Noname manuscript No.
(will be inserted by the editor)

Metadata Management, Interoperability and Linked Data Publishing Support for Natural History Museums

Giannis Skevakis · Konstantinos Makris · Varvara Kalokyri · Polyxeni Arapi · Stavros Christodoulakis

the date of receipt and acceptance should be inserted later

Abstract Natural History Museums (NHMs) form a rich source of knowledge about Earth's biodiversity and natural history. However, an impressive abundance of high quality scientific content available in NHMs around Europe remains largely unexploited due to a number of barriers, such as: the lack of interconnection and interoperability between the management systems used by museums, the lack of centralized access through a European point of reference like Europeana, and the inadequacy of the current metadata and content organization. The Natural Europe project offers a coordinated solution at European level that aims to overcome those barriers. In this article, we present the architecture, deployment and evaluation of the Natural Europe infrastructure allowing the curators to publish, semantically describe and manage the museums' Cultural Heritage Objects, as well as disseminate them to Europeana.eu and BioCASE/GBIF. Additionally, we discuss the methodology followed for the transition of the infrastructure to the Semantic Web and the publishing of NHMs' cultural heritage metadata as Linked Data, supporting the Europeana Data Model (EDM).

Keywords Digital Curation · Preservation Metadata · SKOS · Linked Data · Europeana · BioCASE · GBIF

1 Introduction

Natural History Museums (NHMs) are unique spaces that have only recently come to comprehend the effectiveness of the learning opportunities they offer to their visitors [13]. Their scientific collections form a rich source of knowledge about Earth's biodiversity and Natural History. However, an impressive amount of high quality content available in European NHMs remains largely unexploited due to a number of barriers, such as: the lack of interconnection and interoperability between the management systems used by museums, the lack of centralized access through a European point of reference like Europeana, as well as the inadequacy of current content organization and the metadata used.

The Natural Europe project [8] offers a coordinated solution at European level that aims to overcome the aforementioned barriers, making the natural history heritage available to formal and informal learning processes. Its main objective is to improve the availability and relevance of environmental cultural content for education and life-long learning use, in a multilingual and multicultural context. Cultural heritage content related to natural history, natural sciences, and natural/environmental preservation is collected from six Natural History Museums around Europe into a federation of European Natural History Digital Libraries, directly connected with Europeana.

It is clear that the infrastructure offered by Natural Europe needs to satisfy strong requirements for metadata management, while establishing interoperability with learning applications, cultural heritage and biodiversity repositories. Towards this end, the Natural Europe project offers appropriate tools and services that allow the participating NHMs to: (*a*) uniformly describe and semantically annotate their content according to international standards and specifications, (*b*) interconnect their digital libraries, and (*c*) expose their Cultural Heritage Object (CHO) metadata records to Europeana.eu and BioCASE/GBIF.

The Biological Collection Access Service for Europe (BioCASE) [10] is a transnational network of biological collections of all kinds, while the Global Biodiversity Information Facility (GBIF) [6] is an open infrastructure which provides a single point of access to global biodiversity data.

Lab. of Distributed Multimedia Information Systems & Applications (TUC/MUSIC), Technical University of Crete, Greece
E-mail: {skevakis, makris, vkalokyri, xenia, stavros}@ced.tuc.gr

This article presents the Natural Europe Cultural Environment, i.e., the infrastructure and toolset deployed on each NHM allowing their curators to publish, semantically describe, manage and disseminate the CHOs that they contribute to the project. Additionally, we discuss the methodology followed for the transition of the infrastructure to the Semantic Web and the publishing of NHMs' cultural heritage metadata as Linked Data, supporting the Europeana Data Model (EDM) [4].

2 The Natural Europe Cultural Environment (NECE)

The Natural Europe Cultural Environment (NECE) [16] is a node in the cultural perspective of the Natural Europe project architecture [17]. It refers to the toolset deployed at each participating NHM, consisting of the Multimedia Authoring Tool (MMAT), the CHO Repository and the Vocabulary Server, facilitating the complete metadata management lifecycle: *ingestion*, *maintenance*, *curation*, and *dissemination* of CHO metadata. NECE also specifies how legacy metadata are migrated into Natural Europe. Figure 1 presents the architecture of Natural Europe with a focus on the Natural Europe Cultural Environment.

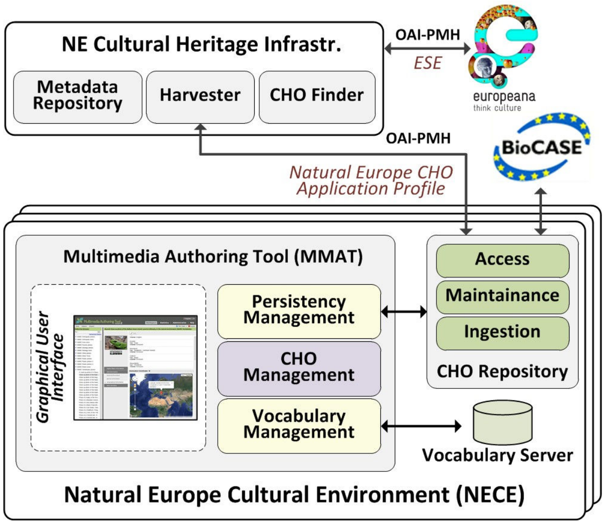


Fig. 1 The Natural Europe Architecture.

In the Natural Europe context, the participating NHMs provide metadata descriptions about a large number of Natural History related CHOs. These descriptions are semantically enriched with Natural Europe shared knowledge (vocabularies, taxonomies, etc.) using project provided annotation tools and services. The enhanced metadata are aggregated by the project, harvested by Europeana (to become available through its portal) and exploited for educational purposes. Furthermore, they are exposed to the BioCASE/GBIF networks, contributing their high quality content to biodiversity communities.

The following sections present the Natural Europe CHO Application Profile, as well as the architectural components of NECE (i.e., Multimedia Authoring Tool, CHO Repository and Vocabulary Server), focusing on their internal functionality.

2.1 The Natural Europe CHO Application Profile

The Natural Europe CHO Application Profile is a superset of the Europeana Semantic Elements (ESE) [5] metadata format. It has been developed through an iterative process involving the NHMs' domain experts and the technical partners of the project, driven by the needs and requirements of the stakeholders and the application domain of the project. The Natural Europe CHO Application Profile describes 3 main element categories for each CHO.

The *Cultural Heritage Object (CHO)* metadata category provides information about the analog resource or born digital object. It is composed of the following sub-categories: (*a*) the *Basic information*, dealing with descriptive information about the Cultural Heritage Object, (*b*) the *Species information* is applicable to describe information related to the species of a described specimen (animals, plants, minerals, rocks, etc.), and (*c*) the *Geographical information* contains metadata about the location in which a specimen has been collected.

The *Digital Object* metadata category provides information about a digital or digitized resource. It contains the following sub-categories: (*a*) the *Basic information* deals with general descriptive information about a digital or digitized resource, (*b*) the *Content information* holds the physical characteristics and technical information exclusive to a digital or digitized resource, and (*c*) the *Rights information* describes the intellectual property rights and the accessibility to a digital or digitized resource.

The *Meta-metadata* category provides metadata information for a CHO record. These include the creator of the record, the languages that appear in the metadata, etc. Additionally, it describes the history of the record during its evolution in the MMAT, including the operations and entities that affected it.

