AGGREGATION AND MANAGEMENT OF METADATA IN THE CONTEXT OF EUROPEANA

by Walter Koch & Gerda Koch

Abstract: *The creation of connected content and the linking of metadata are basic requirements for the realisation of the semantic web. Semantic linkage of data enables the joint search of heterogeneous databases and facilitates future machine learning. The present article outlines the metadata management and metadata linking activities of the European Digital Library. A short overview on the current core research areas and implementation strategies in this field is presented. Various projects and metadata services tailored to natural history data, regional cultural heritage data and audio collections are described.*

Keywords: *Linked data; data enrichment; microservices; data aggregation; Europeana; cultural heritage metadata; natural history data; audio collections; data annotation*

AGGREGATION UND MANAGEMENT VON METADATEN IM KONTEXT VON EUROPEANA

Zusammenfassung: *Mit dem In-Beziehung-Setzen und Verlinken von Daten im Internet wird der Weg zur Umsetzung des semantischen Webs geebnet. Erst die semantische Verbindung von heterogenen Datenbeständen ermöglicht übergreifende Suchvorgänge und späteres „Machine Learning". Im Artikel werden die Aktivitäten der Europäischen Digitalen Bibliothek im Bereich des Metadatenmanagements und der semantischen Verlinkung von Daten skizziert. Dabei wird ein kurzer Überblick zu aktuellen Forschungsschwerpunkten und Umsetzungsstrategien gegeben und einzelne Projekte und maßgeschneiderte Serviceangebote für naturhistorische Daten, regionale Kultureinrichtungen und Audiosammlungen werden beschrieben.*

Schlüsselwörter: *Verlinkung von Daten; Anreicherung von Daten; Microservices; Datenaggregation; Europeana; Kulturerbedaten; Naturhistorische Daten; Audiosammlungen; Annotationen*

Contents

1. Linking of heritage data in Europe

The European Digital Library, Europeana[1], was first launched in 2008 and is the cross-domain central portal and single entry point to Europe's digitized cultural heritage. The Europeana Data Model (EDM)[2] is the metadata model that is used for aggregating and ingesting data from the diverse European cultural and natural heritage repositories into the joint digital library. EDM describes the data using the Resource Description Framework (RDF)[3] and re-uses the Simple Knowledge Organization System (SKOS)[4] framework for the integration of vocabulary concept terms. The Resource Description Framework (RDF) provides the syntax and rules that build the basis of linked data and structures information as triples of subjects, predicates and objects. In addition, ontologies and vocabularies reduce the complexity of the world and help to put vast amounts of data into structured forms.

Popular use cases often list the following benefits of publishing Linked Open Data:

- Being findable on major search engines and social platforms in the first place
- Cross-language retrieval based on vocabularies
- Vocabulary based facets with multilingual facet labels
- Auto completion with semantic disambiguation and suggestions of related searches
- Cross-linked entity pages for concepts, agents, places and periods

Recognizing these advantages, Europeana has started to strongly encourage its partners to provide open metadata, so called Linked Open Data (LOD), with an increasing proportion of links to content marked Public Domain or bearing Creative Commons licenses.

A basic premise of the LOD approach is to reuse, wherever possible, terms from existing standard vocabularies, rather than reinvent them when publishing Resource Description Framework (RDF) data. This maximizes

the probability that the data can be used without additional modifications by applications that may be tuned to well-known vocabularies. Vocabulary standards for creating (ISO 25964)[5] and publishing vocabularies (SKOS) on the web offer solutions to interconnect isolated data silos and support cross search and data comparison. SKOS (Simple Knowledge Organization System) has become an accepted standard model for expressing the structure and content of concept schemes like vocabularies, authority lists and thesauri on the web using RDF.

Data enrichment with vocabulary terms can be done right when the data object is registered in the local cataloguing system, or "afterwards", when the metadata is processed further. Vocabulary web services provide the functionalities to enrich metadata (catalogue data) with standard vocabulary information that further allow the automatic linking of data from various online data sets. When data is ingested into joint virtual catalogues (like Europeana), these automated enrichment functionalities are often used in order to support semantic linking of heterogeneous data and easy browsing through the entire repository. The following paragraphs describe recent tools and services for the semantic enrichment of data that have been developed by European research projects.

2. Natural history data in Europeana

OpenUp![6] started in 2011 as a three year project of the European Commission with the main goal to provide online access to a wide range of natural history collection objects and to connect the cultural heritage and natural history domains. Today OpenUp! is a constantly growing network and Europeana's aggregator for the natural history domain. OpenUp! is using international vocabularies within a common names web service that draws common names from about 25 different vernacular names lists from various countries. The enrichment with common name information opens up the natural history research data to a considerable number of other user communities, like tourists, pupils or nature enthusiasts. The OpenUp! technical aggregator uses the open source Business Intelligence tool Pentaho Kettle[7] for data integration and ETL (Extract, Transform, Load) mechanisms. The chosen tool provides the ETL functionalities needed for the processing of the distributed natural history datasets across Europe. As a result, the entire data management process for the Europeana data provision was tailored to three steps: transform, validate, and OAI[8]-import.

