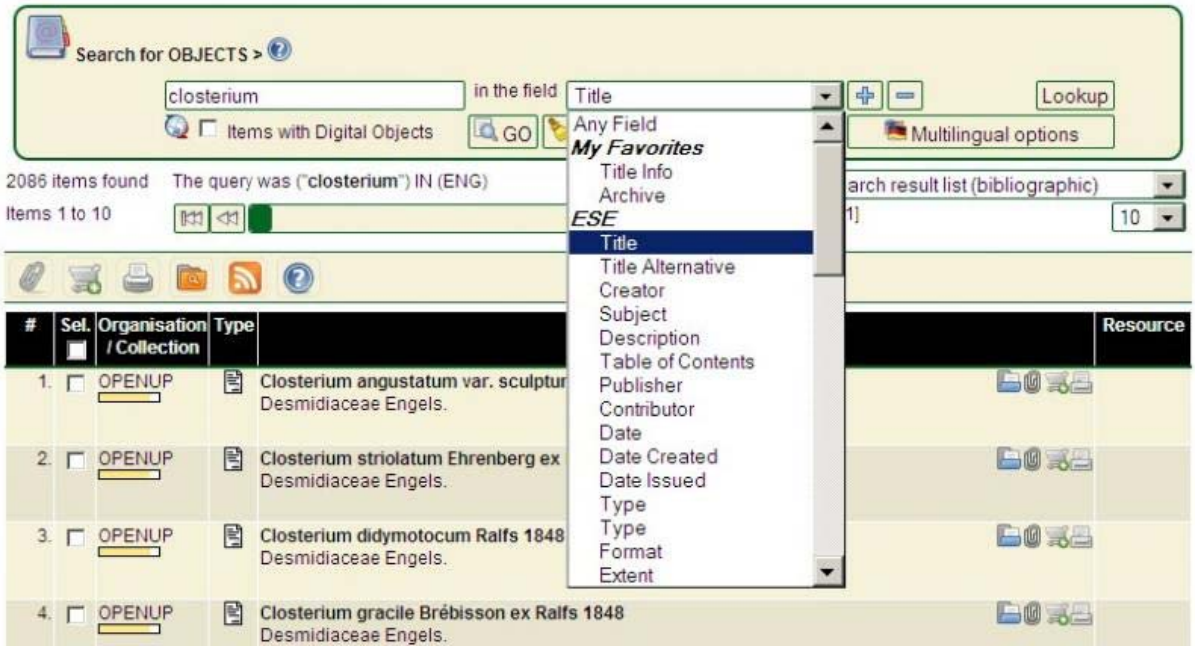


Figure 13: Search possibilities using the ESE Schema

An - index - look up for the ESE field "Title" for example delivers an alphabetic index containing the frequency of terms used:

