

DISMARC and BHL-Europe: multilingual access to two aggregation platforms for Europeana

Walter Koch¹, Henning Scholz²

¹ AIT Angewandte Informationstechnik Forschungsgesellschaft mbH, Klosterwiesgasse 32, 8010 Graz, Austria

² Museum für Naturkunde – Leibniz Institute for Research on Evolution and Biodiversity at the Humboldt University Berlin, Invalidenstrasse 43, 10115 Berlin, Germany

Abstract. Digital audio content and digitized biodiversity literature are aggregated in two platforms and delivered to Europeana, the European Digital Library. The audio platform which is already fully operational was developed in course of the DISMARC project and provides the baseline for multilingual data mapping and access modules of Biodiversity Heritage Library (BHL) for Europe, an eContentPlus project which complements the successful BHL operations started in the United States. Multilingual vocabularies which are exposed as web services are used for semantic enrichment of data during the input process, for the query expansion and when presenting search results.

Keywords: DISMARC, audio, BHL-EUROPE, biodiversity, SKOS, WebServices, semantics, Europeana

1 Introduction

Within the European Digital Library (Europeana)¹ concept the access to domain specific or national/regional aggregation platforms is of paramount importance. Several projects funded by the European Commission through the eContentPlus Programme are targeting to set up aggregation systems which provide metadata to Europeana and links to digital resources. An “audio-pillar” for the Pan-European Library has been developed during the implementation of the DISMARC (DIScovering Music ARChives)-Project which started in 2006 and finished successfully in August 2008. After half a year fully operation the DISMARC² meta store (aggregation platform) contains nearly 2 million metadata records in different languages. Due to the fact that the project was lead by the “Multi-Kulti” department of RBB (Radio Berlin Brandenburg) it was possible to develop multilingual vocabularies and word lists in over 25 languages since for all these languages

¹ Europeana – a single access point to Europe's cultural heritage (15 June 2009), http://ec.europa.eu/information_society/activities/digital_libraries/europeana/index_en.htm

² DIScovering Music ARChives (15 June 2009), <http://www.dismarc.eu>

translators have been available. The vocabularies, implemented as SOAP³/WSDL⁴ based WebServices, are used during the harmonization and input processes of archival records as well as at the user side when accessing the DISMARC meta store via the multilingual query and presentation interface. The DISMARC technical system will be further enhanced for the management of audio content related metadata within the EuropeanaConnect⁵ project and forms the basis for a first pilot system supporting the Europeana Aggregation Platform for biodiversity literature. This work is carried out within the framework of the BHL-Europe project which can be considered as an “European Complement” to BHL⁶, the Biodiversity Heritage Library, and started as a three year eContentPlus project in May 2009.

2 The DISMARC audio aggregation platform for Europeana

As stated in the Description of Work: “DISMARC uncovers large amounts of under-exposed European cultural, scientific and scholarly music audio. Content providers archives, broadcasters, museums, universities, research institutes, private collectors will be able to open up their collections to the wider world”[1], the project acts as an “audio aggregator” and guarantees operations for five years beyond the official project end which was in autumn 2008. The technical components of the DISMARC (DM) system consist of the “DM-meta store node” and the “DM-ontology node”, both of them can be accessed via the DM-portal at www.DISMARC.eu. DM nodes offer functionalities for managers, domain experts, and end users either in the “back office” or “front office” mode of the portal; the different roles of DM actors within the different processes (input, aggregation, data mapping, vocabulary management, data access, etc) have been defined in the “DM-workflow” which has been elaborated using BPMN⁷ the Business Process Modelling Notation.

2.1 The DISMARC nodes

The DM meta store node can be a constituent part of a “DM aggregation network” functioning as “DM-sub node” to another “DM-meta store node” and can be delivered in SaaS⁸ - Software as a Service - mode as image of a Virtual Engine (“DISMARC-on-a-Stick”). This node includes sub components like: OAI⁹ provider and harvester,

³ SOAP Version 1.2 Part 1: Messaging Framework (Second Edition) (15 June 2009), <http://www.w3.org/TR/soap12-part1/>

⁴ Web Services Description Language (WSDL) Version 2.0 (15 June 2009), <http://www.w3.org/TR/wsdl20>

⁵ EuropeanaConnect (15 June 2009), <http://www.europeanaconnect.eu/>

⁶ BHL-Wiki (15 June 2009) <https://bhl.wikispaces.com/>

⁷ Business Process Modeling Notation (BPMN) (15 June 2009), <http://www.omg.org/spec/BPMN/1.2/PDF/>

⁸ Software as a service (15 June 2009), http://en.wikipedia.org/wiki/Software_as_a_service

⁹ Open Archives Initiative (15 June 2009), <http://www.openarchives.org/>

search and browse subsystem (browsing supported by index lookup for all meta data elements or controlled vocabularies implemented as lists or trees (taxonomies), query expander, data mapping tools, administration and user management subsystem.

