Metadata Management and Interoperability Support for Natural History Museums

Konstantinos Makris, Giannis Skevakis, Varvara Kalokyri, Polyxeni Arapi, and Stavros Christodoulakis

TUC/MUSIC, Lab. of Distributed Multimedia Information Systems and Applications, Technical University of Crete, University Campus, 73100, Chania, Greece
{makris,skevakis,vkalokyri,xenia,stavros}@ced.tuc.gr

Abstract. Natural History Museums (NHMs) are 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. This paper presents 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 biodiversity networks like BioCASE and GBIF.

Keywords: digital curation, preservation metadata, Europeana, BioCASE.

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 [9]. 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 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, as well as the inadequacy of current content organization and the metadata used.

The Natural Europe project [15] 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

T. Aalberg et al. (Eds.): TPDL 2013, LNCS 8092, pp. 120–131, 2013.
© Springer-Verlag Berlin Heidelberg 2013

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 a number of strong requirements for metadata management, and establish 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, as well as (b) interconnect their digital libraries and expose their Cultural Heritage Object (CHO) metadata records to Europeana.eu and biodiversity networks (i.e., BioCASE [2] and GBIF [10]).

This paper 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.

2 The Natural Europe Cultural Environment (NECE)

The Natural Europe Cultural Environment (NECE) is a node in the cultural perspective of the Natural Europe project architecture [13]. It refers to the toolset deployed at each participating NHM, consisting of the Multimedia Authoring Tool (MMAT) and its underlying repository that facilitate the complete metadata management life-cycle: ingestion, maintenance, curation, and dissemination of CHO metadata. NECE also specifies how legacy metadata are migrated into Natural Europe.

In the context of Natural Europe, 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 (shared 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 network, contributing their high quality content to biodiversity communities.

The following sections present MMAT along with its underlying repository (i.e., CHO Repository), identifying their basic architectural components and their internal functionality.

2.1 The MultiMedia Authoring Tool (MMAT)

The Multimedia Authoring Tool (MMAT) is the first step towards allowing the connection of digital collections with Europeana. 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. Moreover, it establishes the interoperability between museums and Europeana and the seamless ingestion of legacy metadata. MMAT supports a rich metadata element set, the Natural Europe CHO Application Profile [14], which is a superset of the Europeana Semantic Elements (ESE) [8] metadata format, as well as a variety of the most popular multimedia

formats. The development of the Natural Europe CHO Application Profile was 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. The main features of MMAT 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.

MMAT adopts the Google Web Toolkit (GWT) [11] technology, which enables the web applications to perform part of their business logic into the client side and part on the server side. The client side refers to 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. 1.

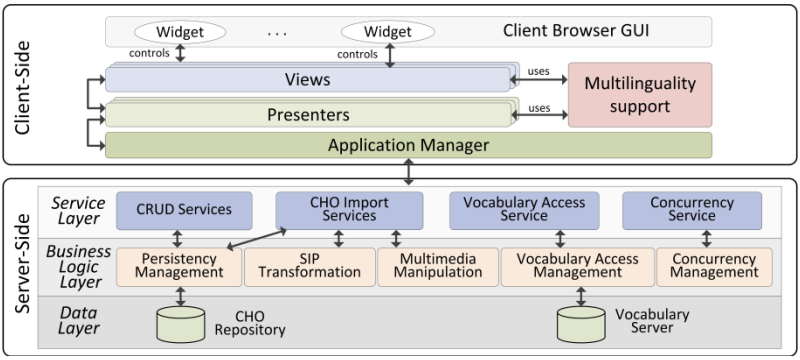

Fig. 1. MMAT Architecture

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) [16] design pattern and accommodates modules with discrete roles in managing and delivering/fetching content and metadata from/to the Client Browser GUI to/from the 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.
- The *View* modules control composite widgets and are responsible for their layout. They dispatch user action events to their corresponding Presenters for 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 *Application Manager* acts as a centralized point of control, handling the communication between the Presenters and the server side by making calls to the services exposed in the Service Layer, and notifying Presenters for their responses.
- 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.