2.2 The MultiMedia Authoring Tool (MMAT)

The Multimedia Authoring Tool (MMAT) [1] is the first step towards allowing the connection of digital collections with

[1] A demo version of MMAT is available at: http://natural-europe.tuc.gr/mmat

Europeana and BioCASE/GBIF. It is a multilingual web-based management system for museums, archives and digital collections, which facilitates the authoring and metadata enrichment of cultural heritage objects. MMAT establishes interoperability between NHMs, cultural heritage and biodiversity networks. Moreover, it supports seamless ingestion of legacy metadata.

MMAT supports a rich metadata element set, the Natural Europe CHO Application Profile, as well as a variety of the most popular multimedia formats. Its main features include the publication of multimedia objects, the semantic linkage of the described objects with well-established controlled vocabularies, and the real-time collaboration among end-users with concurrency control mechanisms. Additionally, it provides the means to directly import the museums' legacy metadata for further enrichment and supports various types of users with different access rights. The three types of user that are currently supported are: (*a*) the administrators, able to manage the user accounts and the application, (*b*) the curators, administering CHO records/collections, and (*c*) the simple users, allowed only to inspect the data.

MMAT has been built as a Rich Internet Application, offering engaging experience and increased productivity. It adopts the Google Web Toolkit (GWT) [7] technology that enables web applications to perform part of their business logic into the client side and part on the server side. The client side refers to the business logic operations performed within a web browser running on a user's local computer, while the server side refers to the operations performed by a web server running on a remote machine. The overall architecture of MMAT is presented in Fig. 2.

The **Client Side** is responsible for the interaction with the user, the presentation of the information as well as the communication with the server when needed. It follows the Model-View-Presenter (MVP) [20] design pattern, separating the responsibilities for the visual display and the event handling behavior into distinct entities. Moreover, it accommodates modules for managing and transferring content and metadata between the Client and Server Side. The main modules on the Client Side are described below.

- The *Client Browser GUI* refers to the Graphical User Interface presented to the user's web browser. It consists of a composite widget set, each of which aggregates multiple simple widgets (e.g., tables, labels, buttons, textboxes, menus etc.) and serving a specific purpose. Two screenshots of the graphical user interface are presented in Fig. 3.
- The *View* modules control the composite widgets that form the User Interface and are responsible for their layout. Each view receives the user action events and dispatches them to its corresponding presenter for further processing.
- The *Presenter* modules are responsible for controlling Views and handling user actions (e.g., user clicks). They communicate with the Service Layer on the Server Side through the Application Manager.
- The *Event Bus implements* the Publish-Subscribe pattern, enabling the decoupling of the user interface components. It is responsible for the message transmission between the entities that reside on the Client Side. It provides mechanisms for publishing events and subscribing to events.
- The *Application Manager* implements the Mediator pattern, handling all the communication between the Presenters and the Server Side. Moreover, it acts as a centralized point of control for the Client Side, enabling its use for caching purposes.
- The *Multilinguality Support* module handles the translation of the user interface elements. While the tool loads on the client's browser, the translation corresponding to the user language preferences is transferred along with the user interface components. It is worth to mention that the graphical user interface has been translated in 7 languages (English, Greek, German, Portuguese, Estonian, Hungarian and Finnish), while it can be extended to support any additional languages with minimum effort.

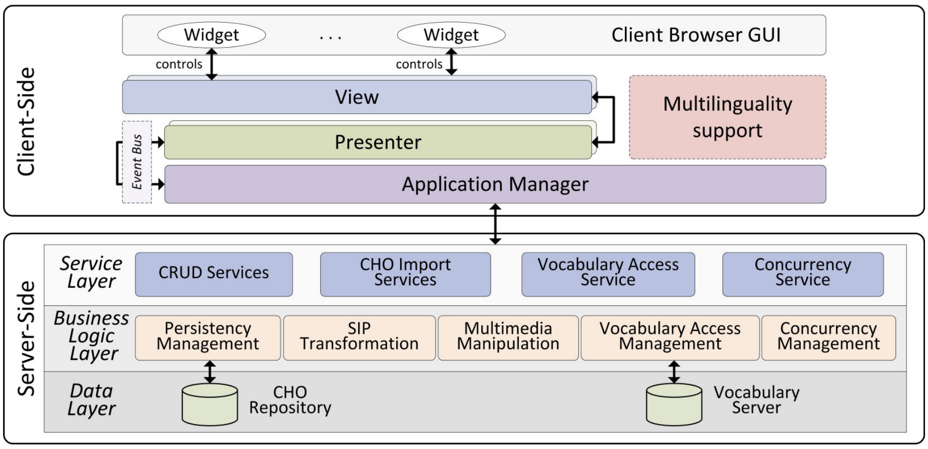
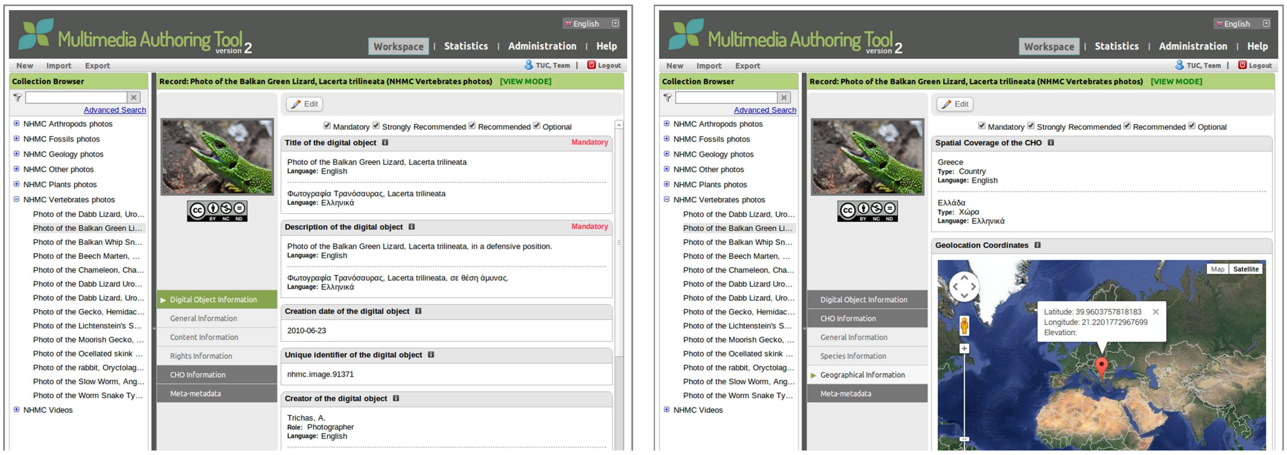


Fig. 2 The Multimedia Authoring Tool Architecture.