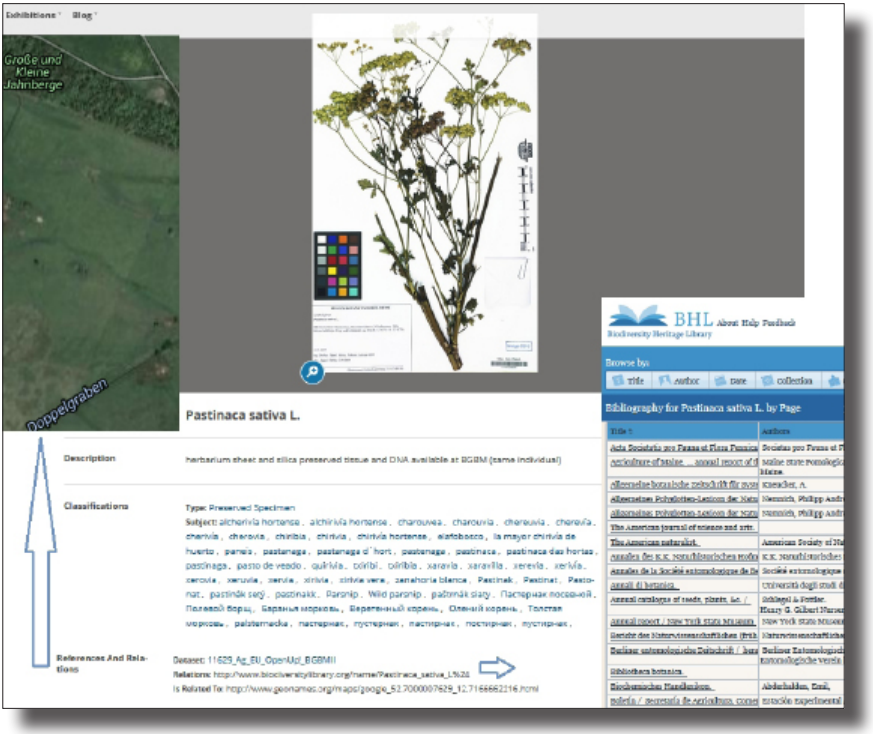

Fig. 1: Biodiversity data enriched with common names and related information

3. Serving the data development and aggregation of regional and cultural heritage institutions

Regional and local cultural heritage institutions may use the support services and tools that were developed within the LoCloud EU-project[9] as well. The LoCloud tools and services simplify ingestion of data into Europeana and help to improve metadata quality by means of cloud-based services. LoCloud uses a microservice architecture that consists of a suite of small and independent services that may be compiled into a single application. Microservices are independently deployable and scalable and they could even be written in different programming languages by different software teams. The microservices are built upon business capabilities and communicate via web service requests or remote procedure calls. LoCloud data providers and cultural heritage technicians may choose among the following microservices:

- Geo-location API (provides geoinformation to data)

- Vocabulary service (enriches data with various vocabulary terms)
- Historic Place Names service (adds historic place names information)
- Geo-coding application (tool to manually enrich data with geoinformation)
- Vocabulary matching service (automated vocabulary matching)
- Background linking service (automated enrichment with Wikipedia information)
- Wikimedia application (capture data from selected Wikimedia resources)

The LoCloud vocabulary microservice, for instance, provides the business capability for vocabulary management and enrichment of metadata. The LoCloud vocabularies have been imported into the TemaTres[10] tool in order to use them via web services in the aggregation process. That way the LoCloud microservice "Generic enrichment" automatically receives the vocabularies available in the vocabulary tool and uses them during the automated enrichment process conducted in the LoCloud aggregation facility MORe. Further on, instead of matching the vocabularies automatically to the metadata, it is also possible to select individual terms from the LoCloud vocabularies that should be added to the metadata.

One major benefit of microservices is that these small and independent services may be plugged in into other applications and can be used via the APIs they provide. The LoCloud vocabulary tool can also be used to create vocabularies from scratch or to import already existing vocabularies. By importing existing vocabularies to the tool the vocabularies become available in the SKOS format which has its own web presence that can be used for further semantic linking afterwards. Moreover, the LoCloud TemaTres vocabulary installation allows online collaboration when creating and extending vocabularies (for example in order to create new translations of existing vocabularies).

