DELIVERABLE

Project Acronym:	**Europeana Cloud**
Grant Agreement number:	**325091**
Project Title:	**Europeana Cloud: Unlocking Europe's Research via The Cloud**

D4.5 Research metadata and content available in Europeana Cloud

Revision: FINAL

Authors:

Marian Lefferts, Consortium of European Research Libraries
Adina Ciocoiu, Europeana Foundation
Anastasia Gasia, Europeana Foundation
Nuno Freire, Europeana Foundation
Peter Vos, VU Amsterdam
Henning Scholz, Europeana Foundation
Nienke Schaverbeke, Europeana Foundation
Els Jacobs, Europeana Foundation

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	P
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	25/1/2016	Adina Ciocoiu	Europeana Foundation	Revised table of metadata contributions
0.1	25/1/2016	Lucas Anastasiou	OU	Confirmation on number of CORE contribution
0.1	19/1/2016	Peter Vos	VU Amsterdam	Uploading data into Europeana Cloud using the API
0.1	27/01/2016	Marian Lefferts	CERL	Putting together all elements into draft deliverable
0.2	29/1/2016	Marian Lefferts	CERL	Executive Summary, Introduction, Conclusion – shared with Europeana Team
1.0	1/2/2016	All authors		Editing
FINAL	4/2/2016	Els Jacobs	Europeana Foundation	Final Edit

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

D4.5 Research metadata and content available in Europeana Cloud

Executive summary

This Deliverable 4.5 gives an overview of the metadata and content processing that took place in the context of the Europeana Cloud project and the workflows that were applied. It presents the state of affairs per 31 January 2016, the original end date of the project. The KPI for metadata ingestion was set at 2.4 million items and this target was met. The KPI for content ingestion was set at 5 million items and this target was not yet met in January 2016. However, close to 5 million newly ingested digital objects will be stored in the Europeana Cloud infrastructure by the time the project formally ends per 30 April 2016, after a three months extension.

This Deliverable 4.5 should be read in combination with Europeana Cloud D4.3 *A report and a plan on future directions for improving metadata in the Europeana Cloud* and D4.4 *Recommendations for enhancing EDM to represent digital content*. The three documents together provide invaluable recommendations for the data management facilities in the Europeana Cloud infrastructure to be further developed. They identify a clear need for guidance on how to support the processing of both the metadata and content in the Europeana Cloud.

The D4.5 reports how the thinking progressed about ingestion workflows, especially those for content ingestion. It presents lessons learnt which will contribute to the future-proof metadata and content ingestion workflow that is currently under development at Europeana.

The report also explains how the project's progressive thinking itself created difficulties in meeting the project's ingestion KPI's. The focus shifted from heritage institutions to domain and national aggregators as the first adopters of the Europeana Cloud infrastructure. Consequently, the heritage institutions, which were involved in the WP4 ingestion efforts, refrained from storing content in the Europeana Cloud infrastructure, also due to their lack of the required technical skills.

The metadata and content ingestion work reported in this D4.5 was part of Work Package 4, which also addressed topics of the contextualisation of data (through linking to external datasets and internal linking within the Europeana dataset), and the processing of datasets in the Cloud infrastructure, with the aid of the Europeana Cloud API.

Table of contents

Introduction

This Deliverable reports on the ingestion of Metadata and Content under the umbrella of the Europeana Cloud project. In the project we aimed to ingest a great variety of data of interest to academics.

While it is therefore natural for ingestion to cover digitised maps, manuscripts, incunabula, archival materials, pamphlets, playbills, dissertations and journals, and visual materials such as portraits, architectural drawings, photographs, images of plaster casts, films and videos, further datasets have also been included for their special relevance to scholars in the Humanities and Social Sciences (the core target audience of Europeana Research).

Oliva, Franciscus, n.d. Portolan Charts. (s.l.): (s.n.) [Marseille, 1650]. Image by the University of Edinburgh, CC-BY

For example, the project ingested the Directory of Open Access Books,[1] which brings together metadata of Open Access books contributed by publishers who publish academic, peer reviewed books. Aggregators can integrate the records in their commercial services and libraries can integrate the directory into their online catalogues, helping scholars and students to discover the books.

The European Library also aggregated the Bielefeld Academic Search Engine (BASE)[2] – one of the world's most voluminous search engines for academic open-access web resources. BASE collects, normalises and indexes data repository servers that use OAI-PMH , and currently supports access to over 60 million documents from over 3,000 sources.

Research organisations DANS[3] and CESSDA[4] also contributed their datasets. DANS provides access to thousands of scientific datasets, e-publications and other research information in the Netherlands, while CESSDA is an umbrella organisation for the European national data archives (including DANS). Its major objective is to provide seamless access to data across repositories, nations, languages and research purposes, and to encourage standardisation of data and metadata, data sharing and knowledge mobility across Europe.

The project aimed to aggregate metadata for digital objects (including the all-important link to the object) but also the actual digital object. Both are stored within the Cloud, with the aim of creating a supportive environment for innovative exploration and analysis of Europe's digitised content. Europeana Cloud is not to become an archive but aims to support active storage and direct access to European cultural heritage content for its manipulation and reuse.

[1] http://www.doabooks.org/
[2] http://www.base-search.net/about/en/
[3] http://www.dans.knaw.nl/nl
[4] http://cessda.net/

Task 4.1 Ingestion of Metadata into Europeana [M1-30]

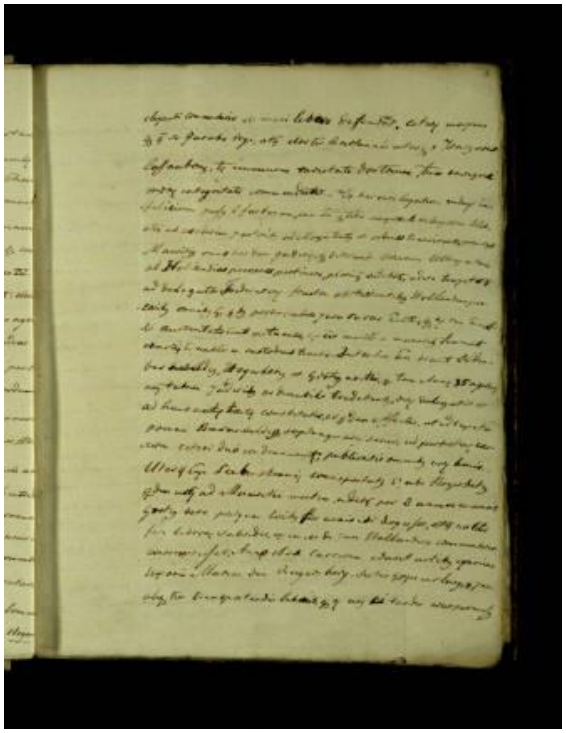
Ingesting Metadata

The project partners' metadata sets were being harvested and processed by The European Library as per the Metadata Ingestion Plan (available via Europeana Cloud website). By the end of January 2016 than 2.4 million meta data records had been processed by The European Library. The full set is available as part of the Europeana Research facility for scholars, academics and researchers that was developed in the project.