English
Multimedia Authoring Tool version 2
Workspace | Statistics | Administration | Help
New Import Export
TUC, Team | Logout
Collection Browser
Advanced Search
NHMC Arthropods photos
NHMC Fossils photos
NHMC Geology photos
NHMC Other photos
NHMC Plants photos
NHMC Vertebrates photos
Photo of the Dabb Lizard, Uro...
Photo of the Balkan Green Li...
Photo of the Balkan Whip Sn...
Photo of the Beech Marten, ...
Photo of the Chameleon, Cha...
Photo of the Dabb Lizard Uro...
Photo of the Dabb Lizard, Uro...
Photo of the Gecko, Hemidac...
Photo of the Lichtenstein's S...
Photo of the Moorish Gecko, ...
Photo of the Ocellated skink ...
Photo of the rabbit, Oryctolag...
Photo of the Slow Worm, Ang...
Photo of the Worm Snake Ty...
NHMC Videos
Edit
Digital Object Information
General Information
Content Information
Rights Information
CHO Information
Species Information
Geographical Information
Meta-metadata
Map
Satellite

Record: Photo of the Balkan Green Lizard, Lacerta trilineata (NHMC Vertebrates photos) [VIEW MODE]

☑ Mandatory ☑ Strongly Recommended ☑ Recommended ☑ Optional

Title of the digital object — Mandatory

Photo of the Balkan Green Lizard, Lacerta trilineata
Language: English

Φωτογραφία Τρανόσαυρας, Lacerta trilineata
Language: Ελληνικά

Description of the digital object — Mandatory

Photo of the Balkan Green Lizard, Lacerta trilineata, in a defensive position.
Language: English

Φωτογραφία Τρανόσαυρας, Lacerta trilineata, σε θέση άμυνας.
Language: Ελληνικά

Creation date of the digital object

2010-06-23

Unique identifier of the digital object

nhmc.image.91371

Creator of the digital object

Trichas, A.
Role: Photographer
Language: English

Spatial Coverage of the CHO

Greece
Type: Country
Language: English

Ελλάδα
Type: Χώρα
Language: Ελληνικά

Geolocation Coordinates

Latitude: 39.9603757818183
Longitude: 21.2201772967699
Elevation:

Fig. 3 Screenshots presenting the graphical user interface of MMAT.

The **Server Side** of MMAT follows a multi-layered architecture consisting of the following layers:

- The *Service Layer* controls the communication between the client and server logic by exposing a set of services to the client side components. These services comprise the middleware concealing the application's business logic. The basic system services are: (*a*) the CRUD Service, facilitating the creation, retrieval, update and deletion of a CHO, a CHO record/collection, a user etc., (*b*) the CHO Import Service, supporting the ingestion of XML metadata records to the CHO Repository through the Persistency Management module, (*c*) the Vocabulary Access Service, enabling the access to taxonomic terms, vocabularies, publicly sourced authority files of persons, places, etc., through the Vocabulary Access Management module, and (*d*) the Concurrency Service, providing the basic methods for acquiring/releasing/refreshing locks on a CHO record/collection.
- The *Business Logic Layer* contains the business logic of the application and separates it from the Data Layer and the Service Layer. It consists of five basic modules: (a) the Persistency Management module, managing the submission/retrieval of information packages to/from the CHO Repository, (*b*) the SIP Transformation Module, transforming XML metadata records to Submission Information Packages (SIPs), (*c*) the Multimedia Manipulation Module, creating thumbnails and extracting metadata from media files used for the creation and enrichment of CHO records, (*d*) the Vocabulary Access Management Module, providing access to indexed vocabularies and authority files residing on the Vocabulary Server, and (*e*) the Concurrency Management Module, applying a pessimistic locking strategy to CHO record/collection metadata in order to overcome problems related to the concurrent editing by multiple users.
- The *Data Layer* accommodates external systems that are used for persistent data storage. Such systems are the CHO Repository and the Vocabulary Server.

2.3 The CHO Repository

The CHO Repository handles both content and metadata and adopts the OAIS Reference Model [9] for the ingestion, maintenance and dissemination of Information Packages (IPs). To this end, it accommodates modules for the ingestion, archival, indexing, and accessing of CHOs, CHO records/collections etc. This functionality refers to a complete information preservation lifecycle, where the producer is the MMAT and the consumers are the MMAT, the harvester application of the Natural Europe federal node and the BioCASE/GBIF networks. Figure 4 presents the overall architecture of the CHO Repository with emphasis to its internal software modules.

The **Ingestion Module** is responsible for the ingestion of an information package (i.e., CHOs, CHO records, CHO collections, and user information) in order to store it as a new Archival Information Package (AIP) to the repository, or to update/delete an already existing AIP. Any submitted information package should be validated and processed in order to identify and create the required AIPs that should be transferred for archival. The only actor on this module is the MMAT, which serves as a SIP producer.

The **Archival Module** receives AIPs from the Ingestion Module for storage purposes, as well as AIP retrieval requests from the Access Module for dissemination purposes. In order to support storage and retrieval operations, it employs a DB Storage/Retrieval Manager component which is implemented in a flexible way for supporting any DBMS (relational, XML). A dedicated eXist DB Storage/Retrieval Manager has been implemented, supporting database specific storage and retrieval operations in an eXist XML DB

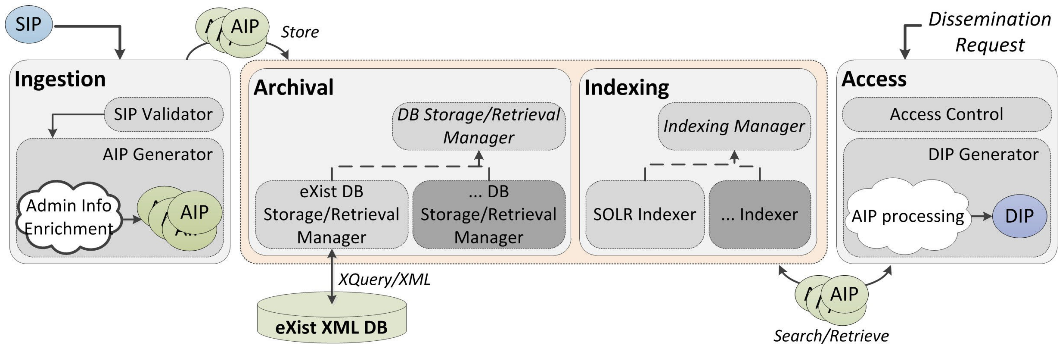


Fig. 4 The CHO Repository Architecture.

instance, using XQuery/XML. After the storage, update, or deletion of an AIP, the Archival Module notifies the Indexing Module of the changes. Furthermore, an advanced logging mechanism has been implemented, keeping track of any actions/changes (full history) on the CHO records or collections, while ensuring data authenticity and chain of custody. This mechanism allows the restoration of any CHO record or collection to previous states when required.

The **Indexing Module** receives AIPs from the Archival Module in order to build and maintain AIP index structures, as well as AIP retrieval requests from the Access Module for dissemination purposes. In order to support both the maintenance and retrieval index operations, it employs an Indexing Manager component which is flexibly implemented to support any search platform. Currently, a dedicated Apache SOLR Indexer component has been implemented, supporting platform specific maintenance and retrieval operations.

The **Access Module** provides services allowing Dissemination Information Package (DIP) consumers (i.e., MMAT, the harvester application of the Natural Europe federal node, the BioCASE/GBIF networks and other external applications) to request and receive information stored in the CHO Repository. It provides functionality for receiving information access requests, while applying access control policies through the Access Control component. Furthermore, it exploits any available indices maintained by the Indexing module in order to retrieve the requested AIPs. The AIPs retrieved from the Archival and/or Indexing Modules are transferred to the DIP Generator component so as to be further processed for creating the final DIP that will be delivered to the DIP consumer. Finally, the module exposes the following services: (*a*) the OAI-PMH interface, supporting the selective harvesting of the contributed CHO metadata by the Natural Europe federal node and subsequently by Europeana, (*b*) the BioCASE protocol interface, establishing the connection to the BioCASE/GBIF networks, and (*c*) the OpenSearch endpoint, enabling the search of CHO metadata in a standard and accessible format.

