DELIVERABLE

Project Acronym:	**Natural Europe**
Grant Agreement number:	**250579**
Project Title:	**Natural Europe: Natural History & Environmental Cultural Heritage in European Digital Libraries for Education**

D4.5 – Deployment of Natural Europe Interplay with External Platforms

Revision: [final]

Authors:

Stauros Gkinis (GRNET)

Giannis Stoitsis (GRNET)

Anastasios Koutoumanos (GRNET)

Nikos Manolis (GRNET)

Nikos Manouselis (GRNET)

Bram Vandeputte (KULeuven)

Joris Klerkx (KULeuven) – Subcontract with GRNET

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	


natural
europe

Revision history:

Revision	Date	Author	Organization	Description
0.1	1/6/2012	S. Gkinis, N. Manolis,	GRNET	First Draft
0.3	15/6/2012	N. Manouselis, G. Stoitsis	GRNET	1st Revision
0.5	1/7/2012	S. Gkinis, N. Manolis, N. Manouselis, G. Stoitsis	GRNET	2nd Revision
1.0	19/7/2012	J. Klerkx; B. Vandeputte	KULeuven	Final version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Executive Summary

This deliverable consists of the presentation of the deployment of the Natural Europe infrastructure and the supported technologies for interconnection with external platforms

Table of Contents

List of Figures

1 Introduction

1.1 Scope

This document provides an overview of the deployment of the Natural Europe infrastructure and the supported technologies for interconnection with external platforms.

1.2 Audience

This report is addressed to Natural Europe project partners, the European Commission, and the general public

1.3 Abbreviation & Definitions

API: Application programming interface

ESE/EDM: Metadata schemas used by Europeana

Open Archival Initiative -Protocol for Metadata Harvesting (OAI-PMH):
A protocol for metadata harvesting that supports selecting metadata records from repositories based upon their identity, the date of their last modification, and their membership in predefined sets

NHM: Natural history museum

Representational State Transfer (REST): An approach for getting information content from a Web site by reading a designated Web page that contains an XML (Extensible Markup Language) file that describes and includes the desired content

Simple Publishing Interface (SPI): A language for ingesting new metadata and/or resources in repositories with LOM descriptions

Simple Query Interface (SQI): A conceptual protocol for searching in repositories with LOM descriptions

Search/Retrieve via URL (SRU): A REST style protocol that encodes the search method and parameters as a URI and returns an XML instance.

Search/Retrieve Web Service (SRW): It binds the SRU protocol to a SOAP implementation

1.4 Structure

Chapter 1: contains information about the content of the deliverable, the scope and audience, together with the necessary definitions.

Chapter 2: describes the Natural Europe infrastructure as this has been deployed. Also it describes the technologies that allow the interplay with external platforms.


natural
europe

2 The Natural Europe Infrastructure Components

2.1 Overall architecture & Toolset

The Natural Europe Infrastructure for metadata aggregation has been based on the ARIADNE technologies and services (ARIADNE Foundation, 2012). **ARIADNE** is a standards-based technology infrastructure that allows the publication and management of digital learning resources in an open and scalable way.

The components of the ARIADNE infrastructure that have been deployed as Cultural Heritage Infrastructure Components are the following:

- **Harvester**: An application to manage harvesting of one or more OAI-PMH targets in a way that is automated as much as possible.
- **Validator**: A web application for checking both the syntactic and semantic validity of metadata instances against multiple standards, specifications and metadata schemes.
- **Repository Management**: A tool to for the management of learning objects in an open and scalable architecture.
- **Registry Management**: A web application to manage and provide up-to-date information on learning repositories.
- **ARIADNE Finder**: A web tool for faceted and text-based search.

The above components will be also deployed and configured for supporting the Educational Infrastructure.

Figure 2.1 depicts the overall architecture and how the aggregated metadata could be provided to external platforms and tools. The chapters to follow, present briefly the components deployed and the supported services and technologies in the architecture.