Summing up, all LoCloud cloud-based SaaS (Software as a Service) tools serve the cultural heritage community in many ways. The enrichment services help to improve data quality for the content providers and Europeana. The data capturing service allows re-use of already existing online cultural content and its proper ingestion into Europeana. The scalability and various options for deploying and using the microservices help tailor the services offered to the needs and capabilities of cultural heritage institutions.

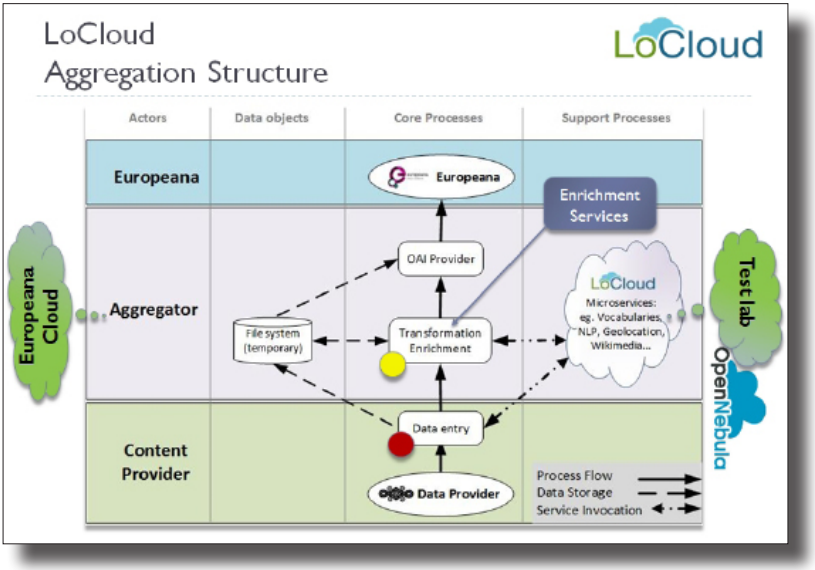

Fig. 2: The LoCloud Aggregation structure

4. Annotating content

Several Europeana related projects have recently dealt with the annotation of content. Among these projects are the PATHS project, investigating automatic semantic enrichments, the DM2E project (scholarly annotations using the Pundit toolset), the SEALINCmedia project (expert annotation "nichesourcing", developed the Accurator tool), the Europeana Creative project (pre-alpha version of the Annotations API), Europeana 1989 (annotations in HistoryPin.org), Europeana V3 (roundtripping of annotations with HistoryPin.org), Europeana Food and Drink (annotations in HistoryPin.org) and Europeana Sounds with a wider range of user scenarios for annotations. Within these activities Europeana adopted the W3C Web Annotation Data Model[11] which is based on RDF and defines JSON-LD[12] as its default serialization format. This offers a model for exchanging annotations across platforms but needs further investigation into its flexibility to support complex scenarios.

The "Moments of Interest" Approach of the audio aggregation platform DISMARC[13] adopts the W3C recommendation on Media Fragments[14] and defines annotations as "has annotation" relationships between data objects instead of following the complex W3C Web Annotation Data Model (see figure 3).

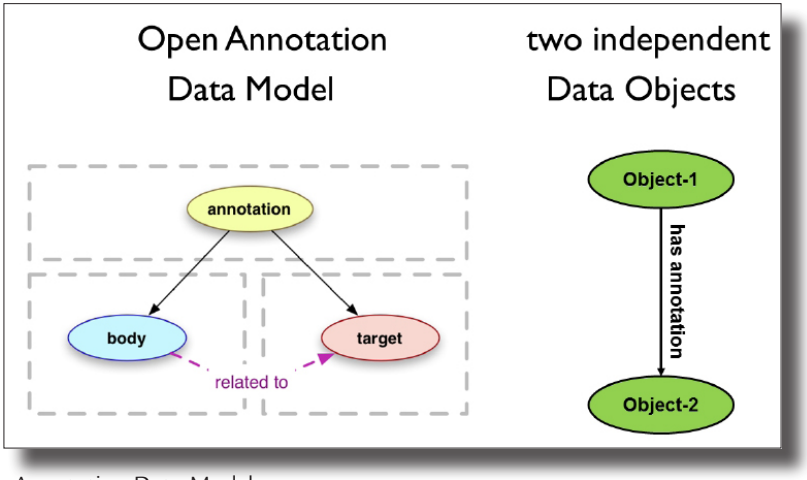

Fig. 3: Annotation Data Models

Moments of Interest can be chosen from a timeline, and therefore annotations may refer to media fragments that can be selected from audio, video or text streams etc. (see figure 4).