The supported response formats for metadata harvesting through the OAI-PMH interface are DC and the Natural Europe CHO Application Profile.

2.4 The Vocabulary Server

The Vocabulary Server manages the vocabularies and authority files used during the semantic annotation process. Authority files refer to information about organizations, persons and places, while vocabularies refer to the taxonomies used for the enrichment of the CHO metadata. The Vocabulary Server is also responsible for the indexing and retrieval of authority and taxonomic information, allowing us to provide fast auto-complete functionality to MMAT end users. This saves time from the curation process, increasing the user productivity and providing error prevention during semantic annotation. For this purpose, the server employs a Lucene/Solr instance, managing the indexing and querying of data.

The semantic annotation of resources is a strong requirement for any system supporting metadata editing within a museum. Apart from the use of controlled vocabularies during the annotation process, it provides great cross-institution interoperability. To this end, the Vocabulary Server has been developed to support any taxonomic classification that the museums might use. This is achieved through the ingestion of taxonomies represented in SKOS format.

Simple Knowledge Organization System (SKOS) [18] is the leading format for the representation of thesauri, classification schemes, taxonomies, or any other types of controlled vocabularies. It is based on the Semantic Web principles and therefore enables the smooth transition of data to the RDF format. The exploitation of well-established vocabularies in SKOS format during the annotation process, provides a solid basis for the production of linked data, and subsequently an additional dissemination channel towards the Linked Data communities.

A vocabulary that was extensively used in the context of Natural Europe was the Catalogue of Life (CoL), which is the most comprehensive catalogue of living species, containing over 1.4 million species along with their relationships. It is widely used in biological classification and serves as an important point of reference for many institutions, including Natural History Museums. CoL offers a web-based system for browsing the species taxonomy, as well as a set

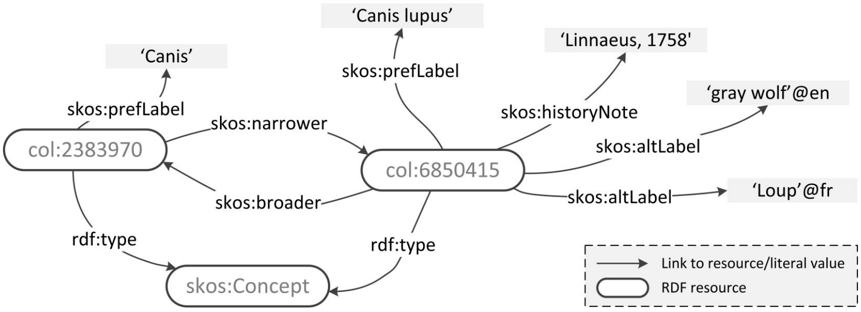


Fig. 5 An example of the Catalogue of Life SKOSified data in the form of a graph.

of web services for searching. However, CoL lacks support for persistent URIs able to be referenced by external applications, as well as SKOS representation of its data. Towards this end, we worked on a method of exposing the taxonomy of CoL to SKOS, using the CoL annual checklist and a D2R Server [12]. The features of the SKOS model that we employed are: (*a*) the class Concept, and (*b*) the properties broader, narrower, prefLabel and altLabel.

The first step in the process was the representation of all the taxonomy nodes as Concepts. The scientific name of each node was transformed into a prefLabel, and the common names into altLabels. Finally, the hierarchy of the taxonomy was retained by connecting the parent and children nodes with the properties broader and narrower. An example of the CoL SKOSified data in the form of a graph is shown in Fig. 5.

3 The metadata management life-cycle

The complete lifecycle that NECE defines for the metadata management comprises four phases: (*a*) pre-ingestion, (*b*) ingestion, (*c*) maintenance, and (*d*) dissemination.

During the **pre-ingestion phase (preparatory phase)** each NHM selects the CHO records/collections that will be contributed to the project and ensures that they will be appropriately migrated into Natural Europe. This includes the web publishing of CHOs along with their respective thumbnails, and the metadata unification of existing CHO metadata. Publishing of CHOs refers to the uploading of CHO descriptions on the museum's website, or simply to the uploading of digital object thumbnails to a web server. The most important part of this step is the acquisition of a persistent URI for each resource. The web publishing of media files, the creation of thumbnails and the assignment of persistent URIs can be undertaken by MMAT. On the other hand, the metadata unification of existing CHO metadata is performed by preparing XML records conforming to the Natural Europe CHO Application Profile. This step can be easily carried out by any well-known legacy database system and even from Excel documents.

During the **ingestion phase** any existing CHOs and CHO metadata are imported to the Natural Europe environment. The latter are further enriched through a semantic annotation process. MMAT provides functionality for loading metadata conforming to the Natural Europe CHO Application Profile, as well as CHOs into its underlying repository. Afterwards, museum curators have the ability to inspect, modify, or reject the imported CHO descriptions. As far as the ingestion through the normal metadata curation/annotation activity is concerned, MMAT allows museum curators to maintain (create/view/modify/enrich) CHO metadata. This is facilitated by the access and concurrency control mechanisms, ensuring security, integrity, and consistency of the content.

The **maintenance phase** refers to the storage and management of CHOs and CHO metadata using MMAT and the CHO Repository. It addresses policies related to the integrity, authenticity and chain of custody. The integrity of data is guaranteed by performing both full and incremental database backups on a weekly and daily basis respectively. These backups are persisted on remote machines and provide the means to overcome any failure on the servers hosting the systems with minimum loss. Concerning the authenticity and chain of custody, both are controlled by the system's logging mechanism keeping track of any actions/ changes (full history) on the CHO records/collections, along with information about the user that performed each action/ change. This enables the rollback to any previous state of the CHO record/collection when required.

The **dissemination phase** refers to the controlled provision of the maintained metadata to third party systems and client applications. Such systems are the Natural Europe federal node and the BioCASE/GBIF networks. Metadata dissemination is mainly performed by the Access Module of the CHO Repository, which provides functionality for receiving information access requests and replying in several response formats, while applying various access control policies. It exposes (*a*) an OAIPMH interface for the selective harvesting of the CHO metadata, (*b*) a service interface implementing the BioCASE protocol, and (*c*) an OpenSearch endpoint. The OAI-PMH interface supports metadata dissemination in DC and Natural Europe CHO Application Profile format.

4 Connection of the Natural Europe Cultural Environment with BioCASE/GBIF

The Biological Collection Access Service for Europe (BioCASE) [10] is a transnational network of biological collections of all kinds. It enables widespread unified access to distributed and heterogeneous European collections and observational databases using open-source, system independent software and open data standards/protocols. Moreover, the Global Biodiversity Information Facility (GBIF) [6] is an open infrastructure which provides a single point of access to global (world-wide) biodiversity data.

In order for data providers to connect to BioCASE, they have to install the BioCASE Provider Software. This software offers an XML data binding middleware for publishing data residing in relational databases to BioCASE. The information is accessible as a web service and retrieved through BioCASE protocol requests. The BioCASE protocol [2] is based on the ABCD Schema [11], which is the standard for access and exchange of data about specimens and observations. Furthermore, this protocol is supported (among others) by GBIF for accessing data from its providers.

