Title	Intermediary technical report regarding Europeana - CLARIN integration
Version	1
Author(s)	Twan Goosen *(CLARIN ERIC)*
Date	2017-03-22
Status	Final
Distribution	CLARIN (SCCTC), Europeana
ID	CE-2017-0996

Summary

CLARIN's task in Europeana DSI-2 concerns the dissemination of Europeana's cultural heritage data through third-party infrastructures. Within CLARIN's infrastructure the OAI-PMH standard is used to retrieve metadata for further processing, with the aim to make resources discoverable and processable through a user-friendly workflow of finding applicable processing tools. Through experiments and development towards fulfilling the objectives of this task, it has become apparent that metadata available through other channels is not included in the records that can be obtained from Europeana's OAI-PMH provider. CLARIN recommends unifying their metadata sources and making all relevant information available by means of the widely adopted OAI-PMH protocol.

1 Introduction

CLARIN is responsible for task 2.6.3 of DSI-2: "Data sharing with third parties". As phrased in the project's description of action, this task "aims to share Europeana data with third-party infrastructures, for example, CLARIN, DARIAH, CENDARI, EU-DAT and EHRI." CLARIN aims to implement the objectives of this task by harvesting metadata from Europeana, convert it to the CMDI metadata format which is supported within CLARIN's infrastructure, thus allowing for further exploitation (discovery, linguistic processing) within the existing CLARIN infrastructure[1]. This solution depends on a minimal set of relevant information being present in the harvested metadata; for full integration into the CLARIN infrastructure as envisioned, relevant information includes (in order of priority) basic descriptive metadata about the cultural heritage object, links to machine processable (e.g. full text) resources and technical metadata concerning these resources, most importantly data type in the form of an IANA media type[2].

While this requirement is currently met to a degree that allows for the realisation of 'proofs of concept' for the envisioned integration, it is not yet sufficiently met to be able to fully implement this integration. The issue is described in detail in the following sections, along with recommendations for addressing the described issue and a description of presently attempted ways of mitigating it.

[1] A detailed description is available in the work plan for this task.
[2] Also referred to as "MIME type"

2 Issue: omission of technical metadata

Europeana provides a set of related APIs for searching and retrieving resource metadata (as opposed to retrieval on basis of resource content). The Media search API provides the option to filter by availability of "full text" media, which is very useful in the context of the kind of integration CLARIN aims for in their current task. Metadata for a given record can be retrieved through the Record API; in addition to descriptive metadata (title, creator, location, time span etc.), and properties of Europeana's aggregation process, the returned structure also includes references to digital representations of the object ("WebResources") and, depending on the resource, **technical metadata** such as file size and media type.

As described above, direct links to full text resources and exact file type information are a prerequisite for carrying out a full discovery and processing scenario within CLARIN's infrastructure. However, CLARIN's infrastructure is set up to obtain metadata over OAI-PMH. In line with the objectives of CLARIN's task in DSI-2, the current aim is to integrate Europeana data into CLARIN's existing ecosystem. Since Europeana already provides metadata over OAI-PMH, which is a widely-supported standard for the distribution of archival metadata, this provides a natural interface for CLARIN to obtain Europeana metadata. The XML response can be transformed to a format suitable for further processing in a straightforward manner using XSLT.

Unfortunately, the metadata that Europeana provides over OAI-PHM does **not fully reflect** the extensive metadata that is available from the Record API. In particular, the technical metadata required by CLARIN is not present in the harvested metadata. While the available descriptive and structural metadata can be used to make resources discoverable (in CLARIN's *Virtual Language Observatory*), the omission of technical metadata **obstructs the possibility of automatically discovering applicable tools for further linguistic processing** (through the *Language Resource Switchboard*), which depends on the availability of information about the content language and media type.

2.1 Recommendations

CLARIN recommends unifying the metadata exposed through the Record API and the OAI-PMH provider and include technical metadata (i.e. the properties from the 'ebucore' vocabulary) in the serialised EDM in the OAI-PMH results.

2.2 Workarounds

The envisioned integration into the CLARIN ecosystem can be demonstrated by enriching selected metadata records obtained over OAI-PMH, either manually or through a script, with the additional (technical) metadata exposed by the Record API. Even if this is automated, this is not considered to be a sustainable solution by CLARIN since it introduces a dependency on the Record API and the exact format in which the required metadata is presented.

3 Conclusion

The current state of Europeana's public services allows CLARIN to partially implement its task but not beyond making the metadata and resources discoverable and providing proofs of concept (through demonstration cases) for further integration into the

infrastructure. Achieving the envisioned objective of providing cultural heritage data of interest to CLARIN's community and providing an easy to use gateway to further processing of these resource requires the problem of incomplete metadata over OAI-PMH to be addressed.

References

Goosen, T. (2016). *Europeana DSI-2 task 2.6.3 work plan*. https://www.clarin.eu/file/3932

Lagoze, C., & Sompel, H. V. d. (2001, January). *The Open Archives Initiative Protocol for Metadata Harvesting*. https://www.openarchives.org/OAI/openarchivesprotocol.html

Appendix

Sample response

- Record *92076/BibliographicResource_1000056588419*
  - oai-pmh
    - <edm:WebResource rdf:about="http://resolver.kb.nl/resolve?urn=dpo:11005:mpeg21:pdf"><edm:rights rdf:resource="http://creativecommons.org/publicdomain/mark/1.0/"/></edm:WebResource>
  - Record API
    - {"webResourceEdmRights":{"def":["http://creativecommons.org/publicdomain/mark/1.0/"]},"about":"http://resolver.kb.nl/resolve?urn=dpo:11005:mpeg21:pdf","textAttributionSnippet":"...","htmlAttributionSnippet":"...","ebucoreFileByteSize":23114220,"ebucoreHasMimeType":"application/pdf"}