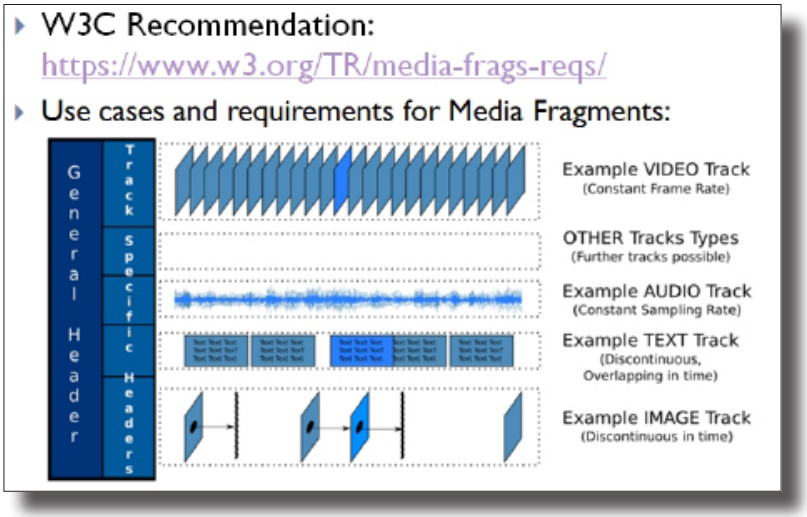

Fig. 4: Media Fragments

Different types of fragments refer to different media types. Video and audio streams lend themselves to temporal fragments, images to spatial fragments, remixed content to track fragments and data objects to named

fragments. According to this approach all Moments of Interest become independent objects and can be recorded using whatever type of metadata scheme that best fits the content.

The DISMARC v2 platform offers novel applications for the development of connected content (see figure 5) and seeks to prove whether this approach could be a better choice for linking independent data objects and annotating parts of existing data objects with additional information.

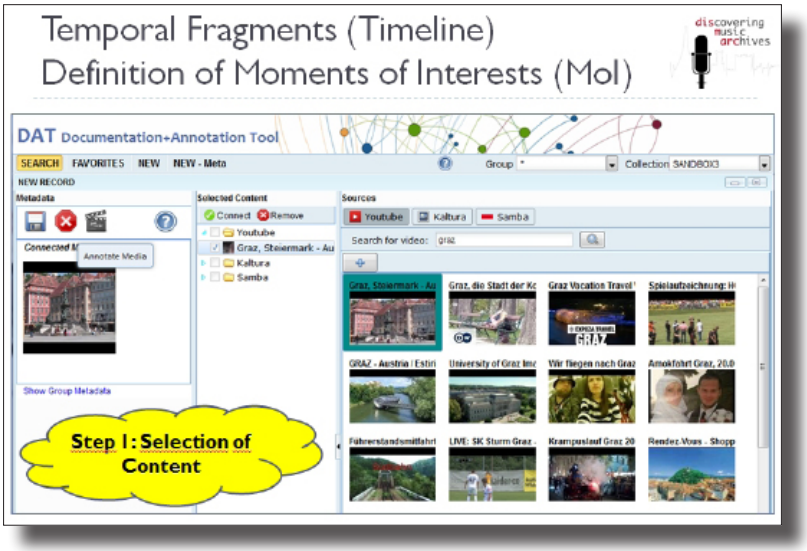

Fig. 5: Selecting content and describing Moments of Interest (DISMARC)

In the future the semantic internet will be capable of developing ontologies and connections between content without the need for human interaction. Linked information will be found and presented even more rapidly. The foundation of this process is the availability of semantically rich and open data sets. This article presented a selection of European tools, services and projects that aim at blazing the trail.

Univ. Prof. Dr. Walter Koch
Steinbeis Innovation Transfer Center IMCHI
E-Mail: walter.koch@stw.de

Mag.ª Gerda Koch, MBA
AIT Angewandte Informationstechnik ForschungsgesmbH Graz
E-Mail: kochg@ait.co.at

Notes

1 The Europeana Data Portal. http://www.europeana.eu
2 Europeana Data Model Documentation. http://pro.europeana.eu/page/edm-documentation
3 Resource Description Framework. https://www.w3.org/RDF/
4 Simple Knowledge Organization System. https://www.w3.org/TR/skos-reference/
5 ISO 25964 – the international standard for thesauri and interoperability with other vocabularies. http://www.niso.org/schemas/iso25964/
6 OpenUp! http://open-up.eu/en
7 Pentaho Data Integration. http://www.pentaho.com/product/data-integration
8 Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/
9 LoCloud. http://www.locloud.eu/
10 TemaTres. http://www.vocabularyserver.com/
11 Web Annotation Data Model. http://www.w3.org/TR/annotation-model/
12 JavaScript Object Notation for Linked Data. http://json-ld.org/
13 DISMARC. http://www.dismarc.org
14 Use cases and requirements for Media Fragments. https://www.w3.org/TR/media-frags-reqs/