Figure 6 presents an overview of the BioCASE architecture. On the top left resides the BioCASE portal, backed up by a central cache database, accessing information from the data providers (bottom). The BioCASE Provider Software (wrapper) is attached on top of each provider's database, enabling communication with the BioCASE portal and other external systems (e.g., GBIF). This wrapper is able to analyze BioCASE protocol requests and transform them to SQL queries using some predefined mappings between ABCD concepts and table columns. The SQL queries are executed over the underlying database and the results are delivered to the client after being transformed to an ABCD document.

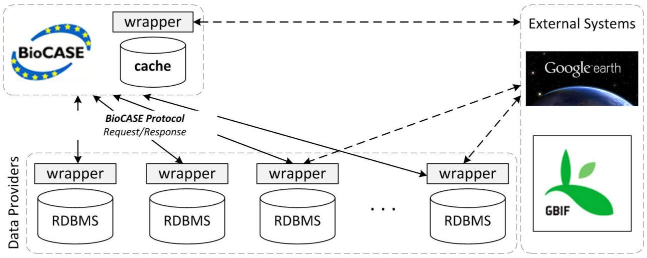


Fig. 6 The BioCASE Architecture.

Although BioCASE supports a variety of RDBMSs, it does not support non-SQL databases (e.g., XML DBMSs). This is also the case of MMAT, which is backed by an eXist XML Database. To address this problem, we have built and installed a customized wrapper to the Access Module of the data providers' CHO Repositories (Fig. 7). The wrapper is able to analyze BioCASE protocol requests and transform them to XQueries, exploiting mappings between the Data Provider's schema ABCD. Towards this end, a draft mapping of the Natural Europe CHO Application Profile to ABCD was produced based on BioCASE practices. The XQueries are executed over the providers' repositories and the results are delivered to the client after being transformed to an ABCD document.

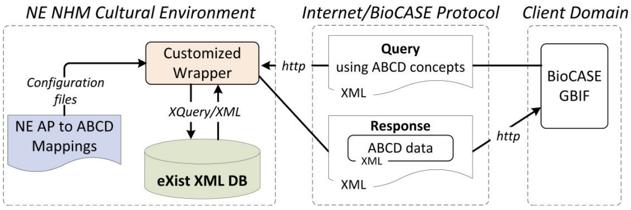


Fig. 7 Connecting Natural Europe Cultural Environment with BioCASE/GBIF.

Although, the wrapper has been implemented for the XML databases of Natural Europe, it is able to support any underlying database either relational or XML with minimum effort. To this end, it follows a modular multi-tier architecture consisting of the following layers:

- The Service Layer controls the communication between the data provider and the BioCASE Portal by implementing the BioCASE protocol. It exposes services that comply to the BioCASE protocol specification, while concealing the wrapper's business logic. The basic system services are: (*a*) the Search Service, enabling complex query execution, based on ABCD concepts, over a data provider's database, (*b*) the Scan Service, supporting the retrieval of unique values for a given ABCD concept, and (*c*) the Capabilities Service, providing useful information about the ABCD concepts that can be used for searching in a data provider's database.
- The Business Logic Layer consists of three basic modules: (*a*) the Query Deserialization Module, handling the deserialization of the submitted queries to database-specific format, (*b*) the Mapping Management Module, administering the mappings between ABCD (used by the BioCASE Protocol) and the data provider's schema, and (*c*) the Results Serialization Module, managing the transformation of query results to an ABCD document, utilizing the mappings provided by the Mapping Management Module.
- The Data Layer provides simplified access to the data stored in the persistent storage.

Changes in the persistent storage can be easily supported by providing a new implementation of the Query Deserialization Module, based on the query language supported by the new persistent storage and the underlying data structure.

[2] http://www.biocase.org/products/protocols/

To this end, the module has been designed using the principles of the plugin pattern and is therefore able to automatically recognize new module implementations. On the other hand, changes in the mappings between the ABCD and the data provider's schema can be addressed by modifying the wrapper's configuration files.

The source code of our implementation has been contributed to BioCASE and our approach has been successfully tested by their technical staff. Until the actual connection of the Natural Europe federated nodes (data providers) to the BioCASE/GBIF networks is established, we have deployed a local BioCASE portal installation able to retrieve CHOs from all federated node CHO Repositories [3].

5 Transition to the Semantic Web

The Semantic Web standards and best practices provide a basis on which interoperable Web systems can be built in a well dened manner. W3C recommendations like RDF(S), SKOS, SPARQL, and OWL are considered as corner-stones for cross-domain and domain-independent interoperability. The use of these standards and practices enables: (*a*) semantically richer content, (*b*) reusability of existing common ontologies, taxonomies and published datasets, (*c*) answering of highly structured and distributed queries, (*d*) creation of large open data repositories, and (*e*) inferencing of new knowledge by performing reasoning. In the Cultural Heritage domain it enables the creation of large national and international Cultural Heritage portals, like Europeana, as well as the massive publications of linked library data [14].

Driven by the motive to expose the Natural Europe content as semantically rich Linked Data and therefore benefit from the aforementioned advantages, we developed infrastructures for the Semantic Web presence of the participating NHMs [22]. Our aim was to develop a semantically rich cultural heritage infrastructure for NHMs, providing a Semantic Web perspective to the Natural Europe cultural content in terms of: (*a*) creating the Natural Europe Ontology in order to introduce semantics to the current Natural Europe Schema for inferring new knowledge, (*b*) using the RDF data model to publish the Natural Europe data on the Web, (*c*) linking the Natural Europe's cultural content to external commonly used vocabularies, thesaurus and published datasets, (*d*) enabling data retrieval through SPARQL, and (*e*) supporting interoperability with the Europeana Semantic Layer by offering the appropriate EDM dissemination mechanisms.

In the following sections we introduce the Natural Europe Ontology, we continue with the presentation of the semantic infrastructure for the transition of the Natural Europe Cultural Digital Libraries Federation to the Semantic Web, and finally, we describe the methodology for ingesting and converting the NHMs' cultural heritage metadata to Linked Data supporting the Europeana Data Model.

5.1 The Natural Europe Ontology

For any rich cultural heritage infrastructure aiming to provide a Semantic Web perspective to its underlying content, it is not sufficient to use a flat schema or a schema providing weak semantics. To this end, we introduced the Natural Europe Ontology [4], exploiting class and property axioms for establishing powerful semantics and enabling the inferencing of new knowledge out of existing data. The Natural Europe Ontology has been described in OWL and it is based on the Natural Europe CHO Application Profile. Notions such as CHO, CHO collection, specimen, observation, multimedia object, person, and organization have been described as OWL classes, while the underlying attributes have been expressed as object/datatype properties. As a result, the contributed flat Natural Europe records can be organized in aggregations of different kinds of objects, e.g., a specimen may be described by multiple observations and an observation may contain multiple multimedia objects. Furthermore, the Natural Europe Ontology references well-known ontologies/schemas (e.g., DC, FOAF, Geonames, SKOS) and has been aligned with Europeana Data Model (EDM), supporting interoperability with the Europeana Semantic Layer.

5.2 Architecture & methodology

In order to achieve the objectives set for the Semantic Web presence of the NHMs, the modules of the federated instances (NECE) and the federal node (NECHI) of the Natural Europe Cultural Federation have been enhanced with software components supporting the Semantic Web technologies, comprising the Semantic Layer of NECE.