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 of the Natural Europe federal node [13].

2.2 The CHO Repository

The CHO Repository handles both content and metadata and adopts the OAIS Reference Model [12] 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 network.

Fig. 2 presents the overall architecture of the CHO Repository with emphasis to the internal software modules (i.e., Ingest Module, Archival Module, Indexing Module, and Access Module), employed by the repository.

The **Ingest Module** is responsible for the ingestion of an information package (i.e., CHOs, CHO records/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.

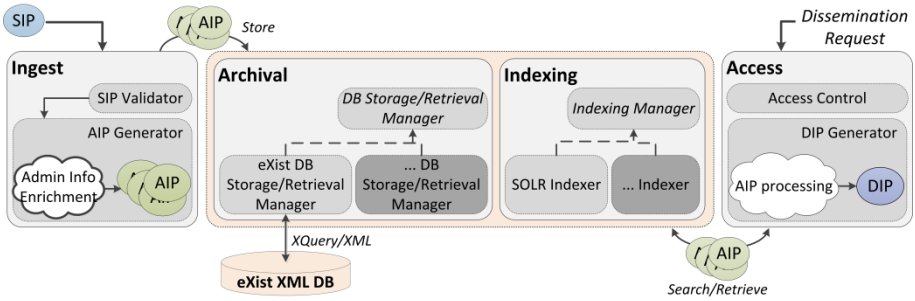

Fig. 2. CHO Repository Architecture

The **Archival Module** receives AIPs from the Ingest 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, etc.). A dedicated eXist DB Storage/Retrieval Manager has been implemented, supporting database specific storage and retrieval operations in an eXist XML DB instance, using XQuery/XML. After the storage, update, or deletion of an AIP, the Archival Module notifies the Indexing Module of the changes.

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 a number of services allowing Dissemination Information Package (DIP) consumers (i.e., the MMAT, the harvester application of the Natural Europe federal node and the BioCASE network) 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 passed to the DIP Generator component so as to be further processed for creating the final DIP that will be delivered to the DIP consumer. Additionally, the Access Module offers an OAI-PMH interface, allowing NHMs to expose their metadata in order to be harvested by the Natural Europe federal node and subsequently to Europeana. Finally, it implements the BioCASE protocol, enabling the connection to biodiversity networks like BioCASE and GBIF.

3 The Metadata Management Life-Cycle Process

The complete life-cycle process that NECE defines for the NHM metadata management comprises four phases: (a) pre-ingestion phase, (b) ingestion phase, (c) maintenance phase, and (d) dissemination phase.

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:

- Web publishing of the CHOs, along with their respective thumbnails (e.g., using MMAT), making them accessible to end users.
- Metadata unification of existing CHO descriptions by preparing XML records conforming to the Natural Europe CHO Application Profile.

During the **ingestion phase** any existing CHOs and CHO descriptions 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. Fig. 3 presents an indicative screenshot of this tool.
- 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.

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, the BioCASE network etc.