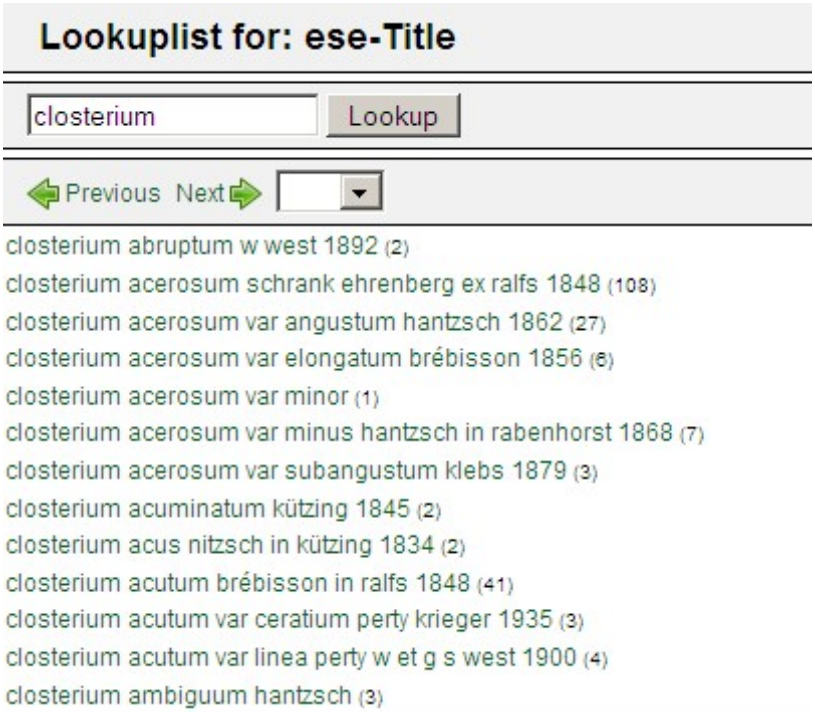


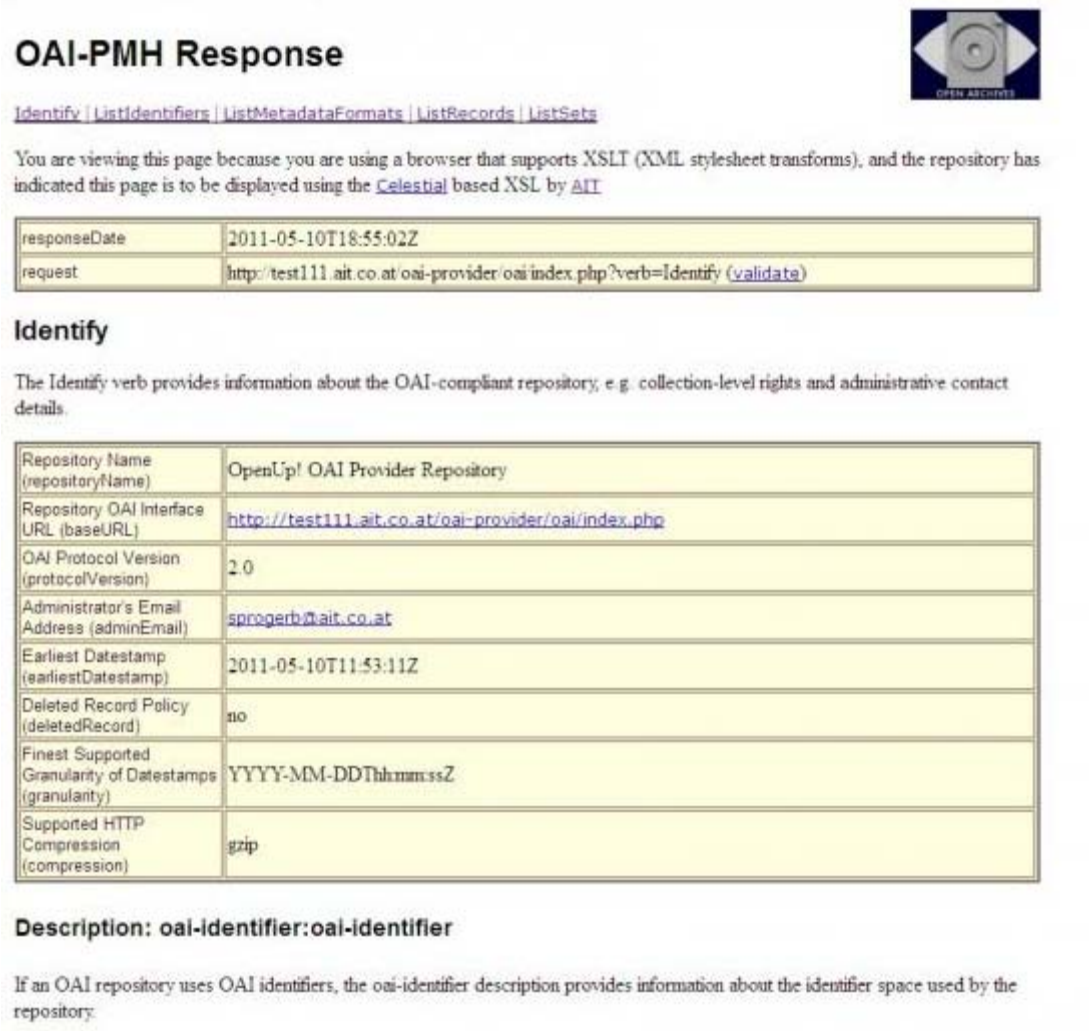
Figure 14: (Index) Look Up (ESE-Title meta data element)