The DM ontology node offers import and export facilities based on SKOS¹⁰ the Simple Knowledge Organisation System, Management tools (including translation services) for multilingual Controlled Vocabularies (CV) and SOAP/WSDL based WebServices which provide generic functionalities as described in ANSI Z39.19-2005¹¹ and Thesaurus specific functionalities (eg specific meta data elements which can be returned in a “query response” of a service); as registered user one can have access to the DM vocabulary WebServices (WS) via the DM-portal and integrate them into other applications. The vocabulary WS calculate “e-points” which can be used for charging service requests on a “pay per view” basis.

2.2 Multilinguality

Multilingual aspects are provided at different levels: the translators can expand a multilingual word list which contain preferred terms used in partner archives and the audio domain, multilingual vocabularies (eg IconClass¹²) can be imported, during the input and mapping processes of raw data provided by an Archive can be expanded by adding relevant terms taken from a DM controlled vocabulary, queries can be expanded by adding terms (via DM controlled vocabularies) in selected languages to search terms, result records show all equivalents of a term in translated languages provided the term is included in a DM CV, the portal language itself can be selected in 20+ languages.

3 BHL-Europe as aggregation platform for biodiversity literature

The Biodiversity Heritage Library (BHL) began as a consortium of 10 natural history, botanical, and research libraries in 2007, working together to digitize the published literature of biodiversity held in their respective collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons.” Discussions are now underway with other nations moving BHL to a global initiative. BHL-Europe has recently started to make the biodiversity knowledge of Europe available to everybody who is interested by improving the interoperability of European biodiversity digital libraries. As a Best Practice Network, this will be done by the innovative application of proven technologies. Resources of other projects like DISMARC will be re-used to not reinvent the wheel. Eventually,

¹⁰ SKOS Simple Knowledge Organization System (15 June 2009), <http://www.w3.org/2004/02/skos/>

¹¹ ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (15 June 2009), http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a

¹² IconClass (15 June 2009), <http://www.iconclass.nl/>

BHL-Europe will provide a multilingual access point for digital content through EUROPEANA providing the first major corpus of science material to the European Digital Library. In addition, it will provide a robust and multilingual biodiversity community portal with sophisticated search tools and open, distributed architecture. This will be done in close collaboration with BHL to also support the internationalization of this initiative.

3.1 Technical aspects

The technical architecture of BHL-Europe is built around an OAIS¹³-compliant repository system. Within the Pre-Ingest processes mapping tools derived from the DISMARC project will be tested and integrated into the first BHL-Europe Pilot System which is foreseen to be available by end of 2009. It will also be checked if ETL¹⁴ – (Extract-Transform-Load) technology as used in BI¹⁵ (Business Intelligence) applications can be applied. The “BHL-Europe Access System” will take input from the existing BHL¹⁶ System, specifications from BHL-Europe partners and the code base and experiences provided by DISMARC. Main activities for the first prototype contain: integration of a meta data scheme fulfilling the requirements of BHL, integration of a gateway for different Vocabulary WebServices (uBio¹⁷, DISMARC, etc), data mapping for selected BHL-Europe partners and wrapping records for METS¹⁸ based bulk load (ingest) into the BHL-Europe repository system, setup of an initial version of the BHL-Europe aggregation platform for Europeana. Further versions of the BHL-Europe systems will provide navigation in semantic networks which implementation is foreseen using XTM-TopicMap¹⁹ standard eventually on top of a RDF triple engine (to be defined in course of the project implementation).

Since August 2009 a first prototype of the BHL-Europe test portal is available²⁰ and already provides access to 40.000 bibliographic items and connected digital resources from several BHL-Europe project partners. The following screen shots demonstrate the use of a multilingual thesaurus (eras) during query formulation and when presenting detailed metadata for a relevant item. This vocabulary (dmEras) was taken from the DISMARC project and is used to extend the metadata during the import process into the aggregation platform.