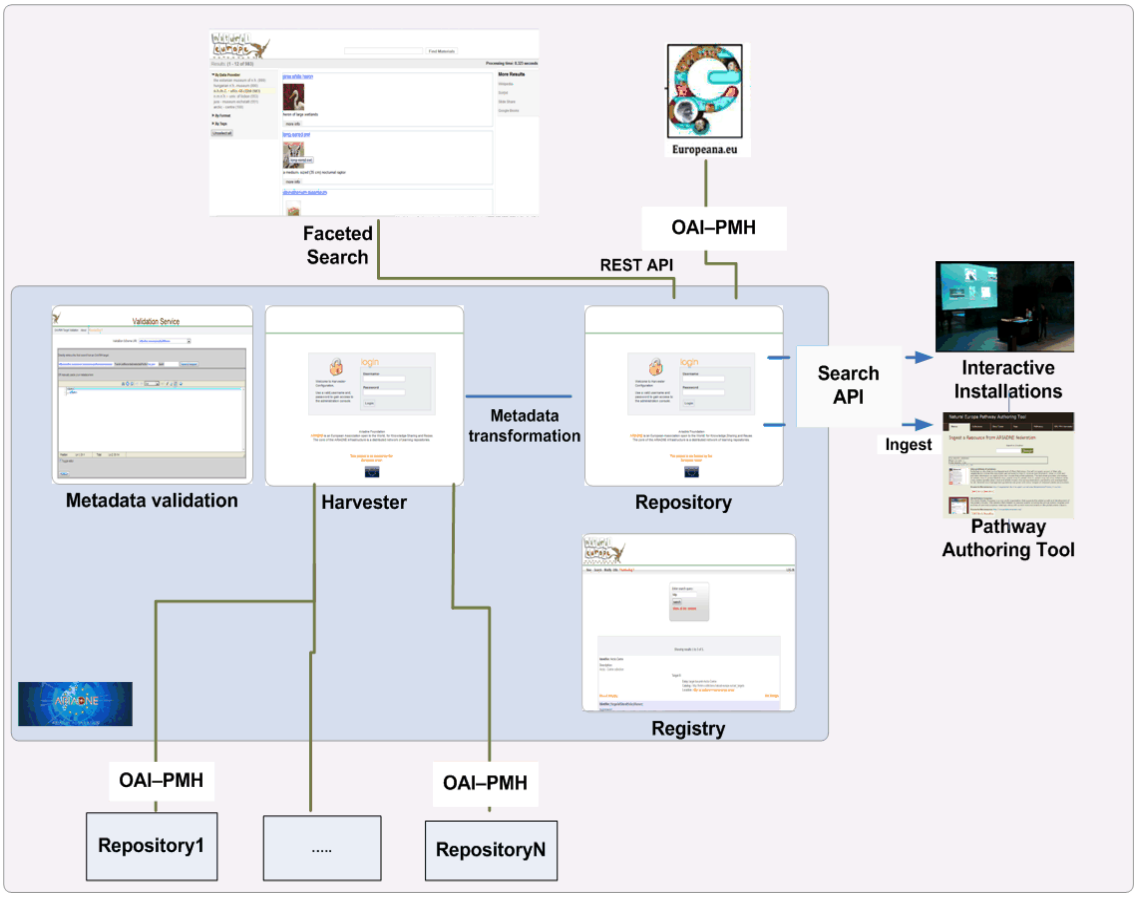

Figure 2.1: The Natural Europe Infrastructure & toolset

2.2 Natural Europe Harvester

In the context of the Natural Europe project, the **ARIADNE Harvester** component has been deployed and configured to manage harvesting metadata records provided by Natural Europe content providers. The **Natural Europe Harvester** uses the OAI-PMH framework for harvesting metadata records from an OAI-PMH target and publishes them to a central repository through the **publish service** that is described in subsection 2.2.2.1.

The **Natural Europe Harvester** is agnostic about the type of metadata to be harvested (DC, LOM, NE) and can be configured and extended for advanced use, such as creating validation reports, pushing metadata through a custom storage connector using an arbitrary protocol, on-the-fly transforming xml metadata to another structure.

2.2.1 Administration

2.2.1.1 Managing Content Providers

The administration page of the harvester, namely “OAI-PMH targets”, provides a set of services for the addition, modification and deletion of:

- **ContentProviders,** i.e the Natural History Museums.
- **Data Sources**, which are the sources of metadata record sets provided by a content provider through appropriate OAI-PMH services.

In specific, for the addition of a Content Provider, i.e. Hungarian Natural History Museum, to the targets of the harvester, the following information is possible to be provided:

- OAI URL: the URL of the OAI server from which the data will be provided
- OAI Set: the name of the Data Source on the OAI server. A set is an optional construct for grouping items for the purpose of selective harvesting in case the repositories organize items into sets.
- Metadata Prefix: The metadata prefix indicating the metadata format of therecords to be harvested.
- Repository Name: The name of the repository

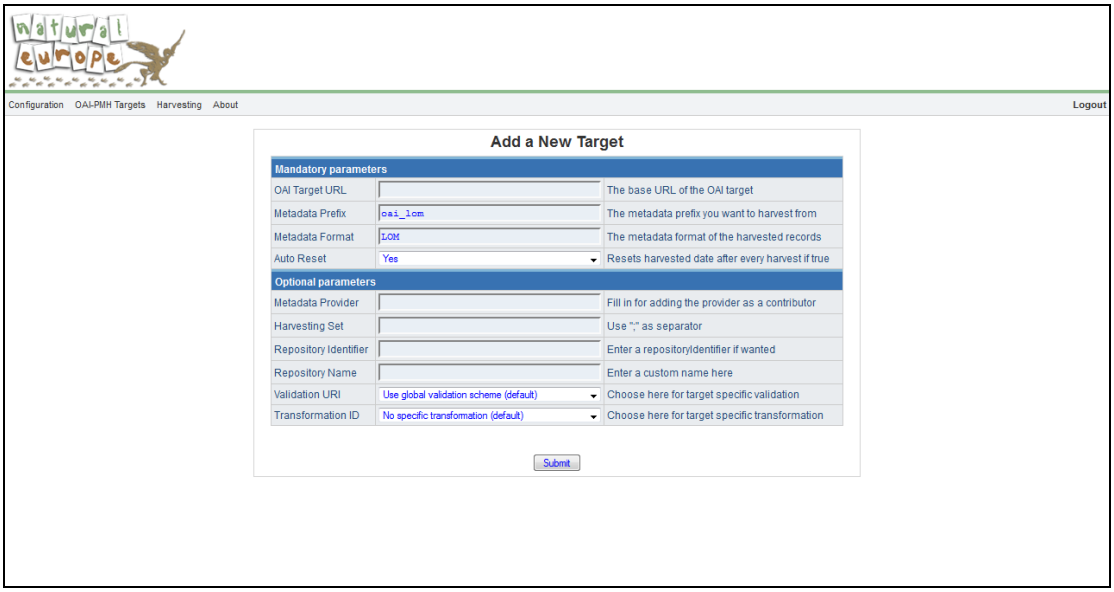

Figure 2.2: Adding a New Target

At the time of addition, every target is being checked if it is well structured according to OAI-PMH. In an opposite case the target is not added to the harvester and error messages are presented to the administrator.