This system (it is intended to substitute it by a migrated version to Drupal7, delivered by the BHLE-Project) provides a final checking facility. [The future Drupal7 based system stores records in "Drupal7 Nodes" following the METS structure which allows further integration of an archival system.]

Only "Items with **"Digital Objects"** are selected and handed over to the "Open Up! meta data" system provided by NHM for further processing, since Europeana only accepts records which can provide links to digital content (images, scanned pages, audios, videos, etc.).

A life version of this system can be accessed at an AIT-Test-Server[29].

This component (Europeana - Natural History Aggregator) provides also access to the "OAI-PMH Service" component:



OAI-PMH Response

Identify | ListIdentifiers | ListMetadataFormats | ListRecords | ListSets

You are viewing this page because you are using a browser that supports XSLT (XML stylesheet transforms), and the repository has indicated this page is to be displayed using the Celestial based XSL by AIT

responseDate	2011-05-10T18:55:02Z
request	http://test111.ait.co.at/oai-provider/oai/index.php?verb=Identify (validate)

Identify

The Identify verb provides information about the OAI-compliant repository, e.g. collection-level rights and administrative contact details.

Repository Name (repositoryName)	OpenUp! OAI Provider Repository
Repository OAI Interface URL (baseURL)	http://test111.ait.co.at/oai-provider/oai/index.php
OAI Protocol Version (protocolVersion)	2.0
Administrator's Email Address (adminEmail)	sprogerb@ait.co.at
Earliest Datestamp (earliestDatestamp)	2011-05-10T11:53:11Z
Deleted Record Policy (deletedRecord)	no
Finest Supported Granularity of Datestamps (granularity)	YYYY-MM-DDThh:mm:ssZ
Supported HTTP Compression (compression)	gzip

Description: oai-identifier:oai-identifier

If an OAI repository uses OAI identifiers, the oai-identifier description provides information about the identifier space used by the repository.

Figure 15: Example of an OAI "Identify" response using the OpenUp! OAI-Data Provider

29 http://test111.ait.co.at/oai-provider/index.php?form=index&db=0

Records can be listed using the command:

http://test111.ait.co.at/oai-provider/oai/?verb=ListRecords&metadataPrefix=ese[30]