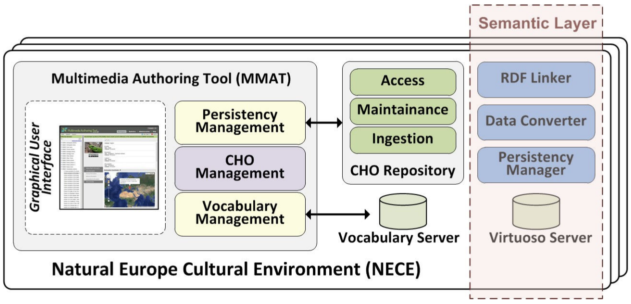


Fig. 8 The NECE Semantic infrastructure.

NECE Semantic Layer (Fig. 8) accommodates modules that enable the periodic batch conversion of CHO Reposi-

[3] http://natural-europe.tuc.gr/biocase

[4] http://natural-europe.tuc.gr/ne-ontology-v01.owl

tory XML data to RDF and support the presence of NHMs' to the Linked Data Cloud. Its main architectural components are:

- The *RDF Linker*, supporting the linkage of CHO Repository data to external well-established RDF datasets (i.e., Geonames, DBpedia, CoL and Uniprot).
- The *Data Converter*, managing the conversion of the CHO Repository's XML data to RDF, taking into account the Natural Europe Ontology. It uses the services provided by the Access Module of the CHO Repository for accessing the NHM data, as well as the RDF Linker functionality for linking the retrieved NHM data to external RDF datasets.
- The *Persistency Manager*, controlling the maintenance of the converted RDF data. The module employs a Virtuoso Open-Source Server supporting the persistence, publication and dissemination of RDF data. In more detail, it provides services for: (*a*) publishing RDF data by providing resolvable URIs, (*b*) persisting data in a native RDF store, (*c*) enabling data access through a SPARQL endpoint, and (*d*) browsing data through a GUI supporting facets on certain data types.

The methodology we followed for the transition of the Natural Europe Cultural Federation to the Semantic Web includes the following stages, supported by the above software components: (*a*) linkage of Natural Europe CHO metadata to established RDF datasets and vocabularies, (*b*) conversion of metadata from XML to RDF, (*c*) archival, publishing and dissemination of the converted RDF data. These stages are described in the following sections.

5.2.1 Establishing links to external RDF datasets

The linking of data to other well-known RDF datasets and vocabularies is a very crucial step in the production of rich Linked Data. The RDF datasets used in the context of Natural Europe include: (*a*) *Geonames*, a geographical database containing over 10 million geographical names, (*b*) *DBpedia*, a knowledge base describing more than 3.64 million things, including persons, places, and species, (*c*) *CoL*, a comprehensive catalogue of all known species of organisms on Earth, compiled by 99 taxonomic databases, (*d*) *Uniprot*, a high-quality database providing (among others) information on protein sequence and taxonomic classification.

More specifically, the CHO spatial information is used to discover links to Geonames, while the CHO species information (e.g., scientific names) are exploited for generating references to CoL and Uniprot. Furthermore, the CHO titles, descriptions and keywords are utilized for the discovery of links to DBpedia. The linkage of the CHOs to the aforementioned datasets is dynamic and performed by exploiting the services that they officially provide. An exception to this is the CoL, which is currently not published in RDF format. This issue is addressed by the Vocabulary Server, persisting the CoL SKOSified version as described in Sect. 2.4.

5.2.2 Conversion of metadata from XML to RDF

The basic operations that need to be performed for the conversion of XML data to RDF data complying with an already specified ontology include: (*a*) the creation of class instances for every resource, as well as (*b*) the use of object/datatype properties (with respect to the Natural Europe Ontology) for describing resource attributes.

For every newly identified resource, unique identifiers have to be assigned in order to enable the exploitation of the ontology's semantic capabilities in full. An example of the Natural Europe RDF data in the form of a graph is shown in Figure 9.

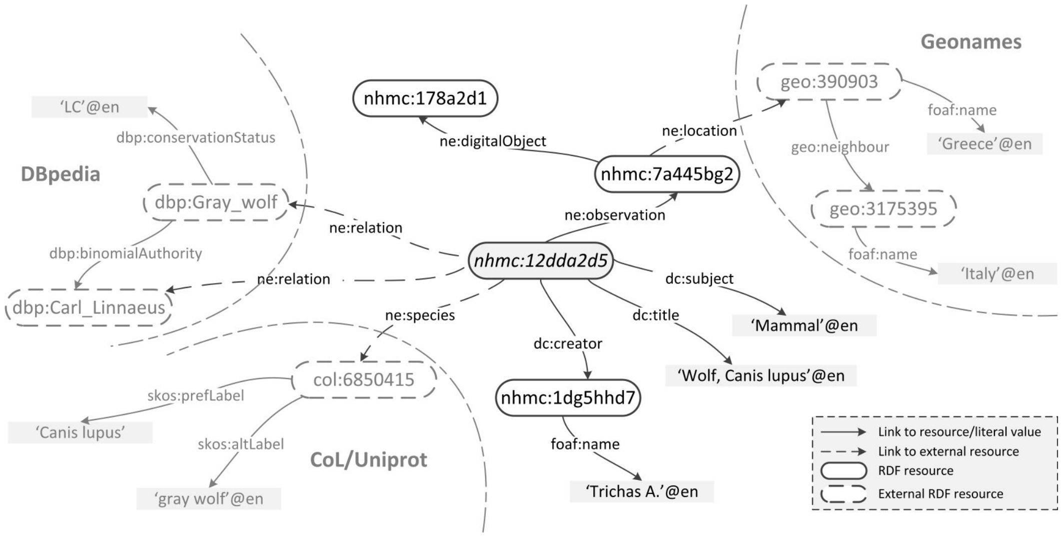


Fig. 9 An example of the Natural Europe RDF data in the form of a graph.

5.2.3 Maintenance, publishing and dissemination of the RDF data

The maintenance of RDF data includes both the archiving and reasoning. The archiving of RDF data is generally supported by the use of RDF stores, while reasoning is performed by applying inferencing techniques on the persisted data.

The publishing and dissemination of RDF data is based on providing resolvable URIs to RDF resources, as well as SPARQL endpoints in order to enable data access and contribute a new node to the Linked Data Cloud. The assignment of resolvable URIs allows any data provider to reference and connect his data to the new node, based on the Linked Data paradigm.

Such an infrastructure allows the execution of highly expressive queries combining knowledge from distributed data sources with the use SPARQL. In the context of Natural Europe, the resulted RDF dataset can be used to answer questions like: “Find photos of endangered species of Genus ’Bufo’ in neighbor countries of Greece”. This query combines information from: (*a*) *Natural Europe*, containing specimen information including photos, (*b*) *DBpedia*, providing species conservation status, (*c*) *CoL/Uniprot*, providing the classification of Genus ’Bufo’, and (*d*) *Geonames*, delivering geographic information regarding neighbor countries of Greece.

6 Deployment, use and evaluation

In order to facilitate the deployment of the MMAT, the CHO Repository and the Vocabulary Server in any museum, we have compiled a packaged version of the whole infrastructure which can be hosted in any web server. This allows for rapid deployment of the tools by less experienced people. Moreover, all the components have been built as separate modules, which means that in the case of a new version they can be updated individually.

