CULTURAL HERITAGE CONTENT RE-USE: AN AGGREGATOR'S POINT OF VIEW

Dimitris Gavrilis [a], Marinos Ioannides [b], Eirini Theofanous [b]

[a] Digital Curation Unit – IMIS, Athena Research Center, Maroussi, 15125, Greece – d.gavrilis@dcu.gr
[b] Electrical Engineering and Information Technology Department, Cyprus University of Technology, 3036 Lemesos, Cyprus - {marinos.ioannides, eirini.theofanous}@cut.ac.cy

KEY WORDS: Metadata aggregation, metadata enrichment, micro-services, linked open data, Omeka, EDM, creative industries, tourism, metadata quality, MORe.

ABSTRACT:

This paper introduces a use case of re-using aggregated and enriched metadata for the tourism creative industry. The MORe aggregation and enrichment framework is presented along with an example for enriching cultural heritage objects harvested from a number of Omeka repositories. The enriched content is then published both to the EU Digital Library Europeana (www.europeana.eu) and to an Elastic Search component that feeds a portal aimed at providing tourists with interesting information.

1. INTRODUCTION

1.1 Related work

The work presented in this paper focuses on a use case of re-using aggregated cultural heritage related content for the tourism industry. The key component in this paper is the aggregation infrastructure that is used and more specifically two main topics are covered: a) its micro-services architecture and b) its enrichment services. These two distinct approaches have been explored partially in the bibliography with the majority of papers focusing on curation micro-services. Enrichment micro-services can be considered as a category of curation micro-services.
In [4], news items are automatically enriched with information from Linked Open Data (LOD) sets and use an ontology based browser to demonstrate the advantages of LOD enabled navigation. In (Rainer Simon et al, 2011) the authors use an annotation tool to help users annotate records with information drawn from LOD thesauri. In (Stephen Abrams et al, 2010), authors propose and present a curation micro-services infrastructure in order to demonstrate the powerful characteristics and flexibility of such an approach. In a micro-services architecture which focuses on digital curation and preservation is presented. A presented in (Kevin Clair, et al 2011), curation micro-services are also used on a thematic aggregator to enrich information and improve the quality of content.

1.2 MORe metadata aggregator

The Metadata & Object Repository (MORe) (Christos Papatheodorou et al, 2012) was established as early as 2011 in the context of the EU CARARE project. Since then has been used to aggregate over two million metadata records and delivered it to Europeana in the EDM (Europeana Data Model) format (Martin Doerr et al, 2011). In the heart of the aggregator lies a repository that is used to store metadata, maintain versions and identity relations between objects, etc. Initially, MORe was based on fedora-commons and provided enrichment services such as: de-duplication, geo-spatial, etc. The cloud based version of MORe that was introduced in the LoCloud project in 2013 replaces the fedora-commons backend of the earlier version with one based on Apache Cassandra, and standalone services with a pluggable and scalable services layer that provides for higher processing capacity and thus reduces aggregation time significantly.
The general architecture of the MoRe aggregator can be seen in Fig. 1. Its primary components are shown with two of them having a centralized role: a) the storage layer and b) the services layer. The first is responsible for handling metadata storage and the latter is responsible for gluing together the various services involved.

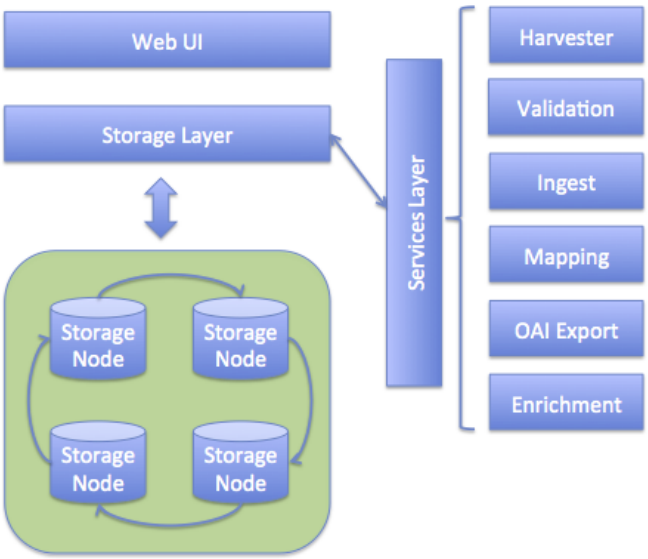

Figure 1. MoRe aggregator architecture

The main aggregation workflow used in this paper is presented in Fig. 2 and depicts all the main steps that users can perform. For each one of these steps a specific service is used to handle the work required and certain steps like the transformation and enrichment require additional work such as validation and indexing of content.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-5-W3-83-2015