However only a limited number of records (67,4%) could be subsequently ingested in the Europeana portal, because the data either a) lacked full and consistent rights labelling, or b) missed direct links to digital contents (see also table below).

The requirements for correct rights information for metadata is made available in the Europeana Publishing Guide.[5] Europeana is committed to the principle that the digitization of public domain content does not automatically create new rights over it. According to the Public Domain Charter 16,[6] works that are in the public domain in analogue form should continue to be in the public domain once they have been digitized. A work is in the public domain when its copyright does not exist or has expired.

Europeana has a clean hands policy and will assume that the data partner has undertaken the correct level of due diligence and labelled the digital objects correctly. However, because it also wishes to help the user and help the data partner to improve and conform to standards, the use of certain rights statements will prompt a manual review during the ingestion process (prior to publication) and Europeana may at this point question some rights statements. This was the case for some datasets provided within the Europeana Cloud project to Europeana.

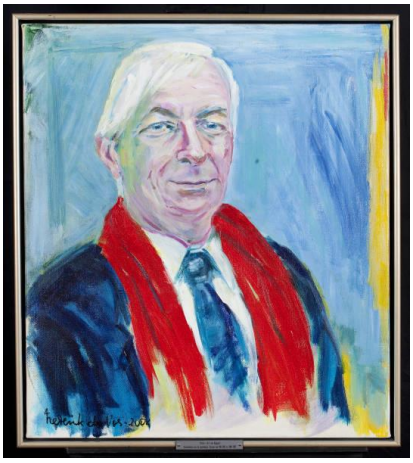


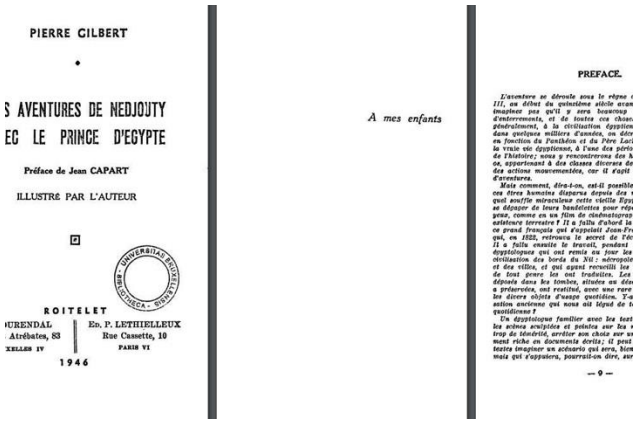
Lecture notes by Johan Melchior Kemper (1776-1824) on Praelectiones in Hugonis Grotii De jure belli et pacis libros III. VU Amsterdam, Heijting, Catalogus hss UBVU, 237

Questioning the rights statements always leads to the start of conversation with the data partner to resolve the issues. In the case of some of the Europeana Cloud datasets (for example, those presented by University College London and the newspapers offered by the National Library of Scotland), this conversation is still ongoing and will not be resolved within the timeframe of the project. Some of the rights issues we discovered are quite complex and involve not only the data partners in the conversation but also legal experts and other authorities in order to resolve them properly. Other rights issues involve another round of due diligence, which takes considerable amount of time in order to be completed. As soon as the issues are resolved publishing into Europeana can proceed.

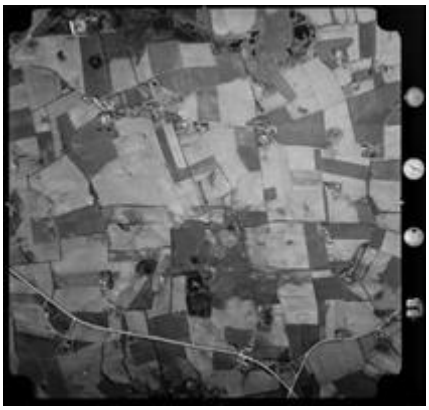
[5] http://pro.europeana.eu/share-your-data/publication/publication-policy
[6] http://www.europeana.eu/portal/rights/public-domain-charter.html

Below is an overview of all datasets processed in the course of the Europeana Cloud project:

	Institution Name	**TEL** *TEL Collection ID (number of records)*	**Europeana** *Europeana Collection ID (number of records)*	**Comments**
1.	OAPEN Foundation (Netherlands)	**6.207** a1139 (2.450) a1193 (3.757)	**6.207** 9200234 (2.450) 9200235 (3.757)	
2.	BASE (Germany)	**758.248** a1141	**0**	Over 36,000 strings with (incomplete/wrong/absent) rights statements that cannot be checked as they are aggregated from many contributors. Proposal to adopt 'unknown' is not acceptable in Europeana as it the user would not have an accurate statement to work with.
3.	DANS (Netherlands)	**31.469** a1161 (26.513) a1162 (2.143) a1163 (2.813)	**0**	Restricted rights records with no previews to digital objects do not conform to the Europeana Publishing Guide and cannot be ingested into Europeana Portal
4.	VU University Amsterdam Library (Netherlands)	**5.237** a1150 (18) a1151 (18) a1152 (749) a1153 (96) a1154 (1.487) a1155 (1.547) a1156 (329) a1157 (292) a1158 (296) a1159 (82) a1160 (229) a1310 (85) a1311 (7)	**5.237** 9200243 (18) 9200245 (18) 9200247 (749) 9200315 (96) 9200246 (1.489) 9200249 (1.547) 9200250 (329) 9200244 (292) 9200407 (296) 9200248 (82) 9200242 (229) 9200371 (85) 9200372 (7)	 *Frederik de Vos (1929), 2002. Rappard, J.H.F. van prof.dr. (1941-). (s.l.): Amsterdam : Vrije Universiteit*