The infrastructure has been already deployed in the six Natural History Museums participating in the project, allowing the curators to publish, semantically describe, manage and disseminate a large volume of CHOs. The participating museums are: (a) Natural History Museum of Crete (NHMC), (b) National Museum of Natural History - University of Lisbon (MNHNL), (c) Jura-Museum Eichstätt (JME), (d) Arctic Center (AC), (e) Hungarian Natural History Museum (HNHM), and (f) Estonian Museum of Natural History (TNHM). Table 1 presents the number of CHOs that have already been published by each NHM using MMAT. Additionally, it holds the total number of triples generated from the metadata of each NHM. These results do not include the triples created by applying any reasoning technique in the RDF data. From an initial dataset of 15,050 CHO records we got 631,220 RDF triples.

Improvements on the user-interface have been made after continuous feedback from museum partners in a number of tool releases. Heuristic evaluation of the MMAT was performed, while extensive usability studies have been performed in a number of curator workshops organized by the participating NHMs.

6.1 Heuristic evaluation

The heuristic evaluation of the Multimedia Authoring Tool was performed by a team of inspectors comprised of 5 current Masters in Human Computer Interaction (HCI) graduates with background and experience in fields such as Computer Science and Information Technology in the context of the HCI course of the Electronic and Computer Engineering Dept. of the Technical University of Crete. In this course, the students had to perform usability evaluation on several products including MMAT. The evaluation was based on Jakob Nielsen’s heuristics [19]; 88 errors (9 major) were detected and fixed [5].

6.2 Curator workshops

A number of curator workshops were organized [21], attracting participants from different professions (presented in Table 2). During these workshops, the participants had the opportunity to interact with the Multimedia Authoring Tool in the means of publishing, annotating, searching and reviewing both their own CHOs and other existing ones.

Concerning the AC, JME, NHMC, TNHM and HNHM curator workshops, twenty out of thirty three curators reported that they had already described items from their collections using metadata. However, most of the participants had seldom or never used any tool for either uploading multimedia files from their museum collections or managing their museum digital collections. In addition, the exploitation of digital collections in education was new for the majority of curators. Regarding the MNHNL curator workshop, all participants had already worked with databases, while most of them occasionally search for or use digital resources from other NHMs (e.g., getting suggestions about metadata management or doing scientific research).

After interacting with MMAT, the participants of NHMC, JME, HNHM, AC and TNHM workshops were administered the satisfaction questionnaire. The results are presented in four parts (Fig. 10): (*a*) usability issues, (*b*) functionality regarding metadata, (*c*) functionality regarding profession, and (*d*) personal aspects.

[5] Results of the Heuristic Evaluation (in Greek): http://natural-europe.tuc.gr/mmat/heuristic

Table 1 The number of CHOs annotated by each NHM using MMAT, along with the number of the generated RDF triples.

Natural History Museums (NHMs)	**Cultutal Heritage Objects (CHOs)**						**RDF Data**
	Images	**Videos**	**Sounds**	**Texts**	**3D**	**TOTAL**	**Triples**
Arctic Center (**AC**)	480	0	0	0	0	**480**	18,715
Jura-Museum Eichstätt (**JME**)	1,214	42	115	287	0	**1,658**	60,371
Natural History Museum of Crete (**NHMC**)	3,840	13	0	157	0	**4,010**	195,905
National Museum of Natural History - University of Lisbon (**MNHNL**)	1,934	37	30	653	32	**2,686**	115,913
Estonian Museum of Natural History (**TNHM**)	1,736	100	0	136	0	**1,972**	85,773
Hungarian Natural History Museum (**HNHM**)	2,418	51	0	1,770	5	**4,244**	154,543
TOTAL	**11,622**	**243**	**145**	**3,003**	**37**	**15,050**	**631,220**

Table 2 Core data of curators participated in the workshops.

NHM	**Participants**	**Gender**		**Mean age**	**Profession**
		M	**F**		
AC	1	1	0	40	Curator
JME	1	1	0	28	Communication
NHMC	7	3	4	46	Curators, Librarian
MNHNL	7	5	2	41	Curators, zoological curator, biologist, Post Doc, Digital resource manager
TNHM	10	4	6	48	Curators
HNHM	14	5	9	45	Researchers, Curators, Librarian

- *Usability Issues:* MMAT was rated positively by the majority of the curators of the NHMC, JME, HNHM, AC and TNHM workshops. Twenty one of the participants found the MMAT easy to learn to operate, while only six identified the interaction with the system as not clear/ understandable.
- *Functionality regarding metadata:* In general, the use of metadata elements in MMAT has met the expectations of the curators; only three of the curators found the elements not sufficient for describing their collections' items.
- *Functionality regarding profession:* The functionality regarding the profession of curation has been rated as satisfying. Creation of CHO records/collections is sufficient, while exporting metadata, searching and reviewing CHO records were rated adequately.
- *Personal aspects:* The overall impression of the tool was positive. Most of the curators felt competent using the MMAT and secure in providing their personal information.

7 Related work

CollectiveAccess [2] is a web-based multilingual cataloguing tool for museums, archives and digital collections. It allows integration of external data sources and repositories for cataloguing and supports the most popular media formats. Although CollectiveAccess supports a variety of metadata standards (Dublin Core, PBCore and SPECTRUM, etc.), direct support for the ESE specification is not provided. Moreover, CollectiveAccess does not implement any harvesting protocol (e.g., OAI-PMH), making impossible to publish the content to Europeanas web portal. Finally, the current version of CollectiveAccess lacks any importing mechanism, crucial in the case of museums having already described their cultural content with metadata in legacy or internal (museum specific) formats.

Collection Space [1] is a web-based application for the description and management of museum collection information. Collection Space does not support the ESE specification and its metadata dissemination mechanisms are limited (REST-API). Moreover, it does not support any harvesting protocol.

Custodea [3] is a system mainly intended for historical and cultural institutions that need to deal with digitization. Custodea covers harvesting of digital content and representations, data transformation, creation and storage of metadata, vocabulary management, publishing and provision of data for Europeana and other institutions. However, the front-end application is desktop-based, complicating the collaboration of museum curators.