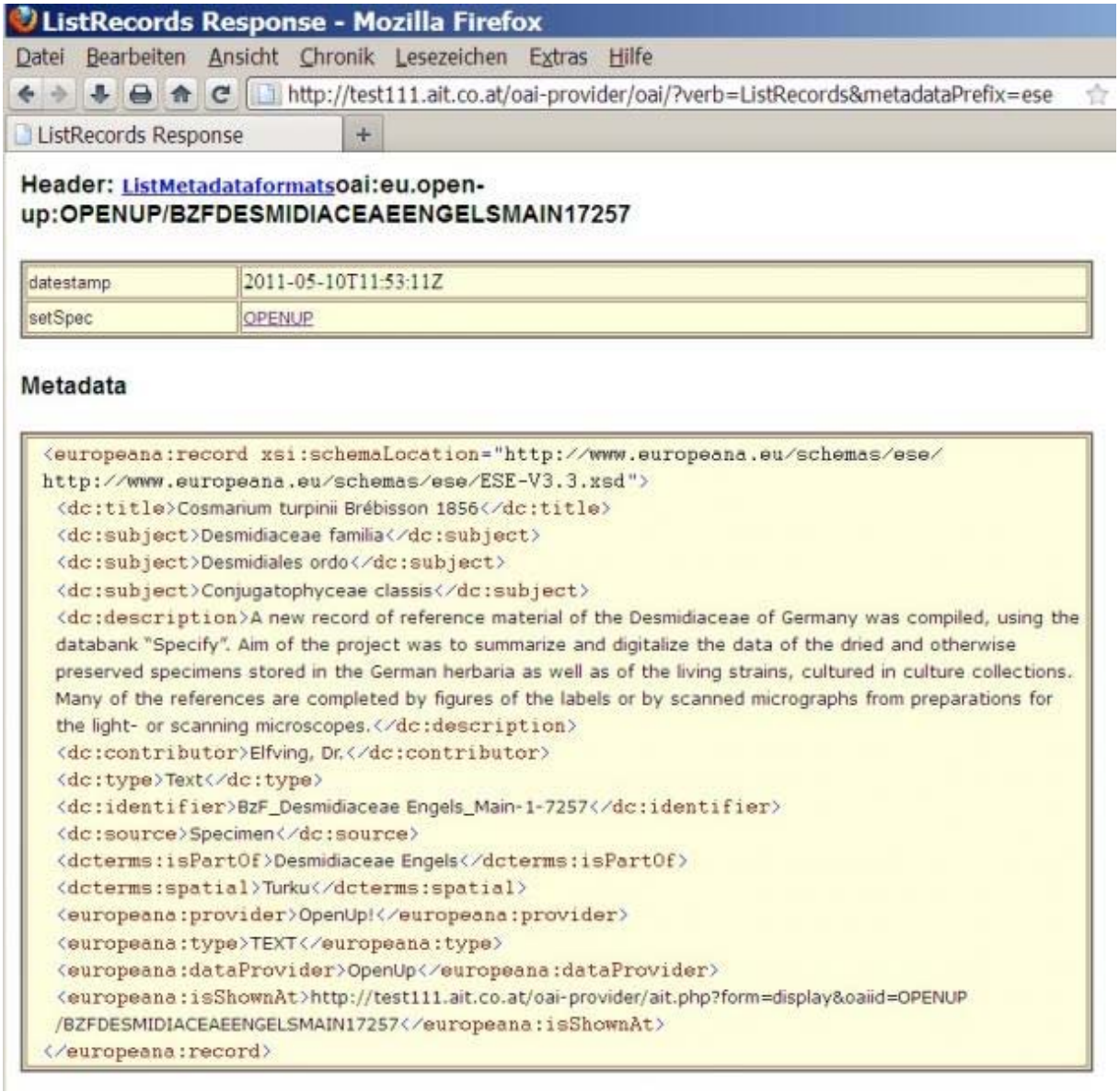


Figure 16: Sample records provided by the OpenUp! OAI-provider (test mapping only)

Please note that each command issued to the OpenUp! OAI-Provider has to include the parameter "&metadataPrefix=ese" to enable presentation of the transformed ABCD records in ESE format.

The "RAW XML+ESE Data Exchange" component prepares the export of the records to be collected by the Europeana Harvester and unloads data into a directory which can be accessed by the OpenUp!

30 http://test111.ait.co.at/oai-provider/oai/?verb=ListRecords&metadataPrefix=ese

Meta Data Store. This data store is located at and managed by a SQL-Server RDBMS. A possible DB structure can be taken from following ERD:

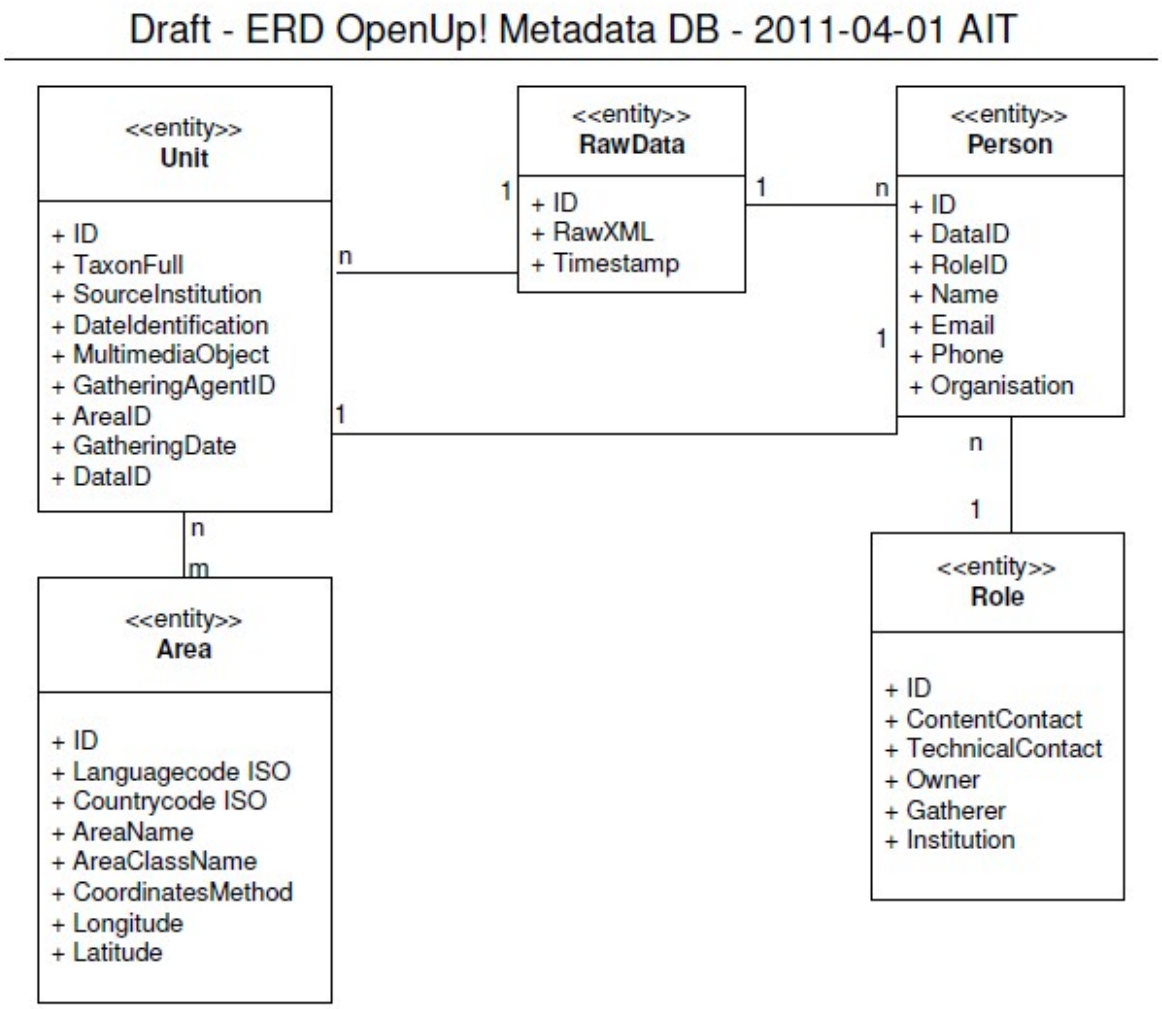


Figure 17: Possible data structure of the Meta Data Base

The Availability Checker finally determines what records are handed over to the OAI-PMH service (Data Provider) component.

The OAI Data Provider located at the Drupal7 execution environment will be based on a refactored PHP application which is already used in different Europeana projects (eConnect/DISMARC[31], EuropeanaLocal[32]). [A first version of this refactored OAI-Provider will be available by end of May.]