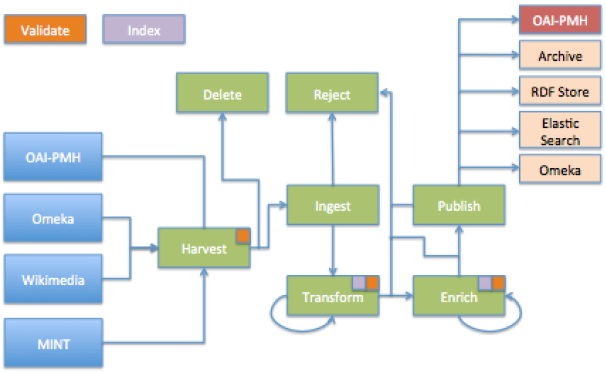

Figure 2. Aggregation workflow

1.3 Motivation

The main motivation behind building a flexible aggregation framework that includes the enrichment layer lies mainly in ensuring metadata quality and interoperability. This is based on the premise that automatic enrichment leads to metadata records that are richer, more comprehensive and provide links to other related resources such as:

- other related records
- thesauri terms
- Wikipedia lemmas

This increases the visibility of the records, improves the search results and ultimately the user experience. It also allows the records to be machine readable and thus to be automatically used by additional services as needed.

Furthermore, content re-use helps build more sustainable models because as it will be made clear, with the appropriate technologies little extra work is required.

1.4 The Omeka Repository & Content model

The work reported in this paper relies on content that has been harvested from Omeka (http://omeka.org/). Omeka is an open source repository that provides out of the box functionality for individuals and institutions that wish to publish collections on the web using international standards. Omeka provides two schemas through OAI-PMH and REST: a) a Dublin Core based metadata schema and b) a custom schema named Omeka-XML. Both use an internal Dublin Core based representation which allows administrators to extend it with new elements. It is possible to get the content in other formats (through metadata crosswalks) such as: mods, CDWA-Lite and METS. METS is used to provide the structure of each record. As we are interested in metadata aggregation, we focus here on metadata, and more specifically on OAI_DC. The OAI_DC metadata contains 15 elements. These elements were provided in a completely unqualified format, meaning that even language information (the xml:lang attribute) was not present. This is typical of the scenario of harvesting from diverse cultural heritage collections addressed here.

Available metadata is used to present records though the Omeka web portal (Fig. 3).

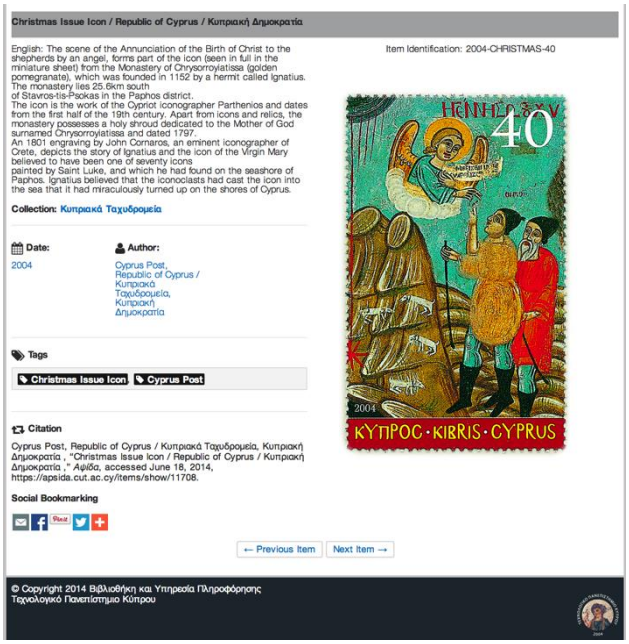

Figure 3. Sample view of a record in the Omeka repository from one of the selected stakeholders: the Postal Services in Cyprus (a stamp representing a fresco in one of the ten painted-churches in Cyprus).

