DELIVERABLE

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

D4.2 Content Ingestion Plan

Revision: Initial Plan February 2015, (to be updated in M30 and M36)

Authors:

Marian Lefferts (CERL)
Adina Ciocoiu (TEL / EF)
Markus Muhr (TEL / EF)
Anastasia Gasia (TEL / EF)
Alastair Dunning (TEL / EF - Review)

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

Revision History

Revision	Date	Author	Organisation	Description
0.1	2015/01/21	M.Lefferts	CERL	Summary of information gathered from partners and institutions outside the project wishing to contribute content Summary of discussions with TEL team re feasible timetable
0.2	2015/01/28	M.Lefferts	CERL	Incorporated suggestions by A. Ciocoiu
0.3	2015/01/29	A.Dunning	TEL / EF	Review by Coordinator
0.4	2015/02/03	A. Gasia M. Muhr	TEL / EF TEL / EF	updated table 1 Added aggregation and ingestion process
Final	2015/02/04	M. Lefferts A. Dunning	CERL TEL / EF	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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Table of Contents

Executive Summary

The plan shows that there are various paths along which content (accompanied by appropriate metadata) is prepared and ingested into the Europeana Cloud.

The project is making use of a combination of existing workflows, which will ensure speedy delivery and the achievement of the target of loading 5 million objects, but will also experiment with new workflows (e.g. ensuring that newly sourced content is properly linked to metadata that had already been processed) and new tools (e.g. the Europeana Cloud API).

At the end of the Ingestion Period, which coincides with the end of the project, relevant information should be available to inform on a scalable and efficient workflow for adding further content (plus metadata) to the Europeana Cloud infrastructure.

Introducing the Content Ingestion Plan

Content is being sourced from three areas:

1. Project Partners

Prior to the writing of this plan, all project partners were asked if they would be willing to share content within Europeana Cloud. The partners listed in Tables 2a and 2b (with the exception of the Swedish National Heritage Board) are those that responded positively to the call

2. Associate Partners from Europeana Newspapers

The Europeana Newspapers project runs from February 2012 to end of March 2015. As part of this, The European Library was funded for the aggregations of metadata and content from 12 Full Partners, and, in an addition to the Description of Work, to aggregate the metadata from 11 Associate Partners.

The funding from Europeana Cloud will allow for the aggregation of more full text from the Associate Partners (who within the funding of Europeana Newspapers only provided metadata). See table 1 for the list of partners

As the list of available content far exceeds what is required for the project, the ingestion team will work with the providers to select a subset of newspapers. This will be done according to chronological and rights concerns, eg selecting public domain newspapers from nineteenth century.

3. Other External Partners

The Europeana Cloud Description of Work identified non-funded partners who may be interested in sharing their content via Europeana Cloud. From these, the Swedish National Heritage Board and the National Library of Scotland have offered to share their content.

Timings

As, per the Description of Work, the ingestion of content is scheduled for the final year of the project. Ingestion of Content is planned on a quarterly basis starting with February 2015 (Q1 2015) and ending on December 2015 (Q4 2015).

For many datasets, the metadata was processed through The European Library earlier in the project, which in 2015 will be followed up by processing the content, and the administrative metadata linking the object with the description. This requires the design of an appropriate workflow to support the process. For the institutions offering newspaper content the same workflows as used in the Europeana Newspaper project will be applied. This started in Q4/2014, when Associate Partners from Europeana Newspaper project were invited to contribute their content to the Europeana Cloud project.

In Q1 2015 the Ingestion Team of The European Library will start contacting all institutions to enquire how we can best obtain a copy of the electronic content, and what is required in terms of metadata to link object to description. Actual ingestion will commence in the latter half of Q1 2015 and will continue in Q2 2015. Q3 and Q4 are left blank, but can be used as buffers should there be problems.