During the Transformation Process at the Pentaho execution environment data can be added by looking up vocabularies and consuming webservices providing access to (multilingual) vocabularies. This process has been successfully implemented in the aggregation platform described above as well

31 http://www.dismarc.eu/

32 http://www.europeana-local.at/

tested in other projects like BHL-Europe[33] (publication[34] [1], presentation[35][2]) and in collection management systems (attached[36][3]). The interaction with uBio's Taxon Finder is outlined in a BHLE presentation (August 2009)[37][4].

To harmonize the input coming from different "vocabulary WebServices" an "Ontology Gateway" will be installed in front of the transformation environment. As reference for such a component AIT's TGN-Vocabulary[38] WebService[39] can be used.

Harvesting and Transformation component prototype

The Harvesting and Transformation component prototype consists of a first implementation of the components as described in the "Nodes and Components" paragraph (Figure 2:) above and following the workflow included in the process diagram (paragraph "Process Analysis" Figure 1:) above. These components are deployed on a test server and include:

1. GBIF-HIT[40] Harvester (Demo Application available at: http://test111.ait.co.at:8080/hit[41] , user:admin - password:ait111).

2. Pentaho Data Integration[42] component (supporting the ABCD -> ESE mapping, including consumption of Ontology Services).

3. OpenUp! Meta Data Data Manager (SQL-Server[43] and Data Import Utility).

4. OAI-Provider[44] (Demo Implementation available at: http://test111.ait.co.at/oai-provider/oai/ [45]) .

These components will elaborated and more detailed in course of the project. The management of the data sets to be harvested by the Europeana Harvester will be supported by the "Data Manager" and needs close co-operation with the BioCASe (GBIF) data providers to align organisational structures and necessary data transformation processes.

33 http://www.cacaoproject.eu/at4dl

34 http://open-up.eu/sites/open-up.eu/files/u22/AT4DL_BHLEurope_final.pdf

35 http://open-up.eu/sites/open-up.eu/files/u22/AT4DL_Koch_Scholz.pdf

36 http://open-up.eu/sites/open-up.eu/files/u22/mai2007bulle.pdf

37 https://bhl.wikispaces.com/file/view/AIT_WP3_Pentaho.pdf

38 http://www.getty.edu/research/tools/vocabularies/tgn/index.html

39 http://demo.ait.co.at/thesaurus/

40 http://code.google.com/p/gbif-indexingtoolkit/

41 http://test111.ait.co.at:8080/hit

42 http://www.pentaho.com/products/data_integration/

43 http://en.wikipedia.org/wiki/Microsoft_SQL_Server

44 http://www.openarchives.org/

45 http://test111.ait.co.at/oai-provider/oai/?verb=Identify

Notes

- The "Harvesting and Transformation component prototype" uses in a first step as "execution Environment" a Z39.50 based infrastructure (Zebra-Server) which is intended to be substituted in the final version by a Drupal7 based environment which includes also a Control & Administration component.

- The prototype software is installed and demonstrable by May 10th 2011.

List of References

[1] Koch W., Scholz H., *DISMARC and BHL-Europe: multilingual access to two aggregation platforms for Europeana*. Available from: <http://open-up.eu/sites/open-up.eu/files/u22/AT4DL_BHLEurope_final.pdf> [13 May 2011].

[2] Koch W., Scholz H., *DISMARC and BHL-Europe: multilingual access to two aggregation platforms for Europeana.* Available from: <http://open-up.eu/sites/open-up.eu/files/u22/AT4DL_Koch_Scholz.pdf> [13 May 2011].

[3] Bulle K., 'Wer die Kirche im Dorf lässt, braucht das Rad nicht neu zu erfinden: Verschlagwortung mit Web Services'. Paper presented at the *Mai-Tagung 2007 on 10/11 May 2007*, ZMK | Zentrum für Kunst und Medientechnologie, Karlsruhe.

[4] Koch W., August 2009, 'Transformation of BHL Europe data using Pentaho Data Integration'. Paper presented at a BHLE Meeting, Ort.