Although the mods, omeka-xml and cdwalite schemas are more expressive, Omeka provides them without qualifiers, thus presenting the same problems as with the oai_dc. Hence, the approach illustrated here for OAI_DC (selected by virtue of its popularity) is applicable for these other schemas as well.

The primary elements harvested are depicted in the following table (Table 1):

Element	Description	Issues observed
dc:title	The title of the record	
dc:creator	The creator of the record	
dc:description	The description of the record	Multiple descriptions are received in two languages (English and Greek)
dc:subject	The subjects related to the record	
dc:publisher	The publisher of the record	
dc:contributor	The contributor of the record	
dc:date	The date of the record	
dc:type	The type of the binary representation of the record (e.g. Image)	
dc:format	The format of the binary representation of the record (e.g. JPG)	
dc:identifier	The identifier of the record	Multiple identifiers are received with different semantics
dc:source	The source of the record	
dc:language	The language of the record	In cases where e.g. multiple

		descriptions are received, a single string with all the language codes is received.
dc:coverage	The coverage of the record	Semantically ambiguous. In this case the spatial coordinates are received in various formats
dc:rights	The rights associated with the record	

Table 1. Elements harvested by Omeka

2. ENRICHMENT SERVICES FRAMEWORK

2.1 Overview

The enrichment services framework presented in this paper consists a generic enrichment service that orchestrates a series of enrichment micro services into simple workflows referred to as: enrichment plans. The enrichment process involves executing one or more enrichment micro services in a specific sequence (referred to as enrichment plan). Each micro-service enriches each record in a specific way (e.g. by inferring coordinates out of a place name or by adding language identifiers). Each enrichment plan can be applied to one metadata schema (e.g. OAI_DC) and each one of the enrichment services support specific schemas.

2.2 Available enrichment micro-services

At the moment there is number of available enrichment micro-services support the OAI DC and EDM (Europeana Data Model) schemas. These micro-services are:

- Language identification: this service is responsible for identifying the language of text and adding the proper qualifiers to the corresponding element. The Apache Tika Language Tools are used.
- Spatial identification & normalization: this service is responsible for identifying spatial information provided through the Coverage element and normalizing it.
- Temporal identification: this service is responsible for identifying the temporal information provided through the Coverage element. This involves dates normalization.
- Reverse geo-coding: this service is responsible for reverse geo-coding an address and place name description out of the coordinates provided.
- Spatial translation: this service adds the ability of providing a textual description of Place in various languages. The Geonames Web service is used.
- Spatial coordinate transformation: this service is based on open-geo libraries and allows transformation between coordinate systems.
- Thesauri enrichment: this service allows the ingestion manager to associate one or more thesaurus concepts to each item. Concepts, describing the collection in general are drawn from standard thesauri or authorities, such as the Library of Congress Subject Headings, GEMET, etc.

2.3 Enrichment plans

The architecture shown in Fig. 4, includes a number of enrichment micro-services combined within an overall framework in order to provide sophisticated enrichment in a variety of metadata schemas. This architecture allows configuring on the fly (with no coding at all) an enrichment plan for each content provider and collection.

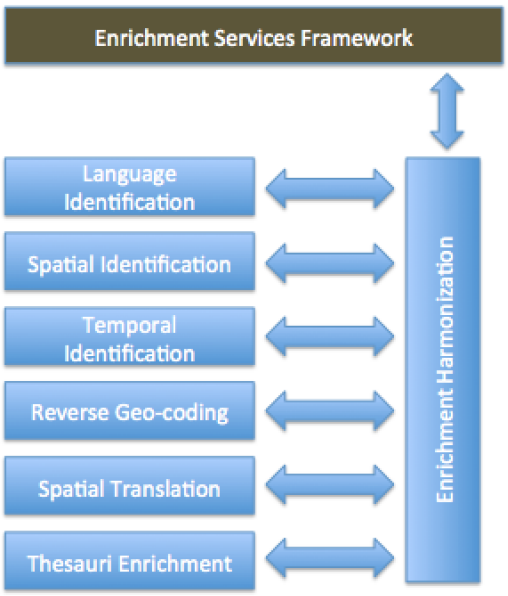

Figure 4. Enrichment services framework

The diverse enrichment micro-services are applied on each metadata record in a predefined sequence, according to rules specified by the aggregation manager. Not all micro-services are applied to all packages harvested. Each content provider primary repository or collection has its own specific characteristics, possibly requiring only a subset of the micro-services. Enrichment micro-services need to be applied in a specific order (for instance, as in Fig. 5) so that:

- Micro-services that provide information useful to other micro-services are executed first
- Micro-services with a highest degree of confidence are executed in a higher order

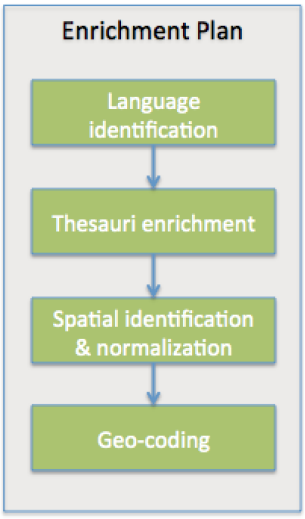

Figure 5. Enrichment services framework

The language identification step takes precedence in order to provide qualifiers to textual elements.

The thesauri enrichment, adds dc:subject terms from standard thesauri.

The spatial identification & normalization extracts the coordinate information out of a string (in this case: dc:coverage) and splits and identifies Lat/Lng coordinates.

The geo-coding service takes the coordinates and provides a place name and geoname-id which it then populates to the record.

2.4 Example of an enriched record

In this section an example of a Dublin Core (OAI DC) record before (Fig. 6) and after (Fig. 7) an enrichment plan is applied is shown. In this particular record, the xml:lang attributes are filled, one subject terms is added from library of congress subject headings, the non-parsable (due to cataloguing error) coordinates are fixed and augmented by textual descriptions and a geonames URI.

```
▼<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:sch
 http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <dc:title>Wine skin</dc:title>
  ▼<dc:creator>
     Ministry of Agriculture Natural Resources and Enviroment
   </dc:creator>
   <dc:subject>Winery</dc:subject>
  ▼<dc:description>
     Made of goat skin. They were made during the winter, beca
     safe transport of wine, zivania and other liquids. The wi
   </dc:description>
  ▼<dc:description>
     Μικρός ασκός κατασκευασμένος από δέρμα αίγιας.Οι ασκοί κατασ
     ο ψυχρός καιρός επέτρεπε την καλύτερη διατήρηση και επεξεργασ
     εμπειρίας τα κατάλληλα ζώα και η σφαγή και το γδάρσιμό τους
     αποθήκευση και ασφαλή μεταφορά κυρίως κρασιού, ζιβανίας, ξ
     ώστε να μην αλλοιώνεται η ποιότητα του περιεχομένου, αλλά κα
     ασκός μεγάλου μεγέθους λειτουργούσε και ως μέτρο όγκου του κ
     χρήση. Απο το χωριό Λαγουδερά.
   </dc:description>
   <dc:publisher>Library of Cyprus University of Technology</d
  ▼<dc:contributor>
     Ministry of Agriculture Natural Resources and Environment
   </dc:contributor>
   <dc:date>2009-2010</dc:date>
   <dc:type>Image</dc:type>
   <dc:format>JPG</dc:format>
   <dc:identifier>MKY788</dc:identifier>
   <dc:identifier>https://apsida.cut.ac.cy/items/show/11722</d
  ▼<dc:identifier>
     https://apsida.cut.ac.cy/files/original/05c37cce45d85d3c7
   </dc:identifier>
  ▼<dc:source>
     Ministry of Agriculture Natural Resources and Environment
   </dc:source>
   <dc:language>EN EL</dc:language>
   <dc:coverage>35.138873.33.396045</dc:coverage>
  ▼<dc:rights>
     The publication or reproduction in an electronic form or
   </dc:rights>
  ▼<dc:rights>
     Απαγορεύεται η δημοσίευση ή αναπαραγωγή, ηλεκτρονική ή άλλη
   </dc:rights>
</oai_dc:dc>
```