For certain datasets we have to establish a tailored workflow (these are not project partners), an example for this is the Swedish National Heritage Board. They have their own data model, which does not make it suitable for processing via The European Library, but could be uploaded directly into the Europeana Cloud via its API. This would require the Cloud to be in place to ready to receive content – schedule to be confirmed.

The combination of newspaper material and the material offered by the Swedish National Heritage Board (minimum of 1.2 million) are sufficient to reach the stated target of 5 million objects. Indeed the quantity of available newspaper far surpasses the target. In the project we would like to test a number of ingestion paths with a number of different institutions in Europeana's network to be able to develop a scalable and efficient workflow for adding further content (plus metadata) to the Europeana Cloud infrastructure, after the project has ended.

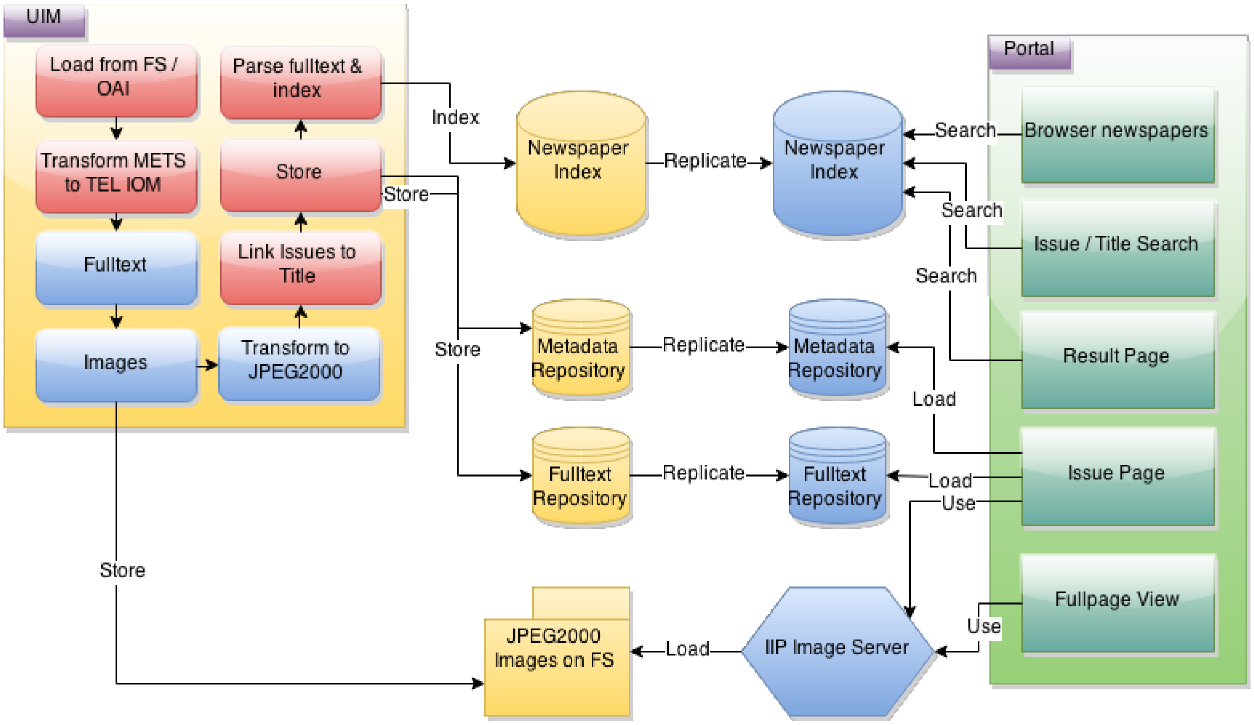
Aggregation and Ingestion Process

The process of ingesting content for Europeana Cloud must be split into two different workflows, one for dealing with the newspaper content and another for dealing with other content.