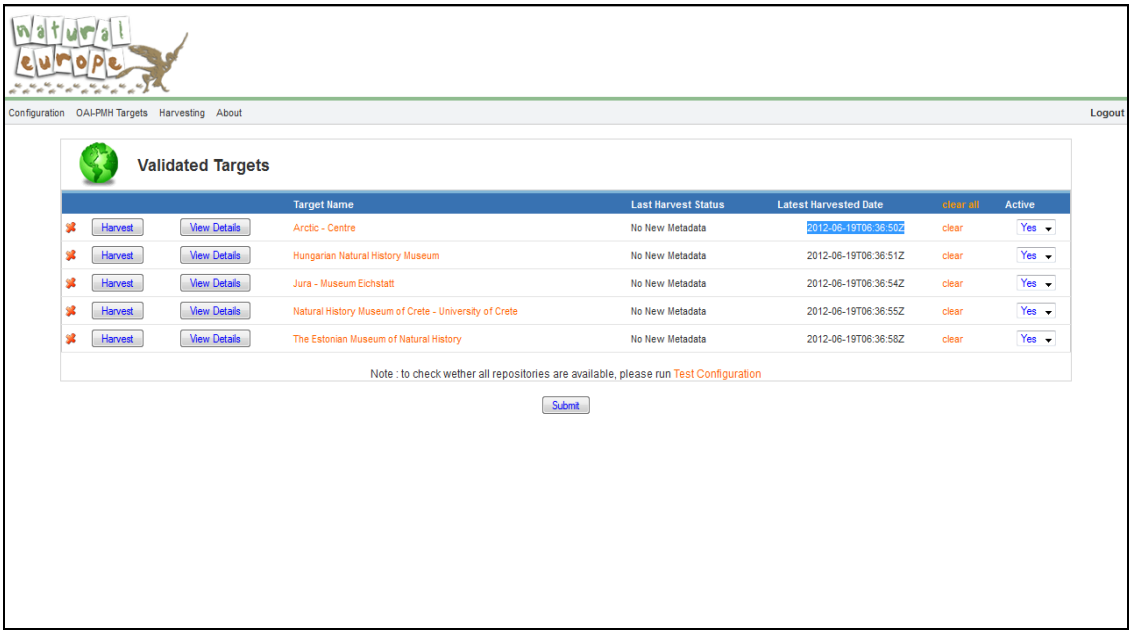

Figure 2.3: Natural Europe Harvester

2.2.1.2 Managing the Harvesting Process

There are two ways provided for harvesting metadata records from the Data Providers (OAI targets) that have been selected:

- **Harvest on demand.** The system administrator is given the ability to start the harvesting process at any desired time by invoking the corresponding process
- **Scheduled harvesting.** This option refers to the automated harvesting process, using a predefined schedule. For the creation of a schedule an appropriate page is offered ("View/Edit harvesting schedule") to the system administrator in order to create edit or delete a scheduled task. While defining a scheduled task the system administrator is able to choose the date and the time that the task will take place along with the task's frequency. Regarding the frequency of each task, the system administrator can choose at least among the following values:
  - **Once**, in order to trigger the harvest process at a specific date.
  - **Daily**, in order to trigger the harvest process every day at the same time.
  - **Weekly**, in order to trigger the harvest process every week in the same day at the same time.

Figure 2.4: Scheduled harvesting

Additionally, apart from the last harvested date, system administrator can be informed by the system for the status of each harvesting task. The status of a harvesting process can take the following values:

- **active**: informing the user that the harvesting process is currently active
- **successful** : the harvest process has completed successfully
- **error**: an error occurred during the harvest process

2.2.2 Integrated Services

2.2.2.1 Publish Service

For publishing the harvested metadata into a central repository we rely on the Simple Publishing Interface (SPI) specification (Ternier & Massart, 2008). The SPI provides a simple lightweight protocol for publishing data and metadata to a repository.

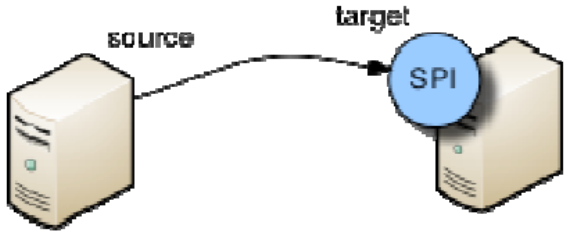

Figure 2.5: Example SPI architecture

Figure 2.6 illustrates how a system, labeled as 'source' issues a publication request to another 'target' system that can be a repository component or a middle layer component.

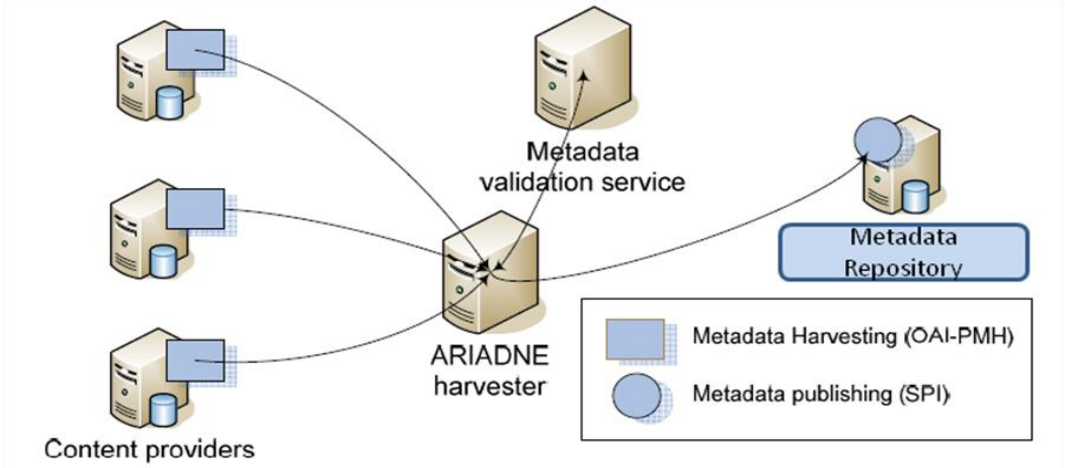

Figure 2.6: Metadata publishing

In our context, the 'source' is the Natural Europe harvester and the 'target' is the metadata repository that acts as an adapter to the publishing API. In order this communication to be possible the following parameters should be inserted to the Natural Europe harvester:

1) The URL address of the SPI target built over the repository (e.g. http://localhost:8080/NaturalEurope/services/SPI)
2) The URL address of the service that is responsible for authorization (e.g. http://localhost:8080/AriadneRepository/services/SqiSessionManagement) and the appropriate credentials.

2.2.2.2 Transformation service

The transformation service converts metadata in one format into another format. This service is needed as the metadata scheme applied in a content provider's repository may be different from the metadata scheme applied in the repository with the aggregated metadata.

For instance, Natural History Museums (NHMs) typically use "The Natural Europe Europeana Semantics Elements – Cultural Heritage Application Profile" for describing their content. This format has to be mapped to Europeana/ESE specification in order the aggregated CHO

metadata records of the participating Natural History Museums to be accessed by the Europeana.eu.

For enabling an easy and rapid migration from the Natural Europe ESE AP to Europeana/ESE specification, the transform service allows an automatic migration.

2.2.2.3 Identification service

The Identification service is used to provide persistent digital identifiers to resources in the ARIADNE infrastructure. The HANDLE system (PILIN Project, 2007) is used as the backend service to create globally unique, persistent and independent identifiers. This system allows the assignment, management and resolution of persistent identifiers in a distributed environment. The lower level API provided by the HANDLE system is connected to an independent service interface that provides the basic functionality for persistent storage: Create, Resolve, Update and Delete (CRUD) identifiers. The identifiers created by the service are compliant with the Universally Unique Identifier standard (ISO/IEC, 1996). For this purpose the Java Uuid Generator (Safehaus, 2010) is used.

2.2.2.4 Metadata validation service

The metadata validation service is available for providing validation of metadata instances against predefined application profiles, for example based on Natural Europe ESE AP. To ensure that only compliant metadata are stored in the repository, we use the validation service to check both the syntactic and semantic validity of the instances against the used profiles. The validation service has a modular approach, and combines different sorts of validation techniques including:

- XSD schema, mainly used for structural validation of xml metadata.
- Schematron rules, which are very flexible and are used for:
    - verifying the presence of mandatory elements
    - checking the presence of empty attribute fields. For example, in the langstring datatype of LOM, the language of a string should be recorded in a non-empty attribute “language”.
    - enforcing inter-field dependencies, like conditional fields.
    - checking for the correct use of the allowed terms in a vocabulary.
- validation of vcards present in the metadata with a separate vcard parser or validator.

The validation service is integrated in the harvester and the administrator can choose to validate all targets against one validation scheme by choosing this scheme in the configuration wizard or choose a validation scheme for an individual harvesting target.

Moreover, the validation service has been deployed as an online service where one single metadata record or all records exposed through OAI can be validated against the appropriate scheme. Reports are automatically generated which give an overview of all validation errors.

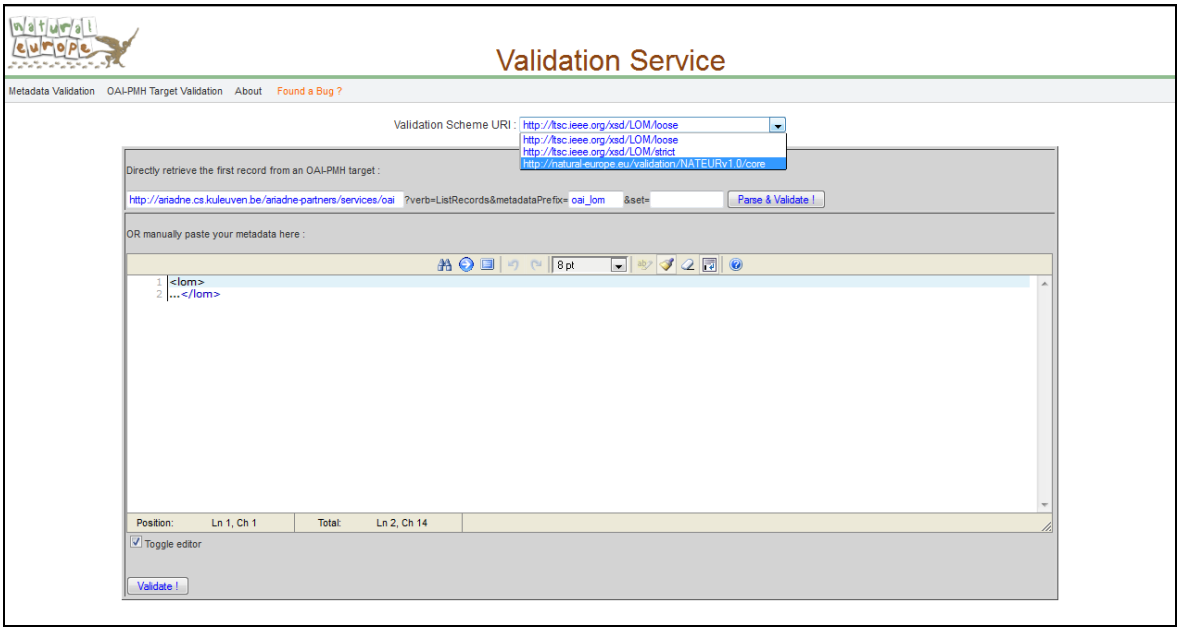

Figure 2.7: Online validation service

2.3 Metadata Repository Software

In the context of the Natural Europe project, the Metadata Repository Software has been deployed, based on the **ARIADNE Next Repository** software. It features both a metadata and file store where learning objects and metadata instances are persistently managed in an open and scalable architecture. Through several open standards and specifications it supports stable querying, publishing and harvesting of learning objects. Also, it provides a flexible building block that can be adapted to different situations and that enables interoperability with external components (e.g. the harvester component). In the following sub-sections we describe the basic characteristics and supported technologies of the Metadata Repository software.

