Europeana DSI 2 – Access to Digital Resources of European Heritage

DELIVERABLE

D4.5 Analysis of IPR implications of other data acquisitions mechanisms and brief for further research

Revision	1.0
Date of submission	31 August 2017
Author(s)	Lisette Kalshoven, Paul Keller (Kennisland)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
0.8	31 May 2017	Lisette Kalshoven	Kennisland	Initial Version
0.9	1 August 2017	Henning Scholz, Julia Fallon, Victor-Jan Vos	Europeana Foundation	Reviewed, added comments
1.0	31 August 2017	Paul Keller	Kennisland	Final Version, resolved comments

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

'Europeana DSI-2 is co-financed by the European Union's Connecting Europe Facility'

Introduction

Europeana is currently reviewing its methods for acquiring data from its Data Providers. This so called "aggregation landscape" has been in place largely unchanged since before 2011 when the Europeana Licensing Framework (ELF) went into effect. This review of the aggregation landscape and possible changes to the methods for acquiring data may have consequences for how the Europeana Licensing Framework operates. Conversely, the way the ELF operates may also impose limitations on new data acquisition mechanisms.

In parallel with the review of the aggregation landscape this document examines the possible consequences of new methods of data acquisition. Given that there is no defined set of new or updated acquisition methods yet, this paper explores likely generic scenarios. This document is intended to flag potential issues at an early stage and may need to be updated as the decision making around new methods for acquiring data advances.

Reviewing Europeana's Aggregation Strategy

Europeana's first strategic priority of the revised 2020 Europeana Strategy is to make it easy and rewarding for Cultural Heritage Institutions (CHIs) to share high-quality content. Until now, Europeana this strategic priority is mainly achieved by aggregating information about digital cultural heritage objects from Cultural Heritage Institutions across Europe. Therefore, aggregation is at the core of Europeana's strategy. Europeana aims to refocus the relationship with cultural heritage institution and aggregation platforms. After some experimentation with Operation Direct and an analysis of what is really feasible in the aggregation landscape, Europeana Foundation has concluded a more complex hybrid strategy, envisioning that by 2020 Europeana will be sourcing content directly from a small number of larger CHIs, as well as indirectly via national and regional aggregators and the expert hubs. By refocusing our relationships onto CHIs directly, and away from third party aggregators, Europeana would, either actively or passively, disintermediate current aggregator partners who are supplying the data on which it critically depends.

This shift away from acquisition methods that rely on aggregators and (at least partially) towards direct sourcing from larger CHIs will likely have implications for the Europeana Licensing Framework. In the short term (by 2018) Europeana expects that it will be able for cultural heritage institutions to directly publish to Europeana, either via push or pull mechanisms. Pull refers to automatic publishing from the websites of the Data

Providers using OAI-PMH. Push refers to allowing the Data Provider to publish on Europeana by using API's provided by Europeana.

The remainder of this paper will explore the consequences of these two ways of direct data acquisition from cultural heritage institution based on generic scenarios. The first one (manual data acquisition) corresponds with the push mechanism in the above quote and the second one (automatic data acquisition) covers the pull mechanism

Direct data acquisition scenarios

These two options of alternative data acquisition will be analysed on a high level. We understand 'manual acquisition' to be a process where a Data Provider directly publishes data on the Europeana platform itself. We understand 'automatic data acquisition' as a process wherein Europeana actively scrapes information of websites (or uses an API to obtain such information). This latter scenario has two sub-scenarios, one where there is a formal relationship with the Data Provider, and the second where there is no such formal relationship.

Assumptions

The analysis in this paper is based on a number of assumptions that are largely derived from the existing technical and institutional arrangements that structure the relationship between Europeana and its Data Providers:

- Europeana continues with the basic principles of the Europeana Licensing Framework. Specifically:
    - All metadata on the Europeana platform must be made available under the Creative Commons 0 Public Domain Dedication.
    - All digital objects published on the platform must be marked with a rights statements from List of Available Rights Statements.
    - The responsibility for clearing rights (obtaining permission to publish metadata and digital objects) lies with the Data Providers who provide such information to Europeana. Under this so-called "clean-hands" policy Europeana is not responsible for rights clearance and relies on the rights information provided by the Data Providers.
- The 'one record, one provider' principle stands, and that there are no clear indications that Europeana intends to combine records about a single digital object from different sources. For example to combine the metadata about the Nachtwacht from the Rijksmuseum with the metadata about the Nachtwacht from the Amsterdam Museum.

- Europeana will continue to work with one legal agreement that structures the relationship between Europeana and its Data Providers (the DEA). All Data Providers will be subject to the same contractual relationship between them and Europeana.

Manual Data Acquisition

In this method of data acquisition the data provider wants to publish data on the Europeana platform. They will be in contact with Europeana. They will be given access to a mapping tool or specialized API (to transform data from their internal format to EDM) where they will get feedback from the tool on whether the data is mapped correctly, and whether they have adhered to the minimum metadata quality standards, such as choosing a rights statement.

When the Data Provider is satisfied with the way the data will look on the platform, they themselves will publish the data (on Europeana) which becomes immediately visible on the platform. Our analysis further assumes that there is no manual check from Europeana on the data before it is published on the platform.