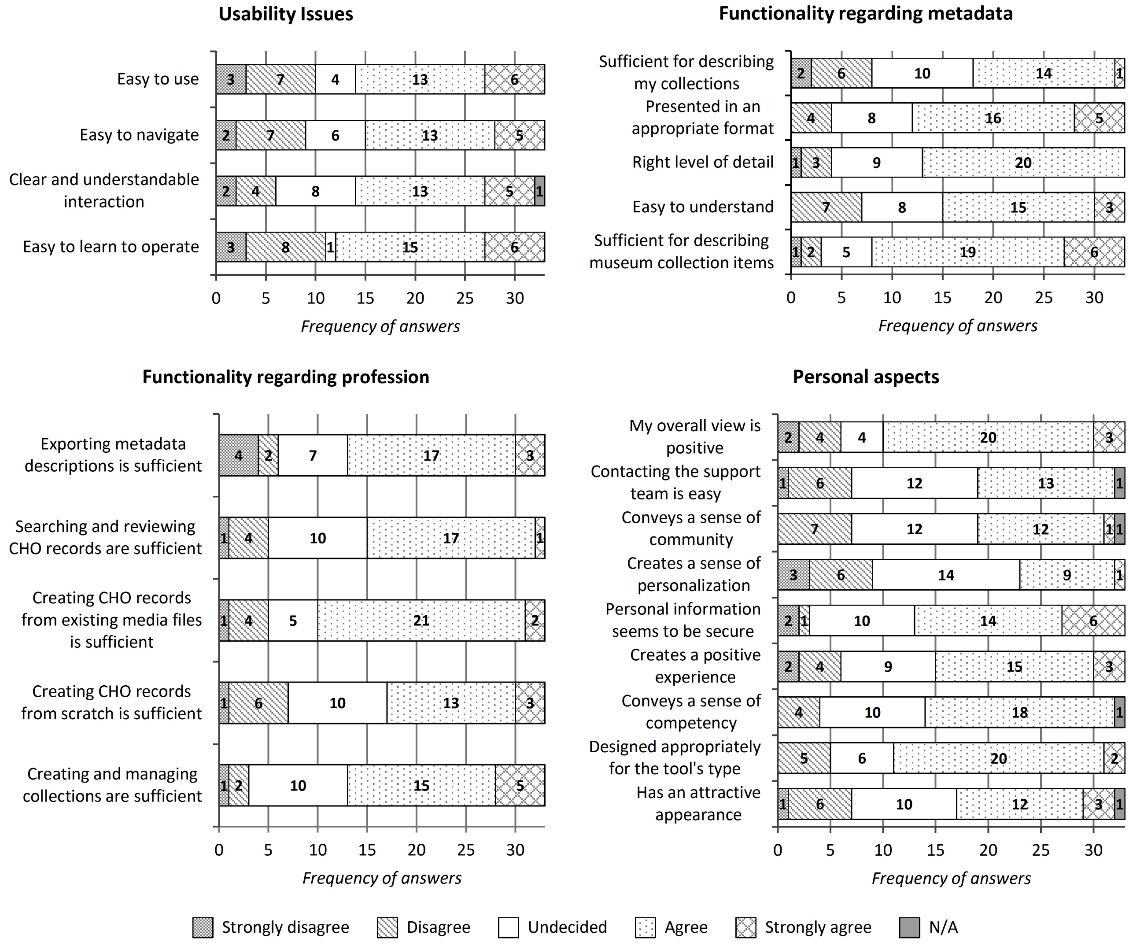


Fig. 10 Results of the satisfaction questionnaire regarding MMAT.

Finally, none of the above tools provides out-of-the-box support for connection to any biodiversity network (e.g., BioCASE, GBIF), or the means for publishing, maintaining and disseminating their underlying data in the RDF format.

8 Conclusion

We presented the architecture, deployment and evaluation of the infrastructure used in the Natural Europe project, allowing curators to publish, semantically describe, and manage the museums' CHOs, as well as disseminate them to Europeana and to BioCASE/GBIF networks. This infrastructure consists of the Multimedia Authoring Tool, the CHO Repository and the Vocabulary Server. It is currently used by six European NHMs participating in the Natural Europe project, providing positive feedback regarding the usability and functionality of the tools. Until today, a large number of CHOs has already been published. A long term vision of the project is to attract more NHMs to join this effort.

Moreover, we presented the infrastructure and methodology that we followed in order to publish the contributed CHO metadata of Natural Europe as Linked Data. Our approach comprised of the following stages: (*a*) linkage of Natural Europe CHO metadata to established RDF datasets and vocabularies, (*b*) conversion of metadata from XML to RDF, (*c*) archival, publishing and dissemination of the converted RDF data. Our methodology can be applied in other contexts as well, exploiting their schemes and domain specific vocabularies/datasets.

Regarding the semantic infrastructure, we currently investigate the integration of the Natural Europe NHM federated nodes with cultural heritage and biodiversity RDF data providers, utilizing different metadata schemas, in an ontology-based mediator system. Such an infrastructure is extremely important for Semantic Web applications and end users, since it will enable the retrieval of up-to-date triples, unlike the data warehousing approaches applied by data aggregators. To this end, the SPARQL-RW Framework [15], developed by TUC/MUSIC Lab, is considered as a cornerstone component for transparently accessing federated RDF data sources complying to different Ontology Schemas.

Finally, we are implementing MoM-NOCS [23], a Framework supporting the management and capturing of observations in real-time using mobile devices like smartphones. This will allow the on-site reporting of observations along with multimedia capturing.

Acknowledgements This work has been carried out in the scope of the Natural Europe Project (Grant Agreement 250579) funded by EU ICT Policy Support Programme.

References

1. CollectionSpace - Collections Management Software for Museums. http://www.collectionspace.org/
2. CollectiveAccess - Web-based software to catalogue, manage and publish museum and archival collections. http://www.collectiveaccess.org/
3. Custodea - Powerful solutions for cultural sector based on open source components. http://www.custodea.com/en/home/
4. Europeana Data Model Definition V.5.2.3. http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22
5. Europeana Semantic Elements Specification V.3.4.1. http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57
6. Global Biodiversity Information Facility. http://www.gbif.org/
7. GWT - Google Web Toolkit. http://www.gwtproject.org/
8. The Natural Europe Project. http://www.natural-europe.eu/
9. ISO 14721:2003 Open Archival Information System (OAIS) Reference Model. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683 (2003)
10. Berendsohn, W., Döring, M., Gebhardt, M., Güntsch, A.: BioCase - A Biological Collection Access Service for Europe. Tech. rep. (2002)
11. Berendsohn, W.G.: ABCD Schema - Task Group on Access to Biological Collection Data. Tech. rep. (2007)
12. Bizer, C., Cyganiak, R.: D2R Server - Publishing Relational Databases on the Semantic Web. In: Proceedings of the 5th International Semantic Web Conference (2006)
13. Falk, J., Storksdieck, M.: Using the contextual model of learning to understand visitor learning from a science center exhibition. Science Education **89**(5), 744–778 (2005)
14. Hendler, J., Ding, Y.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool Publishers series (2012)
15. Makris, K., Bikakis, N., Gioldasis, N., Christodoulakis, S.: SPARQL-RW: transparent query access over mapped RDF data sources. In: Proceedings of the 15th International Conference on Extending Database Technology, EDBT. ACM (2012)
16. Makris, K., Skevakis, G., Kalokyri, V., Arapi, P., Christodoulakis, S.: Metadata Management and Interoperability Support for Natural History Museums. In: Research and Advanced Technology for Digital Libraries, pp. 120–131. Springer (2013)
17. Makris, K., Skevakis, G., Kalokyri, V., Arapi, P., Christodoulakis, S., Stoitsis, J., Manolis, N., Leon Rojas, S.: Federating Natural History Museums in Natural Europe. In: Proceedings of 7th Metadata and Semantics Research Conference, MTSR (2013)
18. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/ (2009)
19. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: CHI, pp. 249–256. ACM (1990)
20. Potel, M.: MVP: Model-View-Presenter. The Taligent Programming Model for C++ and Java. Taligent Inc. (1996)
21. Sattler, S., Bogner, F.: D6.2 Integrated Pilot Evaluation Report. Natural Europe Project. Tech. rep. (2013)
22. Skevakis, G., Makris, K., Arapi, P., Christodoulakis, S.: Elevating Natural History Museums' Cultural Collections to the Linked Data Cloud. In: Proceedings of the 3rd International Workshop on Semantic Digital Archives (2013)
23. Tsinaraki, C., Skevakis, S., Trochatou, I., Christodoulakis, S.: MoM-NOCS: Management of Mobile Multimedia Nature Observations using Crowd Sourcing. In: Proceedings of the 11th International Conference on Advances in Mobile Computing & Multimedia, MoMM (2013)