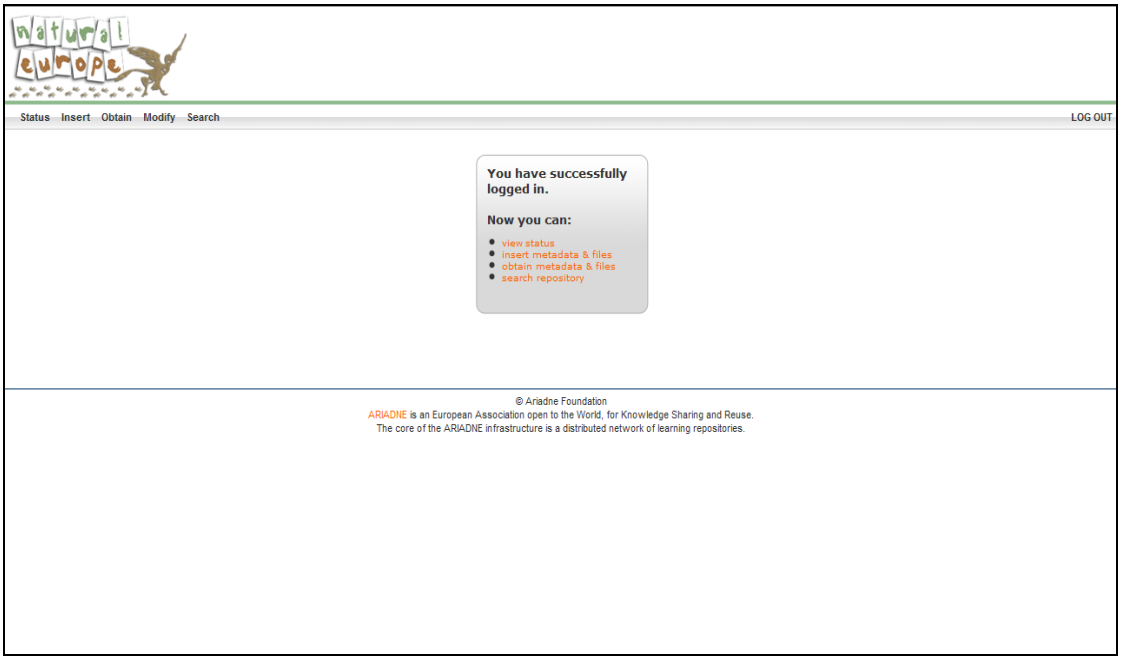

Figure 2.8: Metadata Repository Software

2.3.1 Configuration

The administration page of the Metadata Repository Software, namely "changeConfiguration", provides the administrator with a means to insert all the required parameters for the configuration regarding:

- **Metadata storage**: There are different metadata implementations that can be used for storing the metadata, using either the file system or a database (DB2, Exist, Oracle).

  For insertion using the file system, this parameter must be active:

  mdstore.insert.implementation=
  org.ariadne_eu.metadata.insert.InsertMetadataFSImpl

  Also, this property must be configured according to the local installation:

  mdstore.spifs.dir = /REPOSITORY_FOLDER_PATH/data/mdstore/

- **Search Index**: In order to support a fast search over the metadata instances Apache Lucene is used. By default the Lucene document handler is the LOMHandler and the Lucene document analyzer is the LOMDocumentAnalyzer (both tailored for LOM metadata instances). The only parameter that need to be configured in order to support an index search is:

  search.lucene.indexdir = /REPOSITORY_FOLDER_PATH/data/index/

For querying, these parameters must be active:

mdstore.insert.implementation.1 =
org.ariadne_eu.metadata.insert.InsertMetadataLuceneImpl

mdstore.query.implementation.0 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore.query.implementation.1 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore.query.implementation.2 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore.query.implementation.3 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore.query.implementation.4 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore.query.implementation.5 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore.query.implementation.6 =
org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

mdstore

mdstore.insert.implementation.1	org.ariadne_eu.metadata.insert.InsertMetadataLuceneImpl
mdstore.db.username	admin
mdstore.delete.implementation	org.ariadne_eu.metadata.delete.DeleteMetadataFSImpl
mdstore.insert.implementation	org.ariadne_eu.metadata.insert.InsertMetadataFSImpl
mdstore.db.uri	xmldb:exist://localhost:8082/exist/xmlrpc/db/metadatastore
mdstore.db.xmldb.loc	collection(
mdstore.delete.implementation.1	org.ariadne_eu.metadata.delete.DeleteMetadataLuceneImpl
mdstore.spifs.dir	/home/user/tomcat6/webapps/cultural_repos/data/mdstore/
mdstore.db.password	admin
mdstore.insert.xmlns.xsd	http://ltsc.ieee.org/xsd/LOM
mdstore.query.implementation	org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl
mdstore.query.implementation.6	org.ariadne_eu.metadata.query.QueryMetadataLuceneImpl

Figure 2.9: The administration page of the Metadata Repository Software

2.3.2 Indexing

To enable efficient and fast search on top of large metadata collections the open source Apache Lucene (v2.9.4) has been used. This text search engine library allows to index all the elements of the applied metadata scheme (e.g LOM) in the repository and therefore to enable fast search on top of them. The elements are identified according to their paths (e.g. classification.taxonpath.taxon.entry.string) and indexing elements takes place during the insertion of metadata in the repository. However, the administrator is able to trigger an index recreation and to check if the index has been created appropriately.