Preliminary Considerations

This method is similar to the method of data acquisition currently implemented. A provider expresses interest in publication, and the provider shared data with Europeana. There are some considerations:

- In order to ensure that Europeana can publish the data under CC0 on the platform, it is required that the Data Provider signs the (current or updated) DEA before publishing on the platform. This would suggest that an agreement needs to be reached *before* the potential Data Provider is given access to the mapping tool, since there are no other checks from Europeana Foundation involved.
- In order to give Data Providers direct access to the Europeana Database, the Europeana Foundation might need to provide these parties with a license with the rights to adapt the Europeana Database.
- Giving Data Providers direct access to the Europeana Database (they can, with the tool, add to the database as well as remove their data as they wish) might grant them database rights over the Europeana Database, as their contribution could constitute as substantial investments.

Since the Data Exchange Agreement currently does not contain any provisions relating to the ownership of database rights in the Europeana Database, the need to add such

provisions to an updated DEA should be explored. However this could also be taken care of in the API terms of use or the terms of use for the mapping tool.

Automatic Data Acquisition

Option 1: formal relationship

In this method of acquisition the Europeana scrapes the data or acquires it via an API from a Data Provider website to be included in the Europeana Database. Europeana transforms the data from the format used by the Data Provider into the Europeana Data Model. Our analysis further assumes that Europeana publishes the acquired data on its platform without intervention of the Data Provider, and without manual checks from Europeana (before publication). Europeana does have a relationship with the Data Provider, and has discussed the scraping of data beforehand.

Preliminary Considerations

- In this scenario it is important to clarify the legal situation of the data before the scraped data is published on the Europeana Platform.
- The most obvious way to do so to have the Data Provider sign the DEA, where the provider would agree to publishing the Data under CC0 and provide a rights statement for each of the digital objects to be published on the Europeana Platform, on their own website before the scraping would begin.
- Although this scenario assumes that the data acquisition is authorized by the Data Provider Europeana, if scraping a substantial amount of the Data Provider website (which could be considered a database) could be required to acquire explicit permission to copy the database via either a license or a waiver of Database Rights. Note that the interactions of Database Rights and Web Scraping are at this point unclear, and will require further research.
- It is possible that this method of acquisition would jeopardise the clean hands policy of the Europeana Foundation, as it is Europeana that seeks out data, selects it, and copies it. This will require further research on this topic.

Option 2: no formal relationship

In this method of acquisition the Europeana scrapes the data or acquires it via an API from a Data Provider website to be included in the Europeana Database. Europeana transforms the data from the format used by the Data Provider into the Europeana Data Model. Our analysis further assumes that Europeana publishes the acquired data on its platform without intervention of the Data Provider, and without manual checks from Europeana (before publication). Europeana has no formal relationship (signed DEA

or otherwise) with the Data Provider, and has not discussed the scraping of the data beforehand.

Preliminary Considerations

- In this scenario it is important to clarify the legal situation of the data before the scraped data is published on the Europeana Platform. This will put a high burden on the Europeana Foundation to:
    - Determine that the metadata it acquires can be published under CC0, either because it is made available under CC0 or a comparable license or because it is only of factual nature and not protected by copyright.
    - Determine that the digital resource referenced has a rights statement, compatible with the list of Available Rights Statements.
    - Determine if the Europeana Foundation is scraping a substantial amount of the Data Provider website that could be considered a database. If this is the case it would need to acquire permission to copy the database via either a license or a waiver of Database Rights (for example by applying CC0 to the database).

The last consideration means that scraping of substantial portions of external websites without prior permission is not legally permissible. The only scenario where automatic data acquisition without prior permission is imaginable is a scenario where the target website is made available under CC0 (or a comparable waiver) or where the data acquisition takes place via an API offered by the target website that explicitly authorises the extraction of substantial amounts of data by third parties such as Europeana

Conclusion

This preliminary review of generic scenarios cannot substitute a more in depth analysis of actual mechanisms that will be employed by Europeana in the near future. The main purpose of our analysis is to identify possible incompatibilities between mechanisms that are currently under exploration and the legal framework provided by copyright legislation and the Europeana Licensing Framework.

The preliminary analysis does not show any such systemic incompatibilities. It does however show that the any scenario that would involve automatic (pull) data acquisition without establishing a formal relationship with the operator of the target is difficult to reconcile with the requirements of the ELF and the legislation concerning database rights.

The fact that the ELF requires that each metadata record contains a rights statement from a controlled list of rights statements will likely pose a substantial challenge to any

effort to automatically ingest data on a large scale. This will only be a realistic option if the target website makes use of machine readable rights statements that are well formatted and have a substantial overlap with the rights statements supported by Europeana. While the Creative Commons licenses (and to a lesser degree the statements offered by rightsstatements.org) are increasingly common on the websites of cultural heritage institutions, the vast majority of rights information is still provided in unstructured formats. As a result the inability to automatically acquire the required rights information will likely render such efforts futile.

Based on the preliminary analysis the manual provision of data (push) is much less problematic and will likely require no or only minimal changes to the ELF (possibly in the form of updated API terms of use, or specific terms of use for a newly developed ingestion tool).