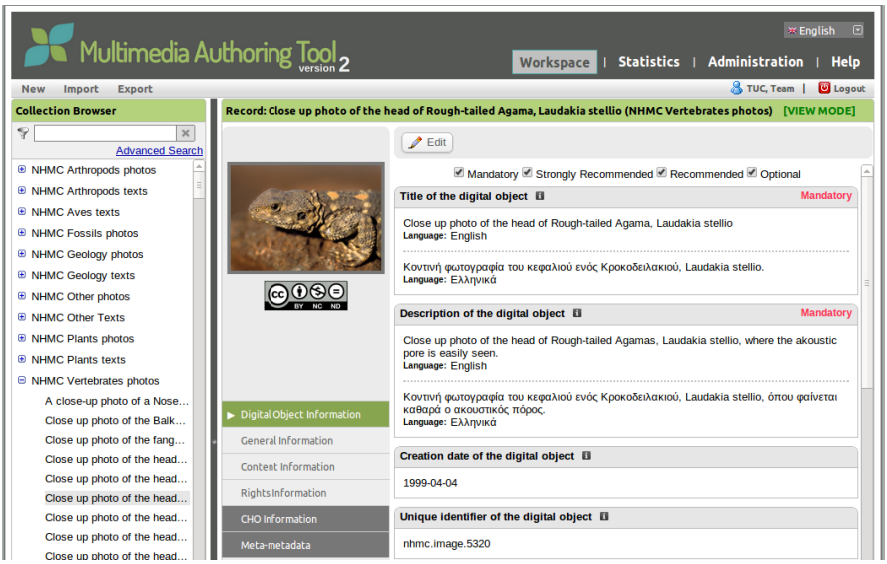

Fig. 3. The Multimedia Authoring Tool in use

4 Connection of the Natural Europe Cultural Environment with BioCASE

The Biological Collection Access Service for Europe (BioCASE) [2] is a transnational network of biological collections of all kinds. BioCASE enables widespread unified access to distributed and heterogeneous European collections and observational databases using open-source, system-independent software and open data standards/protocols.

In order for data providers to connect to this network, 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 is based on the ABCD Schema [1], which is the standard for access and exchange of data about specimens and observations. The ABCD Schema is rather huge, offering nearly 1200 different concepts.

Fig. 4 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.

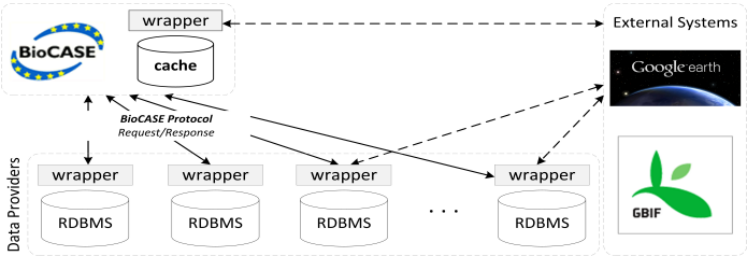

Fig. 4. BioCASE architecture

Although BioCASE supports a variety of relational databases, it does not support non-SQL databases (e.g., XML DBMS). This is also the case of MMAT, which is backed up by an eXist XML Database. To address this problem, we have built a customized wrapper on top of the data providers' repositories (Fig. 5). The wrapper is able to analyze BioCASE protocol requests and transform them to XQueries, exploiting mappings between the Data Provider's schema and the ABCD schema. Towards this end, a draft mapping of the Natural Europe CHO Application Profile to ABCD was produced based on BioCASE practices [3]. The XQueries are executed over the providers' repositories and the results are delivered to the client after being transformed to an ABCD document. The above approach was implemented and successfully tested with a local BioCASE portal installation, retrieving CHOs from all federated node CHO Repositories[1].

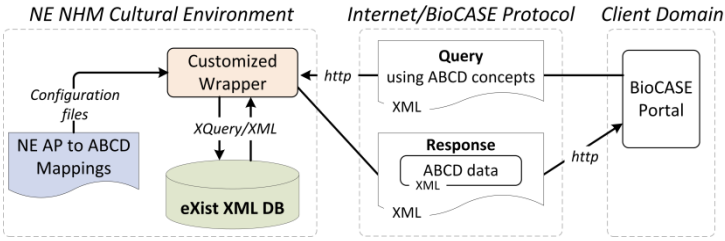

Fig. 5. Connecting Natural Europe Cultural Environment with BioCASE

5 Deployment, Use and Evaluation

The MMAT and the CHO Repository have been deployed on each Natural History Museum participating in the project, allowing the curators to publish, semantically describe, manage and disseminate the CHOs they will contribute to the project[2].

By today (3rd year of the project), a large number of CHOs have been published by each NHM using MMAT, as presented in Table 1. Till the end of the project the total number of CHOs (2nd column) for each NHM will be reached.