5.	University of Edinburgh (UK)	**16.448** a1169 (346) a1170 (701) a1171 (27) a1172 (1.335) a1173 (430) a1174 (693) a1175 (192) a1176 (53) a1177 (1.061) a1178 (2.702) a1201 (2.001) a1202 (137) a1203 (3.710) a1204 (2.052) a1205 (917) a1206 (71)	**16.436** 9200259 (346) 9200260 (701) 9200261 (27) 9200262 (1.335) 9200263 (430) 9200264 (692) 9200265 (192) 9200266 (53) 9200267 (1.061) 9200268 (2.697) 9200269 (2.001) 9200270 (137) 9200271 (3.704) 9200272 (2.052) 9200273 (917) 9200274 (71)	 *Book of Hours, c. 1430* *University of Edinburgh, MS 39*
6.	Université libre de Bruxelles (Belgium)	**479** a1182 (33) a1183 (172) a1184 (74) a1185 (165) Subsets of a1185: a1185a (133) a1185b (32) a1186 (35)	**468** 9200326 (33) 9200327 (170) 9200328 (74) Delivered and published as subsets 9200329 (128) 9200330 (28) 9200331 (35)	 PIERRE GILBERT S AVENTURES DE NEDJOUTY EG LE PRINCE D'EGYPTE Préface de Jean CAPART ILLUSTRÉ PAR L'AUTEUR ROITELET Ed. P. LETHIELLEUX Rue Cassette, 10 PARIS VI 1946 *A mes enfants* PREFACE. *Gilbert, P., 1946. Les aventures de Nedjouty avec le prince d'Egypte. Bruxelles: Ed. Durendal, 1946*
7.	Dialnet (Spain)	**4.512** a1190 (2.364)	**4.086** 9200362 (2.101)	Not all records have links to digital objects, so these were invalidated by Europeana

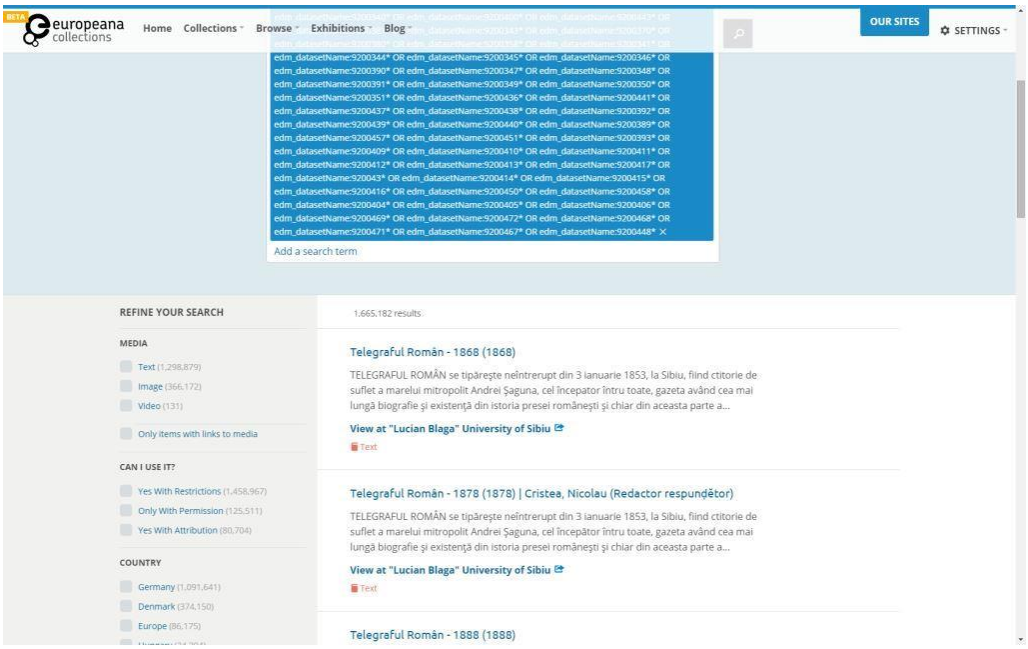
		a1191 (2.148)	9200363 (1.985)	
8.	NL Wales (UK)	**36** a1166 (36)	**36** 9200382 (36)	
9.	Central and Eastern European Online Library (Questa, Germany)	**86.175** a1142 (55.028) a1143 (23.497) a1144 (7.650)	**86.175** 9200418 (55.028) 9200419 (23.497) 9200420 (7.650)	
10	Tilburg University (Netherlands)	**195** a1165 (195)	**195** 9200449 (195)	
11.	Bavarian State Library (Germany)	**1.092.577** a1194 (1.092.577)	**1.091.641** 9200386 (1.091.641)	
12.	"Lucian Blaga" University of Sibiu (Romania)	**274** a1145 (187) a1146 (53) a1417 (34)	**274** 9200340 (187) 9200400 (53) 9200469 (34)	 *Bust (1920): Mitropolitul Andrei Saguna (1808-1873), from the Biblioteca Facultatii de Teologie "Andrei Saguna" din Sibiu*
13.	Library and Information Center, University of Patras (Greece)	**23.983** a1137 (1.407) a1138 (22.576)	**23.983** 9200443 (1.407) 9200381 (22.576)	

14.	NL Denmark (Denmark)	**382.194** a1218 (336.741) a1312 (11.044) a1313 (37.409)	**382.194** 9200343 (336.741) 9200472 (11.044) 9200468 (37.409)	 *Landinspektørens Luftfoto Opmåling, n.d. - 1963-05-07*
15.	National Library of Technology (Prague, Czech Republic)	**228** a1147 (119) a1148 (103) a1149 (6)	**228** 9200370 (119) 9200380 (103) 9200358 (6)	 *Technische Blätter: Wochenschrift für Technik, Baukunst, Industrie und Verkehr, 1869-1921. Prag: J.G. Calve'schen kais. kön. Universitäts-Buchhandlung, Ottomar Beyer 1869 – 1921. Teplitz-Schönau: Verlag technischen Zeitschriften, Jahrg. 1 (1869)-Jahrg. 53, H. 26 (1921)* *Jahrg. 1 (1869)-Jahrg. 53, H. 26 (1921)*
16.	University College London (UK)	**16.757** a1196 (797) a1197 (1.189) a1198 (14.771) a1449 (7.308) a1450 (2.002) a1451 (568)	**797** 9200341 (797)	Disagreement on correct rights label for most of the UCL collections, thus not accepted in Europeana Portal. Conversation ongoing.