¹³ ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model (15 June 2009), http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

¹⁴ Extract, transform, load (15 June 2009), http://en.wikipedia.org/wiki/Extract,_transform,_load

¹⁵ Business intelligence (15 June 2009), http://en.wikipedia.org/wiki/Business_intelligence

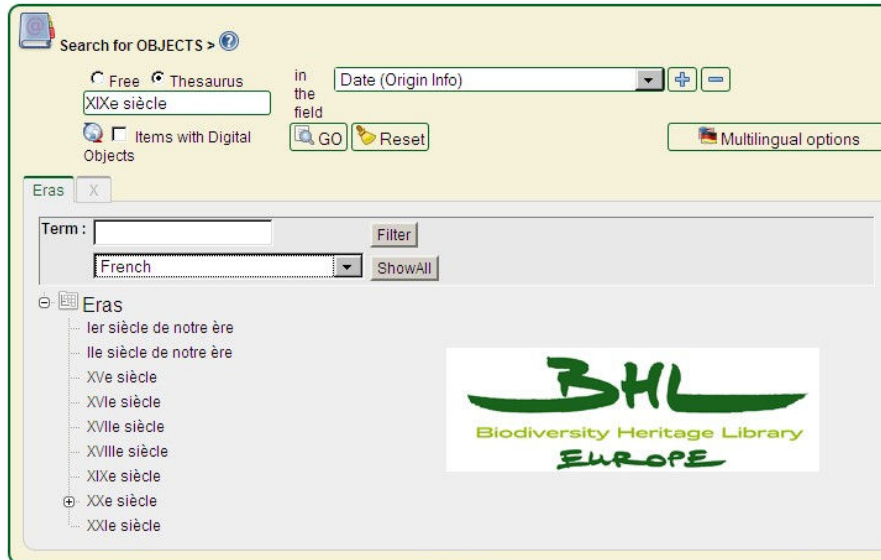
¹⁶ BHL-System (15 June 2009), <http://www.biodiversitylibrary.org/>

¹⁷ uBio (15 June 2009), <http://www.ubio.org/>

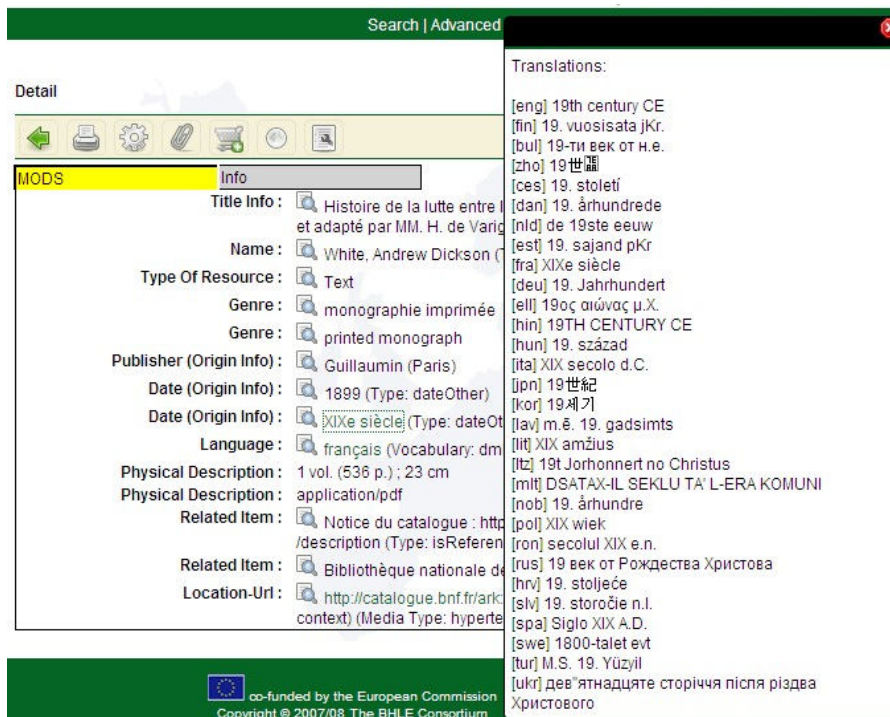
¹⁸ METS (Metadata Encoding & Transmission Standard) (15 June 2009), <http://www.loc.gov/standards/mets/>

¹⁹ XML Topic Maps (XTM) 1.0. (15 June 2009), <http://www.topicmaps.org/xtm>

²⁰ BHL-Europe; Biodiversity Heritage Library Test Portal (25 August 2009), <http://bhl.ait.co.at>



Screenshot 1: Thesaurus supported selection of search terms



Screenshot 2: Translation of the search term used into several languages