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

Table 1. Number of CHOs published and to be published by each NHM using MMAT

Museum	Published CHOs	Remaining CHOs	TOTAL
Natural History Museum of Crete (**NHMC**)	2611	1399	**4010**
National Museum of Natural History – University of Lisbon (**MNHNL**)	1708	902	**2610**
Jura-Museum Eichstätt (**JME**)	1172	478	**1650**
Arctic Center (**AC**)	302	178	**480**
Hungarian Natural History Museum (**HNHM**)	3134	1076	**4210**
Estonian Museum of Natural History (**TNHM**)	1923	0	**1923**

Improvements of the user-interface and the search functionalities 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 and will be performed in a number of curator workshops organized by the participating NHMs.

5.1 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; 88 errors (9 major) were detected and fixed[3].

5.2 Curator Workshops

A number of curator workshops were organized by the NHMs [17], attracting participants from different professions (presented in Table 2), while more are planned for the current year of the project.

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

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

The participants of the workshops carried out by AC, JME, NHMC, TNHM and HNHM were asked about their experience with metadata. Twenty out of thirty three curators had already described items from their collection using metadata. However, most of the workshop participants had seldom or never used any tool to upload multimedia files from their museum collections or manage museum digital collections. In addition, the exploitation of digital collections in education is 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).

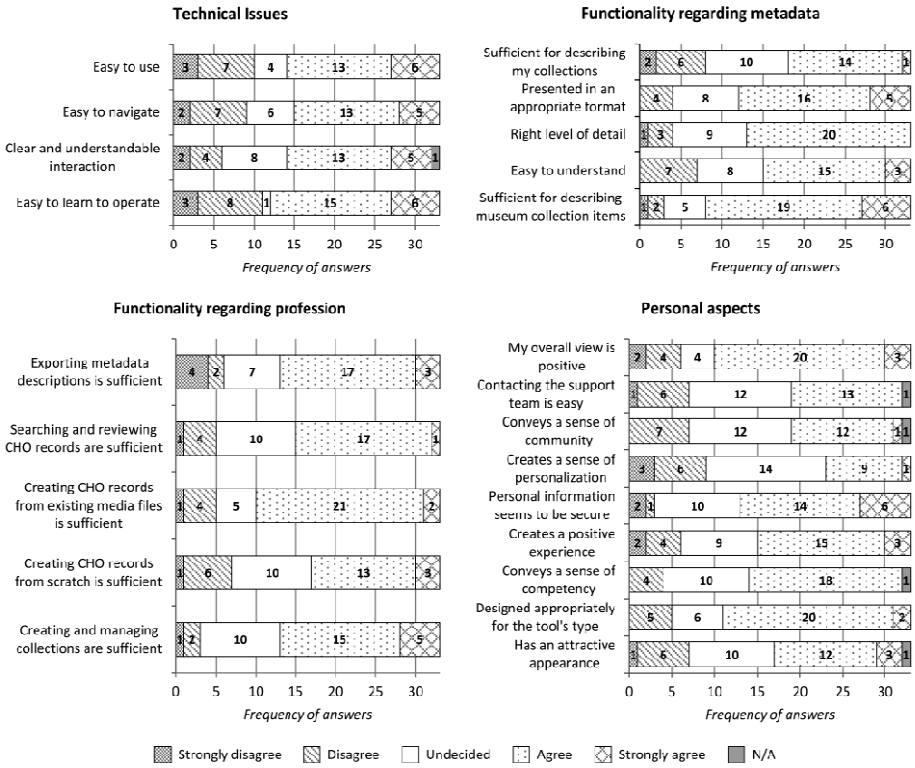

Fig. 6. Results of the satisfaction questionnaire regarding MMAT

After interacting with MMAT, the participants of the NHMC, JME, HNHM, AC and TNHM workshops were administered the satisfaction questionnaire. The results are presented in four parts (Fig. 6): Technical issues, functionality regarding metadata, functionality regarding profession and personal aspects.