17.	Croatian Academy of Sciences and Arts (Croatia)	**24.003** a1167a (393) a1167b (20) a1167c (429) a1167d (3.641) a1167e (69) a1167f (3.963) a1167g (738) a1167h (49) a1167i (527) a1167j (955) a1167k (8) a1167l (1.055) a1167m (4) a1167n (46) a1167o (43) a1167p (871) a1167q (7.395) a1167r (4) a1167s (2.645) a1167t (19) a1168a (30) a1168b (10) a1168c (66) a1168d (78) a1168e (32) a1168f (205) a1168g (164) a1168h	**24.003** 9200344 (393) 9200345 (20) 9200346 (429) 9200390 (3.641) 9200347 (69) 9200348 (3.963) 9200391 (738) 9200349 (49) 9200350 (527) 9200351 (955) 9200436 (8) 9200441 (1.055) 9200437 (4) 9200438 (46) 9200392 (43) 9200439 (871) 9200440 (7.395) 9200389 (4) 9200457 (2.645) 9200451 (19) 9200393 (30) 9200409 (10) 9200410 (66) 9200411 (78) 9200412 (32) 9200413 (205) 9200417 (164) 920043	 *Katalog sedme izložbe umetničke grupe Oblik, n.d. (s.l.): Štamparija Orao, 1932* *From the Croatian Academy of Sciences and Arts - Fine Arts Archives - Exhibition Catalogues*

		(66) a1168i (5) a1168j (60) a1168k (100) a1168l (53) a1168m (261)	(66) 9200414 (4) 9200415 (60) 9200416 (100) 9200450 (53) 9200458 (261)	
18.	The University and National Library of Debrecen (Hungary)	**28.946** a1187 (4.551) a1188 (11.832) a1189 (12.563)	**28.945** 9200471 (4.551) 9200467 (11.831) 9200448 (12.563	 Előadás kezdete ½8 órakor! DEBRECZEN VÁROS CSOKONAI SZINHÁZA Román Sevillai borbély. *Románc: színjáték 3 felvonásban,előjátékkal és utójátékkal - írta Edward Sheldon - fordította Heltai Jenő - rendező Kovács Imre, 1918. Debrecen: Debreczen sz. kir. város könyvnyomda-vállalata*
19.	University of Leuven (Belgium)	**6.851** a1179 (3.128) a1180 (291) a1181 (3.432)	**6.851** 9200404 (3.128) 9200405 (291) 9200406 (3.432)	 *Bon-Secours (Péruwelz). Rue de Condé, ?. Bonsecours: Oeyen*
20.	CESSDA (Sweden)	**2.944** a1314 (2.944)		The legal framework under which CESSDA function does not allow individual archives to expose their collections in Europeana - they would need individual agreements with Europeana, which at this point is not feasible. The collection that was processed for TEL only presence, and was not delivered to Europeana due to reasons mentioned above.
	Total	**2.487.763**	**1.680.754**	

This URL is a query that returns, via the Europeana portal, all collections delivered in this project: http://goo.gl/zluyf0:



In addition, in a direct upload to the Europeana Cloud, The Open University uploaded over 1 million CORE metadata records representing open access research outputs from repositories and journals worldwide. The KPI as stated in the DoW – 2.4 million metadata records ingested in Europeana – was therefore met.

Task 4.2 Ingestion of Content for Research directly into the Cloud [M12-36]

The project's content ingestion plan for 2015, as presented in the YR2 Annual Report, focussed on three of types of potential content providers (project partners, newspapers from Associate Partners in the Newspaper project, and institutions external to the project), combined with a number of workflows (a tool developed by The European Library, a direct upload of data into the Europeana Cloud, and the use of the Europeana Cloud API). In the course 2015, these plans underwent considerable revision. The paragraphs below outline the various approaches taken with the three types of content providers.

1. Project partners

In September 2014, all partners contributing metadata to the project were asked whether, additionally, they would be interested in contributing content to the project. Six project partners (OAPEN, National Library of Technology Prague, UCL with the Bentham Project, the National Library of Wales, VU Amsterdam, and Istituto Luce – Cinecittà) indicated that their content would be available for ingestion.

With the exception of Cinecittà, the metadata for these database had already been processed by The European Library and had mostly been integrated in the Europeana dataset (see section 4.1

above). Cloud WP2 had requested that all providers of content use the Cloud API to upload content into the Cloud.

There were therefore two options:

Workflow 1
The data provider to use the Cloud API to upload both metadata and content from their institution's systems to the Europeana Cloud infrastructure. This would mean that the metadata was in the data provider's native format, and neither converted to EDM nor enriched as per the normal The European Library processes. If the data was not in EDM, it could not be re-used in the Europeana portal and other channels for dissemination of content. This workflow was therefore disregarded.

Workflow 2
Migrate the relevant metadata as held by The Europeana Library to the Cloud. In a separate process the data provider to use the Cloud API to upload their content. Together Europeana Foundation and the content provider to connect the meta data and the content, so that in the Cloud both data sets were properly linked. It became clear that the representation of structured relationships between metadata and content was not a simple issue and required further investigations from WP4. This became the topic of a Research and Development activity the results of which are published as D4.4 **'**Recommendations for enhancing EDM to represent digital content.**'**

It was only possible to investigate the second workflow with the two project partners (VU Amsterdam and NL Wales), who had the necessary technical expertise and Person Months available to participate in the R&D activity, published as Cloud D4.4. Of necessity, this did not result in substantial numbers of content items being added to the Cloud.

1.1. OAPEN

In the D4.2 Content Ingestion Plan, it was explained that The European Library would use its general ingestion pipeline for the ingestion of metadata, augmented with a plugin that downloaded content from embedded links and stored it directly into Europeana Cloud. For this, The European Library planned to implement an additional UIM plugin and connect The European Library infrastructure to Europeana Cloud.

The person with the necessary skill set to undertake this work, Markus Muhr, left The European Library on 31 March 2015. As stated in D 4.2 the workflow described here 'would be only temporary until The European Library will use a new ingestion pipeline completely based on Europeana Cloud. […] this would be heavily aligned with Europeana developments (METIS) currently underway' (now due mid 2017). The loss Mr Muhr's specific skill sets, the conversations about METIS, and the shaping of the Cloud infrastructure, led to a move away from The European Library ingestion pipeline, and its plug-ins, previously described in D4.2.

The Europeana Ingest team and Cloud Work Package 2 are working on an overhaul of the European ingestion workflow, based on the Cloud infrastructure.