Figure 2.10: Recreating Index

Moreover, the Apache SOLR framework (v1.4.1) is being used to provide powerfull, efficient and facetted search algorithms. Solr runs as a standalone full-text search server within a servlet container (Tomcat) and uses the Lucene search library at its core for full-text indexing and search. Also it has REST-like HTTP/XML and JSON APIs that make it easy to use from virtually any programming language.

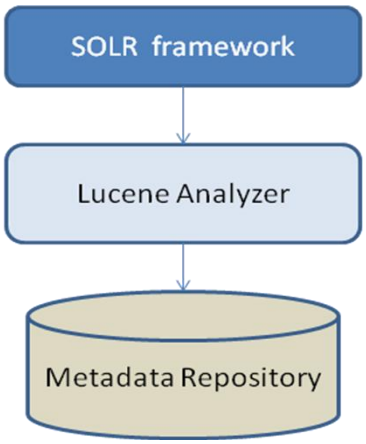

Figure 2.11: Indexing metadata instances

2.3.3 Federated Search

A key technology to offer transparent search to Natural Europe metadata repositories is the **ARIADNE Federated Search Service.** This service relies on Simple Query Interface (SQI) that is designed to support many types of search technologies (Ternier, Massart et al, 2008).

The basic requirements captured by SQI are the following:

- SQI is neutral in terms of results format and query languages. The repositories connecting via SQI can be of highly heterogeneous nature: therefore, SQI makes no assumptions about the query language or results format.
- SQI supports Synchronous and Asynchronous Queries in order to allow heterogeneous networks to connect to each other. In the *synchronous* scenario, the target returns the query results to the source. Results retrieval is therefore initiated by the source through the submission of the query and through other methods allowing the source to access the query results.

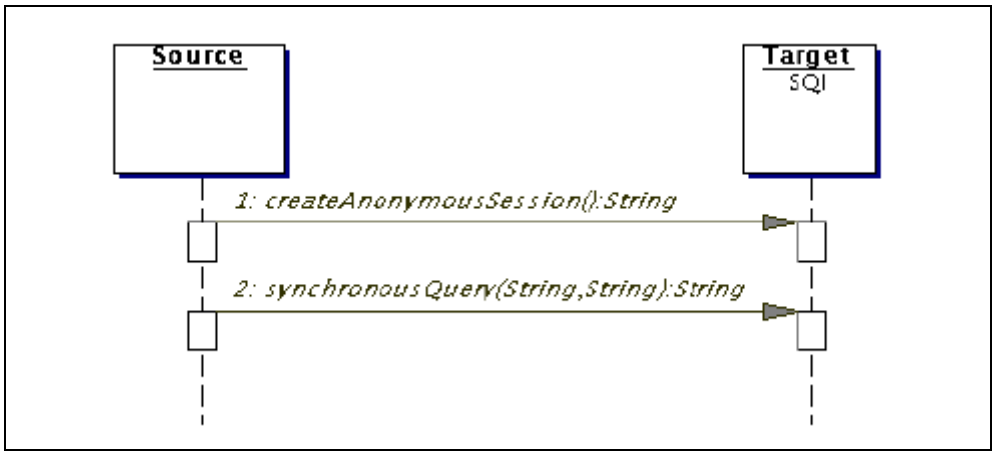

Figure 2.12: Synchronous Query Mode used for querying a single Repository

Once a session has been established between the source and the target, the query interface at the target awaits the submission of a search request. A number of methods allow for the configuration of the interface at the target. According to the SQI specification the following query parameters can be set:

- the query language (setQueryLanguage),
- the number of results returned within one results-set (setResultsSetSize),
- the maximum number of query results1 (setMaxQueryResults),
- the maximum duration of query execution (setMaxDuration),
- and the results format (setResultsFormat)

The provided bindings that implement the Simple Query Interface (SQI) support both SOAP and REST protocols. Some guidelines for using the search service are presented below.

2.3.3.1 SOAP service

Regarding the SOAP binding, the following two addresses are required in order to use the service:

- **SessionManagement**:

http://{SERVER_NAME }/{REPOSITORY_NAME}/services/SqiSessionManagement?wsdl

Example: http://ariadne.cs.kuleuven.be/ariadne-partners/services/SqiSessionManagement?wsdl

- **Target:** http://{SERVER_NAME}/{REPOSITORY_NAME}/services/SqiTarget

Example: http://ariadne.cs.kuleuven.be/ariadne-partners/services/SqiTarget

The main steps for using the service are:

1. Obtain an anonymous session by calling the session management service method: *createAnonymousSession*. The obtained token will be used in the following method calls.
2. Specify the results format required (e.g. "lom") with this method from the target service: setResultsFormat.
3. State and execute your query to the target with the method: synchronousQuery By default the result set size will be 12 and the query language will be PLQL Level 1 that mixes exact and approximate search criteria (Ternier, Massart et al. 2008).