- **Technical Issues:** MMAT was rated positively by the majority of the curators of the NHMC, JME, HNHM, AC and TNHM workshops. Twenty one of the partici-

pants found the MMAT easy to learn to operate and only six identified the interaction with the system as not clear/understandable.

- **Functionality Regarding Metadata:** In general, the use of metadata elements related to MMAT was rated as satisfying; only three of the curators found that the elements are not sufficient for describing their collections items.
- **Functionality Regarding Profession:** The functionality regarding the profession of curation is generally satisfying. Creation of CHO records/collections is sufficient. 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 MMAT and secure in providing their personal information.

6 Related Work

CollectiveAccess [5] 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 Europeana's 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 [4] 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 [6] 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, which greatly complicates the collaboration of museum curators.

Finally, none of the above tools provides out-of-the-box support for connection to any biodiversity network (e.g., BioCASE, GBIF).

7 Conclusion and Future Work

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 biodiversity networks, e.g. BioCASE and GBIF. This infrastructure consists of the Multimedia Authoring Tool and the CHO Repository. 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. A large number of CHOs has already been published and more are to be published till the end of the project. A long term vision of the project is to attract more NHMs to join this effort.

We are currently developing a semantically rich cultural heritage infrastructure for NHMs, as a proof of concept, by supporting EDM [7]. This will give a Semantic Web perspective to the Natural Europe cultural content. Towards this end, the Natural Europe cultural metadata records will be semantically enriched with well-known vocabularies and thesaurus like Geonames, DBpedia, GEMET and CoL/uBio. Part of this procedure is going to be performed through automatic processes by exploiting existing web services. Object aggregations will be created and the semantically enriched metadata records will be transformed to EDM.

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. ABCD Schema, `http://wiki.tdwg.org/ABCD/`
2. BioCASE, `http://www.biocase.org/`
3. BioCASE practices, `http://wiki.bgbm.org/bps/index.php/CommonABCD2Concepts`
4. CollectionSpace, `http://www.collectionspace.org`
5. CollectiveAccess, `http://www.collectiveaccess.org/`
6. Custodea, `http://www.custodea.com/en/home`
7. Europeana Data Model Definition V.5.2.3, `http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22`
8. Europeana Semantic Elements Specification V.3.4.1, `http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57`
9. Falk, J.H., Storksdieck, M.: Using the Contextual Model of Learning to Understand Visitor Learning from a Science Center Exhibition. Wiley InterScience (2005)
10. Global Biodiversity Information Facility (GBIF), `http://www.gbif.org/`
11. Google Web Toolkit (GWT), http://code.google.com/intl/el-GR/webtoolkit/
12. ISO 14721:2003 Open Archival Information System (OAIS) Reference Model, `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683`
13. Makris, K., Skevakis, G., Kalokyri, V., Gioldasis, N., Kazasis, F.G., Christodoulakis, S.: Bringing Environmental Culture Content into the Europeana.eu Portal: The Natural Europe Digital Libraries Federation Infrastructure. In: García-Barriocanal, E., Cebeci, Z., Okur, M.C., Öztürk, A. (eds.) MTSR 2011. CCIS, vol. 240, pp. 400–411. Springer, Heidelberg (2011)
14. Natural Europe Cultural Heritage Object Application Profile, `http://wiki.natural-europe.eu/index.php?title=Natural_Europe_Cultural_Heritage_Object_Application_Profile`
15. Natural Europe Project, `http://www.natural-europe.eu`
16. Potel, M.: MVP: Model-View-Presenter. The Taligent Programming Model for C++ and Java (1996)
17. Sattler, S., Bogner, F.: D6.2 Integrated Pilot Evaluation Report. Natural Europe Project (Ref. No 250579, Area CIP-ICT-PSP.2009.2.5 – Digital Libraries) (2013)