Markus Muhr has tested The European Library plug in with OAPEN data in an early prototype of Europeana Cloud with success. When Europeana Cloud reached the Beta stage, this work was not continued after his departure, for reasons explained above. In order to contribute content to Europeana Cloud, the colleagues at OAPEN resolved to work with the Cloud API. The API became available early December 2015, and due to other commitments OAPEN was unable to start testing until 13 January 2016. In spite of the best efforts from OAPEN and the technical support from Europeana Cloud, the OAPEN content was not successfully uploaded, due to several technical

difficulties and documentation shortcomings identified. A detailed description of the issues faced is provided in Appendix I, as an important source of user experience for the final version of Europeana Cloud currently in preparation. Although, OAPEN content was finally not uploaded into the Cloud, the metadata is of course still part of the dataset of The European Library and there available for reuse.

2. Newspapers

In the D4.2 Content Ingestion Plan, we described that several associate partners of the Europeana Newspapers project had not been able to aggregate their full-text/digitized newspapers under the umbrella of that project. The Cloud project subsequently aimed to load the data for the National Libraries of Wales, Luxembourg, Spain, Belgium, Iceland and Scotland in the course of the Europeana Cloud project.

After an initial show of interest, the **National Libraries of Luxembourg and Spain** failed to respond to further requests for data. The **National Library of Wales** became part of the Research and Development work published in D4.4 – the work on making use of IIIF to present this data holds a great promise for a much better user experience. The **National Library of Scotland** experienced rights issues, just like University College London (described above) and other UK institutions. Europeana will take a collective approach but this is a UK issue, and it was not resolved within the time frame of the project.

The datasets for the **National Libraries of Belgium and Iceland** were processed. The dataset of the Royal Library of Belgium (KBR) resulted in 17,129 metadata records, and 135,330 thumbnails. Full text will be indexed to the TEL portal but will not be searchable. As the KBR image server is not fully supported by the TEL portal the images are not visible in the Newspaper browser that was developed for the Newspaper project.

302,172 records for full text issues[7] from the National and University Library of Iceland (BOK). have been processed by The European Library. As a result of incompatibility between the metadata formats, TEL is not able to publish the full text to the newspaper browser. However, just like the data from the Royal Library of Belgium, the data from the National and University Library of Iceland will be migrated to the Cloud infrastructure and will then be available via the Cloud API in the same way as the newspaper collections that were part of the Europeana Newspaper project.

3. Institutions external to the project

The Description of Work set an ambitious target of 5 million objects to be added to an infrastructure to be developed in the course of the project, sourced from Europeana partners (the DoW explicitly mentions institutions outside the project) that had yet to fully understand the benefits of collaborating in the Cloud infrastructure.

From the deliverables prepared in WP5, it is clear that when the project ended the Cloud infrastructure was ready to be used by the three project partners: Europeana Foundation, The European Library and the PSNC, but that additional partners would come in after the close of the Europeana Cloud project, with a limited number of large aggregators joining the Cloud infrastructure in 2016 under the umbrella of the DSI project.

[7] 303,172 digital objects (number of pages is not measurable in the file format available, which is plain text)

This is mirrored in the efforts undertaken in this WP4. The aggregators that are now the target audience and adopters, of the Cloud Infrastructure are not the same as the institutions that were the original intended contributors in this task in WP4. Fifteen to twenty large (and some smaller) libraries and museums, not themselves aggregators, all carefully considered their contribution of metadata and content to the Cloud project, but indicated that joining the Cloud infrastructure at this point in time was not opportune for them.

The Cloud infrastructure was conceived of, designed and development in the course of this project, and a short project extension of three months was needed to finalise the work on the infrastructure. Adding large quantities of content was at various times technically not possible, or premature. When the cloud was migrated to a new version the OAPEN and OU data that had already been processed were not migrated to the Alpha version. Currently, uploading data from institutions in the Europeana network that are not themselves capable of preparing data in EDM, means uploading data that cannot be reused in the Europeana portal and its other channels. Therefore, for institutions that cannot deliver EDM and are not aggregators themselves, the revised Europeana workflow that is currently under development must be the preferred route. It is expected that this workflow, including the much necessary data enrichment that the Europeana Ingestion team is able to offer, will work for The European Library and Europeana by the end of 2016 and be opened up by mid-2017.

4. Returning to our project partners

We did not want to end the project with at least an attempt to add a sizeable set of content items to the European Cloud infrastructure, to test the system and processes. For this we have returned to our project partners, the Europeana Foundation, Open University and the PSNC.

The **Open University** is responsible for CORE (core.ac.uk). The mission of CORE (COnnecting REpositories) is to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public.

At 31 January 2015, the Open University had uploaded a total of 1,197,317 records (and was still uploading further records at a slow rate). The records consisted of up to 5 representations:

1. An oai-pmh xml representation of the original record (this representation appears always)
2. An enriched json representation of the above metadata (this representation appears always)
3. The .pdf file of the article (where available) (this representation appears for 147,964 records)
4. The extracted text (if the above pdf was available) (this representation appears for 142,840 records)
5. A thumbnail preview of the article (if pdf was available) in png format (this representation appears for 123,381 records)

The **PSNC** worked towards the ingest of 1.880,953 thumbnails. The tool for the migration of PSNC data to Europeana Cloud was completed in the last week of January 2016. Migration started in February 2016.

The Europeana Foundation migrated 10-12TB of Newspaper images. The disks were sent to PSNC in the first week of January, and were being uploaded in February/March 2016.

Recommendations

The work in this work package covered metadata ingestion, content ingestion, contextualisation of data (internally and through linking with external data), and processing of data sets in the Cloud infrastructure. All these tasks in the Work Package together have shown there is a clear need for guidance on how to support the processing of both the metadata and content in Europeana Cloud.