Figure 6 Enrichment services framework

In this enrichment example, it is possible to augment the record with useful information such as a place name, languages and thematic information. As it will be made clear in the next section, this is the critical part that makes the content usable in other domains.

```
▼<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:oa
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schem
 http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <dc:title xml:lang="en">Wine skin</dc:title>
  ▼<dc:creator xml:lang="en">
     Ministry of Agriculture Natural Resources and Enviroment Cyp
   </dc:creator>
   <dc:subject xml:lang="en">Winery</dc:subject>
   <dc:subject xml:lang="en">Distillation</dc:subject>
   <dc:subject xml:lang="en">http://id.loc.gov/authorities/subje
  ▼<dc:description xml:lang="en">
     Made of goat skin. They were made during the winter, because
     safe transport of wine, zivania and other liquids. The wine
   </dc:description>
  ▼<dc:description xml:lang="el">
     Μικρός ασκός κατασκευασμένος από δέρμα αίγιας.Οι ασκοί κατασκει
     ο ψυχρός καιρός επέτρεπε την καλύτερη διατήρηση και επεξεργασία
     εμπειρίας τα κατάλληλα ζώα και η σφαγή και το γδάρσιμό τους γι
     αποθήκευση και ασφαλή μεταφορά κυρίως κρασιού, ζιβανίας, ξιδι
     ώστε να μην αλλοιώνεται η ποιότητα του περιεχομένου, αλλά και γ
     ασκός μεγάλου μεγέθους λειτουργούσε και ως μέτρο όγκου του κραι
     χρήση. Απο το χωριό Λαγουδερά.
   </dc:description>
   <dc:publisher xml:lang="en">Library of Cyprus University of T
  ▼<dc:contributor xml:lang="en">
     Ministry of Agriculture Natural Resources and Environment (F
   </dc:contributor>
   <dc:date>2009-2010</dc:date>
   <dc:type>Image</dc:type>
   <dc:format>JPG</dc:format>
   <dc:identifier>MKY788</dc:identifier>
   <dc:identifier>https://apsida.cut.ac.cy/items/show/11722</dc:
  ▼<dc:identifier>
     https://apsida.cut.ac.cy/files/original/05c37cce45d85d3c773c
   </dc:identifier>
  ▼<dc:source xml:lang="en">
     Ministry of Agriculture Natural Resources and Environment (F
   </dc:source>
   <dc:language>EN EL</dc:language>
   <dc:coverage>35.138873 33.396045</dc:coverage>
   <dc:coverage>http://www.geonames.org/146769/</dc:coverage>
   <dc:coverage xml:lang="en">Athalassa, Cyprus</dc:coverage>
  ▼<dc:rights xml:lang="en">
     The publication or reproduction in an electronic form or oth
   </dc:rights>
  ▼<dc:rights xml:lang="el">
     Απαγορεύεται η δημοσίευση ή αναπαραγωγή, ηλεκτρονική ή άλλη χω
   </dc:rights>
</oai_dc:dc>
```

Figure 7. Enrichment services framework

3. CONTENT RE-USE FOR TOURISM

As it can be seen from Fig. 8, the MORe aggregator is typically used to aggregate content from multiple sources, transform it to a common schema (in our case EDM) and publish it to a single provider (in our case Europeana).

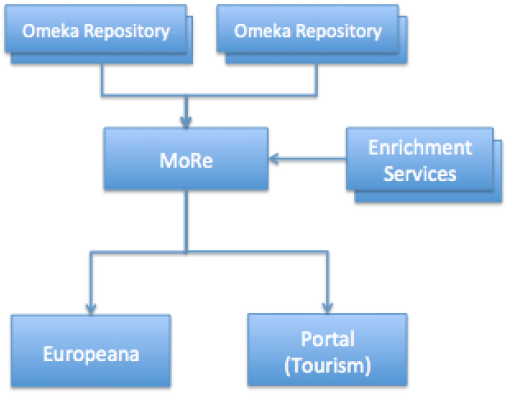

Figure 8. Use case setup

A localized portal for tourism must provide specific functionalities and has specific quality constraints in terms of content. The functionalities include:

- a thematic hierarchical browsing of the content
- the placement of the content on a map
- the ability to search for content
- the browsing of content based on language

3.1 Metadata quality

In order to meet the above requirements, all content must meet certain quality criteria. More specifically, all items must contain: language attributes at least for the titles, descriptions, subject terms and place names. Furthermore, every item must provide a place name and a set of coordinates. These requirements refer to metadata completeness per schema and case and are computed real-time every time information is received or changed.