Workflow for Newspaper content

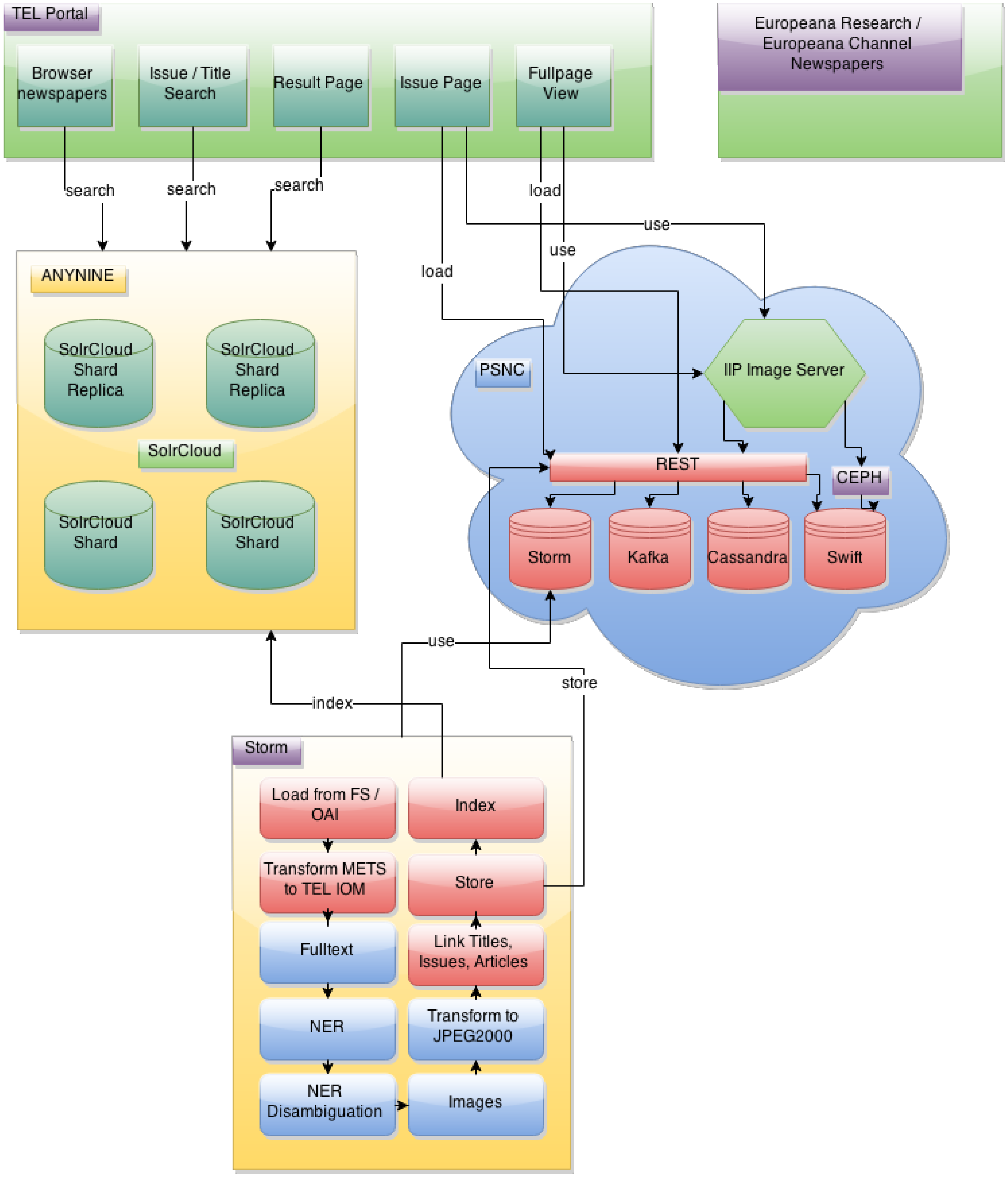
We can reuse the process established in the Europeana Newspapers project, which allows us not only to ingest the data, but also provides a presentation layer embedded into The European Library portal. In the first stage we will still ingest the newspaper content into the local infrastructure of The European Library. However, there is already a migration plan established to migrate the whole newspaper infrastructure to Europeana Cloud and this new content can be migrated as part of the overall plan. After June 2015 a newspapers process will be established that uses Europeana Cloud directly.