To support the process of uploading and structuring of hierarchical datasets over time is not a simple issue and was further explored in D4.3 'A report and a plan on future directions for improving metadata in the Europeana Cloud.' In that report DCG on behalf of CERL, evaluated the use of the Cloud for purely metadata-based applications such as CERL's Heritage of the Printed Book database. This is a repeat of their conclusions published in D4.4:

a) 'The core functionality of Europeana Cloud is storing data. All other aspects must currently be dealt with on the client side. Purely metadata-based applications, such as the HPB would not benefit from data storage that does not also offer means to access the content of a record itself.
b) Functionality that would require client-side implementation includes version control on other levels than the on the individual record level, particularly versions of datasets. Adding this functionality would be recommended.
c) Handling of hierarchically structured metadata and more extensive metadata management options (as described above) might be useful.
d) In the context of the HPB, user rights management would be far easier if it could also be applied on the level of datasets, which would free client implementations from this task. '

As a use case for task WP4.4 the VU University Library, like DCG, used the Cloud API to upload content of two collections to the Europeana Cloud infrastructure. The goal of the test was to have a data provider upload content and metadata to Europeana Cloud which can be picked up by an aggregator (TEL), enriched with EDM and used in the Europeana portal. Peter Vos, VU Amsterdam, prepared a report on his use of the Europeana Cloud API which is made available in Appendix 2.

He notes that 'the eCloud API is very low-level and within the basic CloudId, representation, version structure you can store content any way you like. To facilitate re-use of content in eCloud rules and guidelines have to be implemented about how to store content while still retaining maximal flexibility. It is also a good idea to upload a metadata file describing the Cloud record itself: which representations were used, technical data about the files, etc.

Initially eCloud will be used by aggregators and the Europeana portal, so in principle data providers will not upload and administer their own content. Data providers have to agree on CC0 for the metadata records, but in the majority of cases CC0 isn't suitable for content. This raises technical issues: the provider sets the permissions. Regardless of the way content is harvested in the future I [i.e. Peter Vos] would recommend every data provider has its own providerId in Europeana Cloud. This will ensure the data provider and not the aggregator has ultimate control over the content.'

This report shows that the metadata and content ingestion work reported here took place in the context of the larger Work Package 4, which also addressed topics of the contextualisation of data and the processing of datasets in the Cloud infrastructure, with the aid of the Europeana Cloud API. This Deliverable 4.5 should therefore be read in combination with Europeana Cloud D4.3 *A report and a plan on future directions for improving metadata in the Europeana Cloud* and D4.4 *Recommendations for enhancing EDM to represent digital content*. The three documents together provide invaluable recommendations for the data management facilities in the Europeana Cloud infrastructure.

(Appendix 1)

Testing Europeana Cloud – OAPEN

Introduction

As part of its mission to enhance Open Access publishing of monographs, the OAPEN Foundation hosts the OAPEN Library. The OAPEN Library contains freely accessible academic books, mainly in the area of Humanities and Social Sciences. OAPEN works with publishers to build a quality controlled collection of Open Access books, and provides services for publishers, libraries and research funders in the areas of dissemination, quality assurance and digital preservation. OAPEN works already with Europeana, and is always interested in finding new venues to disseminate monographs. This is why we have chosen to participate in testing Europeana Cloud.

Test setup

As a proof of concept, Ronald Snijder had planned to do the following:

1. Test the connection
2. Create a provider
3. Upload up to five books and upload the corresponding metadata

The test was performed on a Windows 7 PC, using the cUrl program. For guidance, the API manual was used: https://github.com/europeana/Europeana-Cloud/wiki/Europeana-Cloud-API#using-the-api

Result

The results were not good. Only the first task could be completed, before the colleagues at OAPEN ran out of time to execute further tests.

Ad 1. Test the connection. After obtaining login credentials, it was possible to make a connection to the test-server, using this command:

```
curl -X GET -H "Content-Type: application/json" --user ronald_snijder:Jh7ax1UDUL
-d  -i https://195.216.97.96/api/data-providers -k -o "c:\temp\output_bat.xml" --
create-dirs
```

Ad 2. Creating a provider did not work. Not by using the address listed in the manual, and not by using the addresses provided. Therefore, the third goal could not be reached.
Below is the result:

12 January 2016:

```
Microsoft Windows [versie 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Alle rechten voorbehouden.

C:\Users\Ronald>curl -X POST -H "Content-Type: application/json" --user
ronald_snijder:Jh7ax1UDUL -d "{"organisationWebsiteURL":
"oapen.org","officialAddress" : "r.snijder@oapen.org","digitalLibraryURL" :
"oapen.org","organisationName": "OAPEN"}" -i https://195.216.97.95/api/data-
providers?providerId=testOAPEN -g
-k
HTTP/1.1 503 Service Temporarily Unavailable
Date: Tue, 12 Jan 2016 21:08:21 GMT
Content-Length: 323
Connection: close
Content-Type: text/html; charset=iso-8859-1
```

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0//EN">
<html><head>
<title>503 Service Temporarily Unavailable</title>
</head><body>
<h1>Service Temporarily Unavailable</h1>
<p>The server is temporarily unable to service your
request due to maintenance downtime or capacity
problems. Please try again later.</p>
</body></html>

C:\Users\Ronald>
```

13 January 2016:

```
Microsoft Windows [versie 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Alle rechten voorbehouden.

C:\Users\Ronald>curl -X POST -H "Content-Type: application/json" --user
ronald_snijder:Jh7ax1UDUL -d "{"organisationWebsiteURL":
"oapen.org","officialAddress" : "r.snijder@oapen.org","digitalLibraryURL" :
"oapen.org","organisationName": "OAPEN"}" -i https://cloud.europeana.eu/api/data-
providers?providerId=testOAPEN -g
-k
HTTP/1.1 404 Not Found
Date: Wed, 13 Jan 2016 20:24:01 GMT
Server: Apache/2.4.6 (CentOS) OpenSSL/1.0.1e-fips
Cache-Control: no-cache, no-store, max-age=0, must-revalidate
Pragma: no-cache
Expires: 0
Strict-Transport-Security: max-age=31536000 ; includeSubDomains
X-Content-Type-Options: nosniff
X-Frame-Options: DENY
X-XSS-Protection: 1; mode=block
Content-Type: application/xml
Content-Length: 145

<?xml version="1.0" encoding="UTF-8" standalone="yes"?><errorInfo><details>HTTP
400 Bad Request</details><errorCode>OTHER</errorCode></errorInfo>
C:\Users\Ronald>curl -X POST -H "Content-Type: application/json" --user
ronald_snijder:Jh7ax1UDUL -d "{"organisationWebsiteURL":
"oapen.org","officialAddress" : "r.snijder@oapen.org","digitalLibraryURL" :
"oapen.org","organisationName": "OAPEN"}" -i https://195.216.97.95/api/data-
providers?providerId=testOAPEN -g -k
HTTP/1.1 404 Not Found
Date: Wed, 13 Jan 2016 20:24:26 GMT
Server: Apache/2.4.6 (CentOS) OpenSSL/1.0.1e-fips
Cache-Control: no-cache, no-store, max-age=0, must-revalidate
Pragma: no-cache
Expires: 0
Strict-Transport-Security: max-age=31536000 ; includeSubDomains
X-Content-Type-Options: nosniff
X-Frame-Options: DENY
X-XSS-Protection: 1; mode=block
Content-Type: application/xml
Content-Length: 145