2.3.3.2 REST service

Regarding the REST binding, the following address and parameters are required to call the service:

Target: http://{SERVER_NAME}/{REPOSITORY_NAME}/api/SqiTarget

Parameters:

- query: expressed as plql level 1
- size: the cardinality of the result set
- start: a number indicating the starting point of results
- lang: the query language, e.g. plql
- format: the metadata format supported

By default the result set size is 12 and the query language will be PLQL level 1

Example:http://ariadne.cs.kuleuven.be/ariadne-partners/api/sqitarget?query=learning&start=1&size=2&lang=plql1&format=lom

2.3.4 OAI-PMH Service

The metadata repository offers an OAI-PMH target service in order to allow the metadata exposure. In the context of Natural Europe, this service will be used by the Europeana.eu in order to access the aggregated metadata records of the participating Natural History Museums. These metadata conform to ESE.

According to the OAI-PMH V2.0 specification, a target service has to implement the following verbs (i.e. methods). Each verb corresponds to an OAI-PMH request.

Verb	Result
Identify	The identification of the underlying repository
ListMetadataFormats	A list of supported metadata formats
ListIdentifiers	A list of identifiers
ListRecords	A list of metadata records
ListSets	A list of available sets
GetRecord	A metadata record containing the description about an item

These verbs are requested in a REST style encoded as a URL. Through this URL, an application can retrieve metadata stored in the underlying repository, formatted in XML.

The metadata repository with the cultural aggregated metadata records will be accessible from such a URL so that the Europeana.eu can reach and harvest it.

2.4 Registry Service

The Registry service of the ARIADNE infrastructure is a catalog service that provides up-to-date information on learning object repositories (LORs). It provides the information necessary for systems to be able to select the appropriate protocols such as OAI-PMH, SQI, SPI, SRU/SRW supported by a given learning object repository. Also, it facilitates interoperability between learning object repositories.

In the context of the Natural Europe the Registry service has been deployed and configured to keep information about the repositories that are included in the Natural Europe federation. The end user can see all the participating repositories and get information about them, e.g. their OAI-PMH target. Moreover, all metadata related to a registered repository can be shown and thus the user could find information such as the metadata formats supported in a repository.