Diagram 1: Europeana Newspapers Workflow using The European Library infrastructure



The image shows the general system overview and ingestion process currently in use at The European Library. In our ingestion framework (**U**nified **I**ngestion **M**anager - UIM) we established a workflow that reads either from file or OAI, transforms the original format (mostly METS) to our internal data model (**T**he **E**uropean **L**ibrary **I**nternal **O**bject **M**odel - TEL IOM), downloads or loads the embedded full text links, downloads or loads images, transform the images to JPEG2000, links issues to titles, stores the full text in a full text repository and the metadata in a metadata repository and finally parses the full text (mostly ALTO) and indexes the full-text into a search index. The end user portal is served via a replicated repository, search index and uses an image server to serve images stored on a server in a local file system.

Diagram 2: Europeana Newspapers Workflow using Europeana Cloud infrastructure



For Europeana Cloud, we need to make some adaptations to this process. First of all, current capacity of the image storage at The European Library has been reached, and TEL itself will not host the additional images generated in this activity. We still can use the rest of the process including the presentation through the TEL Portal. However, the presentation will be simpler as we cannot readily provide functionality such as the interactive viewer, which relies on the images being hosted on a server that is integrated in the infrastructure.

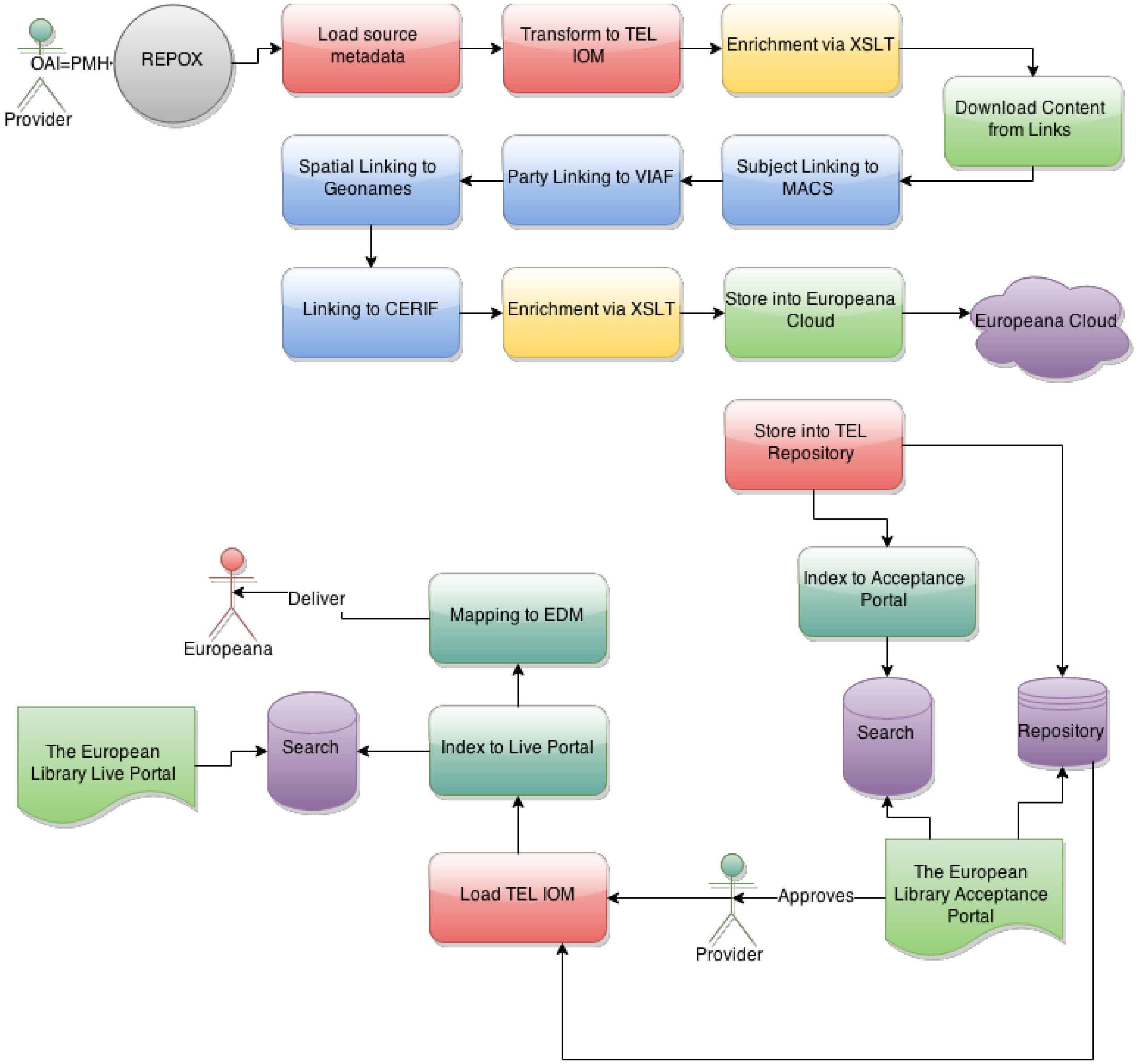
The aim is to migrate the newspapers infrastructure including image storage, so that it will be using Europeana Cloud and AnyNines (a cloud hoster used by the Europeana Foundation) as backend, during the first half year of 2015. The content browser embedded in The European Library will still be the entry point for the content at that moment. The result will be that the new process will be similar to the existing one in use at The European Library, but will in its entirety run on Europeana Cloud infrastructure. There will be an image server service that uses the shared storage layer of Europeana Cloud. The ingestion pipeline will be migrated as a workflow executed by the Europeana Cloud Data Processing Service and the actual search index will be served by scalable, redundant and highly available search service.