<?xml version="1.0" encoding="UTF-8" standalone="yes"?><errorInfo><details>HTTP
400 Bad Request</details><errorCode>OTHER</errorCode></errorInfo>
```

```
C:\Users\Ronald>curl -X POST -H "Content-Type: application/json" --user
ronald_snijder:Jh7ax1UDUL -d "{"organisationWebsiteURL":
"oapen.org","officialAddress" : "r.snijder@oapen.org","digitalLibraryURL" :
"oapen.org","organisationName": "OAPEN"}" -i https://195.216.97.95/api/data-
providers?providerId=testOAPEN -g -k
HTTP/1.1 404 Not Found
Date: Wed, 13 Jan 2016 20:24:36 GMT
Server: Apache/2.4.6 (CentOS) OpenSSL/1.0.1e-fips
Cache-Control: no-cache, no-store, max-age=0, must-revalidate
Pragma: no-cache
Expires: 0
Strict-Transport-Security: max-age=31536000 ; includeSubDomains
X-Content-Type-Options: nosniff
X-Frame-Options: DENY
X-XSS-Protection: 1; mode=block
Content-Type: application/xml
Content-Length: 145

<?xml version="1.0" encoding="UTF-8" standalone="yes"?><errorInfo><details>HTTP
400 Bad Request</details><errorCode>OTHER</errorCode></errorInfo>
C:\Users\Ronald>curl -X POST -H "Content-Type: application/json" --user
ronald_snijder:Jh7ax1UDUL -d "{"organisationWebsiteURL":
"oapen.org","officialAddress" : "r.snijder@oapen.org","digitalLibraryURL" :
"oapen.org","organisationName": "OAPEN"}" -i https://195.216.97.96/api/data-
providers?providerId=testOAPEN -g -k
HTTP/1.1 404 Not Found
Date: Wed, 13 Jan 2016 20:25:09 GMT
Server: Apache/2.4.6 (CentOS) OpenSSL/1.0.1e-fips
Cache-Control: no-cache, no-store, max-age=0, must-revalidate
Pragma: no-cache
Expires: 0
Strict-Transport-Security: max-age=31536000 ; includeSubDomains
X-Content-Type-Options: nosniff
X-Frame-Options: DENY
X-XSS-Protection: 1; mode=block
Content-Type: application/xml
Content-Length: 145

<?xml version="1.0" encoding="UTF-8" standalone="yes"?><errorInfo><details>HTTP
400 Bad Request</details><errorCode>OTHER</errorCode></errorInfo>
C:\Users\Ronald>curl -X POST -H "Content-Type: application/json" --user
ronald_snijder:Jh7ax1UDUL -d "{"organisationWebsiteURL":
"oapen.org","officialAddress" : "r.snijder@oapen.org","digitalLibraryURL" :
"oapen.org","organisationName": "OAPEN"}" -i https://195.216.97.97/api/data-
providers?providerId=testOAPEN -g -k
HTTP/1.1 404 Not Found
Date: Wed, 13 Jan 2016 20:25:57 GMT
Server: Apache/2.4.6 (CentOS) OpenSSL/1.0.1e-fips
Cache-Control: no-cache, no-store, max-age=0, must-revalidate
Pragma: no-cache
Expires: 0
Strict-Transport-Security: max-age=31536000 ; includeSubDomains
X-Content-Type-Options: nosniff
X-Frame-Options: DENY
X-XSS-Protection: 1; mode=block
Content-Type: application/xml
Content-Length: 145
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?><errorInfo><details>HTTP
400 Bad Request</details><errorCode>OTHER</errorCode></errorInfo>
```

Correspondence between the Data Conversion Group and the PSNC revealed the need for a different HTTP-Client. In the context of Deliverable 4.4, DCG worked with a Chrome Browser Add-On Rest-Tool called DHC8. This required to manually save the SSL-certificate of Europeana Cloud and to manually mark it as trustworthy in the Chrome browser. OAPEN did not have the necessary time and PM to complete another round of testing before the end of the project on 31 January 2016.

[8] http://restlet.com/products/dhc/?utm_source=DHC

(Appendix 2)

Uploading data to eCloud – VU Amsterdam

Author: Peter Vos, ICT Developer VU University Library.

As a use case for task WP4.4 the VU University Library uploaded content of two collections to the eCloud infrastructure. The goal of the test was to have a data provider upload content and metadata to eCloud which can be picked up by an aggregator (TEL), enriched with EDM and used in the Europeana portal.

The collections

The VU collections are stored in a local system: OCLC CONTENTdm. The metadata in CONTENTdm are in a custom format mapped to Qualified Dublin Core. The collections were harvested to TEL in an earlier phase of the project using the OAI-PMH interface.

“Portraits of personalities from the history of Dutch Protestantism, period 1600-1900”
http://imagebase.ubvu.vu.nl/cdm/search/collection/prt
This is a collection of digitized portraits. These are simple objects consisting of a JPEG image and metadata.

“Kunstlicht”
http://imagebase.ubvu.vu.nl/cdm/search/collection/klct
Digitized articles of a journal. The issue records were harvested by TEL, but the CONTENTdm records have subrecords containing the PDF of the article, full-text, and article metadata. So in this case we have a lot of extra data and metadata which is not available in TEL and Europeana.

The process

The goal was to create a fully automated process which can automatically upload (part of) a collection in CONTENTdm to eCloud. For the purpose of this test we create a custom script to download the files from CONTENTdm and upload them to the eCloud using the API.

The content was then matched to the EDM records in TEL.

The eCloud API is a web service so you can use any scripting language you like. For the purpose of this test PHP was used.

Creating a data provider

The first step was creating a dataprovider called **VUAlib** for the VU library. On a Linux command line:

```
curl -X POST -H "Content-Type: application/json" --user
peter_vos:UrieGa6y -d '{"organisationWebsiteURL":
"http://www.ub.vu.nl","officialAddress" :
"p.j.m.vos@vu.nl","digitalLibraryURL" :
"http://imagebase.ubvu.vu.nl/","organisationName": "VU University
Amsterdam Library","remarks": "eCloud testing for
WP4","contactPerson": "Peter Vos"}' -i
https://cloud.europeana.eu/api/data-providers?providerId=VUALib
```

Exporting files and metadata

Using the CONTENTdm API the script downloads the Dublin Core metadata to an XML file, the full resolution jpeg image and a thumbnail. For the simple portraits collection this is sufficient.

In the case of the Kunstlicht Journal articles we run into a problem. In the Dublin Core metadata there is no information about the structure.
We need to add two Dublin Core fields with information about the new records in eCloud:

- At the issue level we add a repeated field `dcterms:hasPart`, containing the cloud URIs of the articles.
- At the article level we add the `dcterms:isPartOf` field containing the cloud URI of the corresponding issue record.

With these two changes we still miss the sequence of the articles. The article metadata contains a field with the pages, but this is a free-form field and hard to parse with a computer. EDM does have a field `edm:isNextInSequence` which we can add to the XML metadata file.

This makes the process of uploading structured content different:

- First create CloudId's for all the issues and articles
- Add the URI's to the XML metadata files.
- Upload the files to eCloud

For Kunstlicht the full text of the article was also uploaded, which could potentially be used for searching in Europeana.

Importing in eCloud

For every file to upload we take the following steps:

Create a CloudId

```
curl -X POST --user user:password -i
https://cloud.europeana.eu/api/cloudIds?providerId=VUALib&recordId=oai:imagebase.ubvu.vu.nl:prt/2331
```

This API request takes to parameters, the providerId and a recordId. The recordId can be any unique identifier you like. We use the OAI identifier of the record in CONTENTdm.
In this way the cloud record can be linked to the record in CONTENTdm, the record in TEL and the record in Europeana.
The result of the request is a cloudId, for example:
`NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGSOFTFUZTJIZJRUM4WIQ`
When re-running the script we can use a GET request to the same URL to find out if a record for this recordId already exists in eCloud. If it does the cloudId is returned and files are added to that record.

Create a new version in a representation

Each type of file (full image, thumbnail and metadata) is part of a representation:

```
curl -X POST "Accept: application/json" --user user:password -H "Content-Type: application/x-www-form-urlencoded" -d 'providerId=VUALib'
 https://cloud.europeana.eu/api/records/NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGSOFTFUZTJIZJRUM4WIQ/representations/presentation
```

The response of this request is a version URL, for example:
https://195.216.97.95/mcs/records/NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGSOFTFUZTJIZJRUM4WIQ\representations/presentation/versions/f21aa370-b47a-11e5-9e07-fa163e60dd72

Add a file to the version

In this step we upload the file to the version.

```
curl -X POST --user user:password -H "Content-Type: multipart/form-data"
-F "mimeType=application/png" -F "data=image.png"
https://cloud.europeana.eu/api/records/NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGS
OFTFUZTJIZJRUM4WIQ/representations/presentation/versions/f21aa370-b47a-
11e5-9e07-fa163e60dd72/files
```

The system stores an MD5 hash of the file, this enables us to check if we already uploaded the file in an earlier run of the script.
The result is a temporary file in eCloud. Since we scripted the process we know the file is correct and we can persist the version:

Persist the version

```
curl -X POST --user user:password  https://cloud.europeana.eu/api/
records/NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGSOFTFUZTJIZJRUM4WIQ/representati
ons/presentation/versions/f21aa370-b47a-11e5-9e07-fa163e60dd72/persist
```

This makes the item file permanent and accessible.

Set permissions

Add this point the records are only accessible to members of the VUALib provider. The simplest permission to set is read access for everyone:
curl -X POST --user user:password -i `https://cloud.europeana.eu/api/records/NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGSOFTFUZTJIZJRUM4WIQ/representations/presentation/versions/f21aa370-b47a-11e5-9e07-fa163e60dd72/permit`
This makes the file publicly available.

The URL:
https://cloud.europeana.eu/api/records/NGCXDUTBK7V2NNN5VI2HXC2227S55VGOGSOFTFUZTJIZJRUM4WIQ will now show the record with all the files in an XML representation.

General remarks

The API is fast and stable, especially after the latest upgrade in December 2015. The technical documentation of the API was sufficient for creating the script.

The general structure of objects in eCloud is well thought out and looks like a solid base to the Europeana infrastructure. It is still a low level system and users need programming skills to store and retrieve content. In the future applications can be built on top of the API to enable organizations and end users to enter and view content via a rich interface.

The authorization levels are implemented in the API and can be set on the level of the version. By default objects are only accessible to the uploading provider. So the provider has control over access the content.

Problems

- Deleting records and files

At the moment of testing it was impossible to delete records or versions. Of course it is obvious deleting of records is a special case and an incorrect file can always be replaced by adding a new version. But especially when testing you tend to create a lot of false records.

- No standard interface for harvesting content

Most library systems have interfaces enabling other systems to harvest metadata, for example OAI-PMH. However our repository, CONTENTdm, does not have a standard interface for the exchange of content.

Possibly the SWORD protocol could be used to harvest content to eCloud and in WP4 we also used a IIF interface to access metadata and images.
It is clear harvesting content from all the different systems of data providers will add extra complexity for aggregators. The method the VU used in this test requires programming skills that will not be available at every institutions.

- Representing structure

The upload of structured content adds an extra difficulty as the structure cannot be adequately represented in (Qualified) Dublin Core. We solved this by adding an EDM field to the metadata during the upload process, but in our case and for most providers EDM will not be used in their repository.

Again this adds complexity and extra work for the provider and aggregator. Of course we could imagine a future where the data provider uses eCloud as its main repository, we could then directly add the necessary EDM during the initial entry of the digitized objects.

Recommendations

The eCloud API is very low-level and within the basic CloudId, representation, version structure you can store content any way you like. To facilitate re-use of content in eCloud rules and guidelines have to be implemented about how to store content while still retaining maximal flexibility. It is also a good idea to upload a metadata file describing the Cloud record itself: which representations were used, technical data about the files, etc.

Initially eCloud will be used by aggregators and the Europeana portal, so in principle data providers will not upload and administer their own content. Data providers have to agree on CC0 for the metadata records, but in the majority of cases CC0 isn't suitable for content.
This raises technical issues: the provider sets the permissions. Regardless of the way content is harvested in the future I would recommend every data provider has its own providerId in eCloud. This will ensure the data provider and not the aggregator has ultimate control over the content.