3.2 Conditional publication of content

In an aggregation environment where large amount of content is aggregated, some automated methods and features of screening the content to be published must be provided. For this case, two cases were considered:

- Publication of content based on specific input sources
- Publication of content based on specific spatial criteria (place name / area around a set of coordinates)
- Publication of content that meets specific quality criteria.

When harvesting a large variety of content that comes from different sources and meets different quality criteria, it must be screened properly in order for all content to meet the functionalities of the portal. For example, if the portal contains a map, only items that contain coordinates should be published. Similarly, if a portal has a specific theme, only items that follow that theme should be published.

3.3 Searching: Elastic Search

In order to facilitate search, content that meets the above criteria is automatically published to an Elastic Search (ES) server. Elastic search is Lucene based index server similar to SolR. ES server requires that objects follow a JSON format, thus the common format that metadata is mapped to, is either JSON or XML. In the latter case, ES server will automatically convert the XML record to a JSON representation.

3.4 Experimental setup

To test this software environment a number of selected stakeholders (municipalities, communities etc) from Cyprus have been selected to participate in this benchmark:

- *Agios Athanasios Municipality* *www.agiosathanasios.org.cy*
- *Agios Neophytos Monastery, Paphos* *www.stneophytos.org.cy*
- *Cyprus Police* *www.police.gov.cy*
- *Cyprus Post* *www.mcw.gov.cy/dps*
- *Cyprus Rural Museum, Ministry of Agriculture, Natural Resources and Enviroment* *http://www.agrifair.gov.cy/moa/agrifair/agrifair.nsf/dmlmuseum_gr/dmlmuseum_gr?OpenDocument*
- *Cyprus Tousist Organisation* *www.visitcyprus.com*
- *Holy Metropolis of Limassol* *www.imlemesou.org*
- *Kythrea Municipality* *http://www.kythrea.com*
- *Latsia Municipality* *www.latsia.org.cy*
- *Lefkara Municipality* *www.lefkara.org.cy*
- *Limassol Municipality* *www.limassolmunicipal.com.cy* *(Pattichion Municipal Museum, Historical Archive and Research Centre)*
- *Press and Information Office, Republic of Cyprus* *www.pio.gov.cy*

4. CONCLUSIONS

In conclusion, this paper presents a first attempt at re-using content that is aggregated, transformed and enriched for the European Digital Library (Europeana) for the tourism domain. In order to automate this task, a flexible and very efficient aggregation infrastructure is required. Metadata quality measurement is also critical in order to automate the publication process (this providing a more sustainable model). Automated enrichment services ensure that quality of content will be increased thus leading to the publication of more and richer content.

ACKNOWLEDGEMENTS

This work has been conducted in the context of the EU **LoCloud** (Local Content in a Europeana Cloud) project (LoCloud is a CIP-Best Practice Network under Theme CIP-ICT-PSP.2012.2.2 with Grant agreement no: 325099).

REFERENCES

Martin Doerr, Stefan Gradmann, et.all : The Europeana Data Model (EDM). World Library and Information Congress 76th IFLA General Conference and Assembly.

Christos Papatheodorou, Costis Dallas, Christian Ertmann-Christiansen, Kate Fernie, Dimitris Gavrilis, Maria Emilia Masci, Panos Constantopoulos, Stavros Angelis : A New Architecture and Approach to Asset Representation for Europeana Aggregation: The CARARE Way. Communications in Computer and Information Science Volume 240, 2011, pp 412-423

Stephen Abrams, John Kunze, David Loy : An Emergent Micro-Services Approach to Digital Curation Infrastructure. International Journal of Digital Curation, Vol.5 Issue 1, 2010

Mannens, E., Troncy, R. et.al. Automatic metadata enrichment in news production. 10th Workshop on Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09.

Rainer Simon, Bernhard Haslhofer, Joachim Jung. Annotations, tags and linked data. Metadata enrichment in online map collections through Volunteer-Contributed Information. e-Perimetron, Vol. 6, No. 3, 2011 [129-137]

Kevin Clair. Metadata for a Micro-services-based Digital Curation System. Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

Jason Kucsma, Kevin Reiss, Angela Sidman. Using Omeka to Build Digital Collections: The METRO Case Study. D-Lib Magazine, Volume 16, Number 3/4, March/April 2010

Dimitris Gavrilis, Costis Dallas, Stavros Angelis. A Curation-Oriented Thematic Aggregator. Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science Volume 8092, 2013, pp 132-137