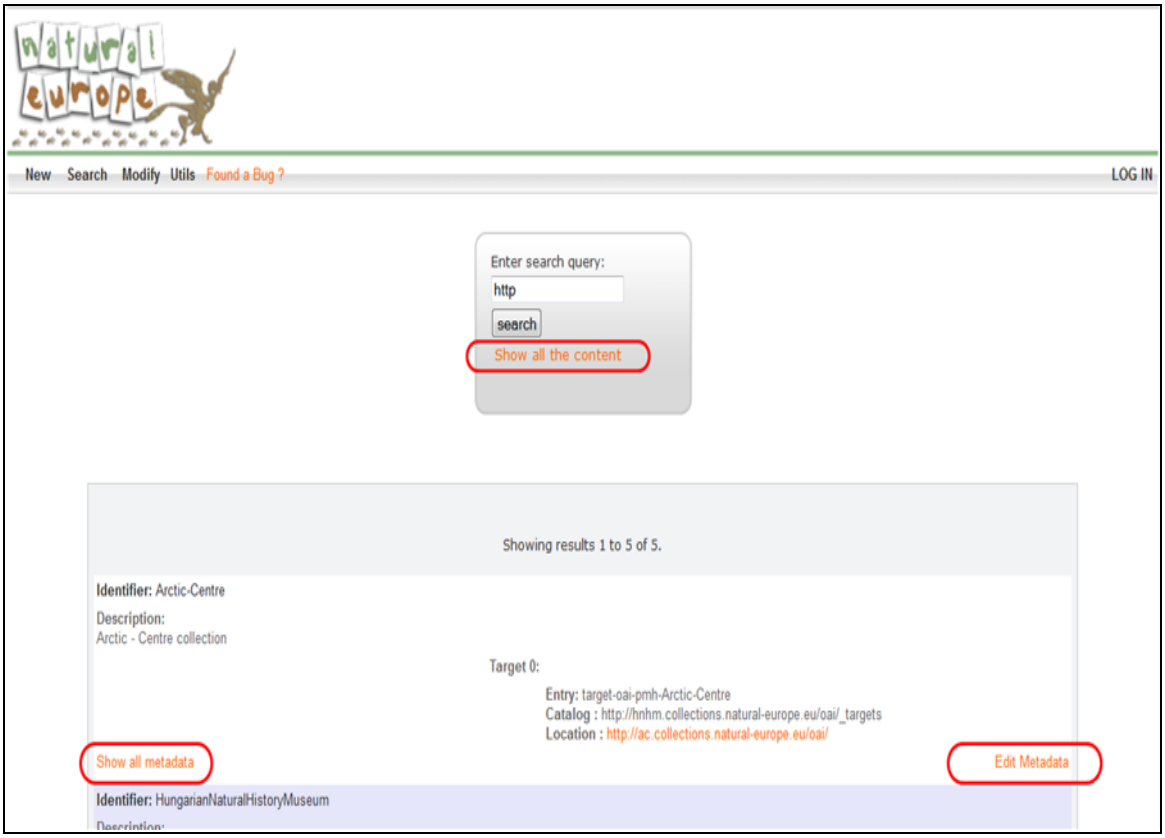

Figure 2.13: The Registry Service

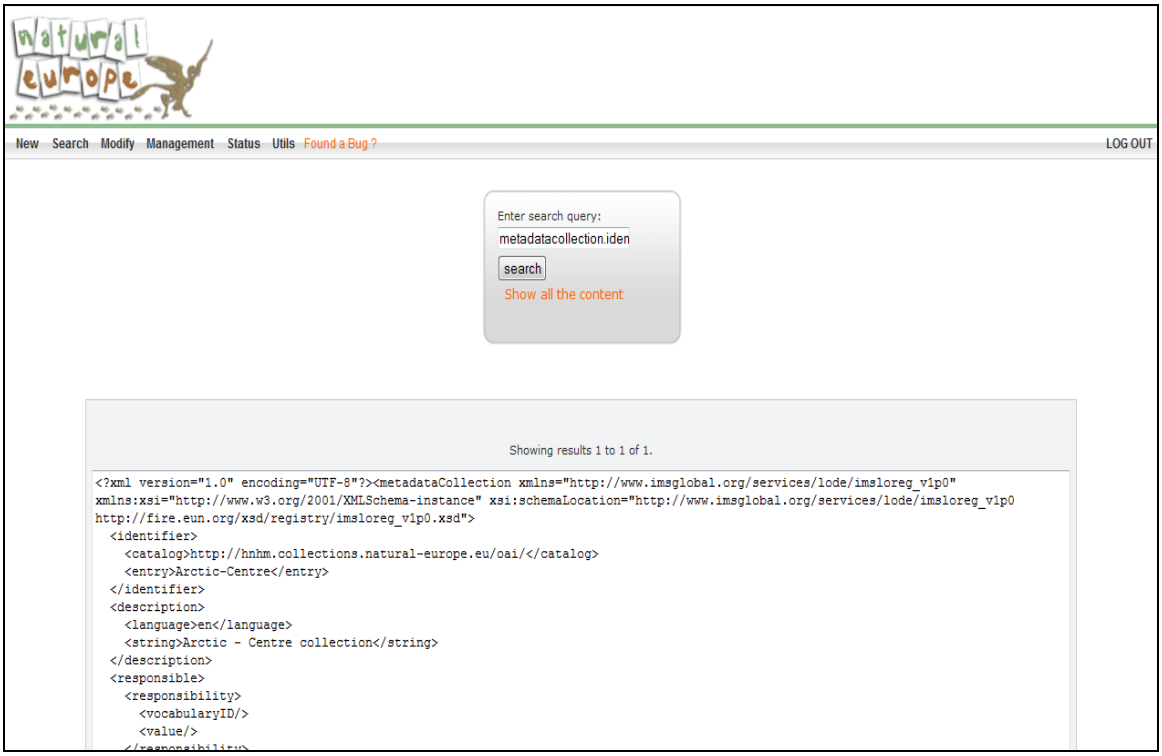

Figure 2.14: Show all metadata related to a registered repository

2.4.1 Managing Repositories

From an administrator's point of view, it is very easy to add a new repository to the registry by using a wizard and thus providing the required information about a repository such as administrative information (e.g. name, description, contact person email) and technical information (e.g. OAI target).

Figure 2.15: Adding a new repository to the registry service

2.5 Searching Natural Europe resources

In order to let users search cultural material and browse the results, the **ARIADNEFinder** has been deployed and configured appropriately. The **ARIADNEFinder** is a standalone and lightweight search tool that lets users search learning material and browse the results. The finder hides protocols and standards that are used in the middle layer. Also, it can be coupled on any repository that supports an implementation of a REST-based API with a JSON binding of LOM. Regarding the Natural Europe instance of finder, this has been coupled on the **Metadata Repository** component. The **ARIADNEFinder** is presented in detail in deliverable D4.4f “Faceted Search Service Web Interface”.

3 Impact on Natural Europe Vision

This section in each deliverable is dedicated to identifying its connections to the Vision of the Natural Europe project. The process in which the vision of the project is affected through the lifetime of the project is clearly documented in D2.1 "White Paper on Natural Europe Vision". For each deliverable, its author answers the questions presented below, summarizing the outcomes in less than a page, which is then incorporated in the D2.1 document.

- How is this deliverable affecting the services that Natural Europe is deploying?
  - *How is it going to affect access to cultural content from museums of natural history?*
  - *How is it going to affect searching for cultural content?*
- How is this deliverable affecting the outreach of Natural Europe to its audiences?
  - *How is this deliverable going to extend the audiences to which Natural Europe is targeted?*
  - *How is this deliverable going to affect Natural Europe outreach to specific audiences (museums' curators, teachers, parents, pupils, etc)*

This deliverable presents the final version of the deployment of the Natural Europe infrastructure and the supported technologies for interconnection with external platforms. The specific deliverable describes the components and services that have been deployed for supporting aggregation and storage of cultural content and that allow the interplay with external platforms. In the near future these technologies will be deployed for educational content too. In this sense, this deliverable affects the way cultural and educational content is aggregated, stored and accessed. Therefore, it will affect the way that audiences search for Natural Europe content and browse the results. Additionally, it directly affects the Natural Europe outreach to audiences by describing the supported technologies for providing content to external platforms and larger networks.


natural
europe

References

ARIADNE (2012), http://www.ariadne-eu.org/content/services

ISO/IEC. Iso/iec 11578:1996 information technology - open systems interconnection - remote procedure call (1996), http://www.iso.ch/cate/d2229.html

Natural Europe Deliverable D4.2. “Specification of the Natural Europe Platform & Tools”.

Natural Europe Deliverable D4.3. “Specification of the Natural Europe Interplay with External Platforms”.

PILIN Project. Pilin Project: Project closure report, http://www.pilin.net.au/Closure_Report.pdf

Safehaus. Java uuid generator(jug) (April 2010), http://jug.safehaus.org/Home

Ternier, S., Massart, D., Campi, A., Guinea, S., Ceri, S., Duval, E. (2008). Interoperability for searching learning object repositories, the prolearn query language. D-Lib Magazine 14(1/2) (January/February 2008)

Ternier, S., & Massart, D. V. (2008). A Simple Publishing Interface for Learning Object Repositories. *World Conference on Education Multimedia, Hypermedia and Telecommunications* (pp. 1840-1845). AACE