Workflow for other content

The European Library can adapt their current metadata ingestion process to automatically upload content into Europeana Cloud. Data thus ingested directly into the cloud (and not via the aggregation process) is by definition not available in the TEL and/or Europeana Portal, but is available for project partners and the Europeana Research platform. The process of how data uploaded directly into the Europeana Cloud can be made available to current end-user services in being explored in Work Packages 1 and 3.

In general, most partners should provide metadata through OAI-PMH to The European Library. The metadata must contain some links to content object that are then downloaded directly by The European Library.

Diagram 3: Other Content Ingestion Workflow using The European Library infrastructure



For other content, we will employ our general ingestion pipeline for metadata, but add a plugin that downloads content from embedded links and stores it directly into Europeana Cloud. For this, we need to implement an additional UIM plugin and connect The European Library infrastructure to Europeana Cloud.

The figure highlights our current general ingestion workflow. It starts by harvesting data from the partner via OAI-PMH (using Repox), and then loading the original metadata. The original metadata is parsed and transformed into our internal data model (TEL OIM). We exploit XSLT as enrichment tool for collection specific enrichments like links to digital objects or rights statements. After that we would download content embedded in the metadata. This would be a new plugin specific for the content ingestion as part of Europeana Cloud.

Afterwards, we run our usual enrichments (MACS, VIAF, Geonames and CERIF Linking). We have an optional second XSLT plugin to perform collection specific updates of the enrichment process. Now we store it into our repository, but a new plugin will be added that also stores into Europeana Cloud. The final step is indexing the data into our acceptance index and serve the collection via our acceptance portal. Once approved by the partner, we publish the dataset on the TEL portal and furthermore deliver the dataset using an automatic EDM mapping to Europeana.

This workflow would be only temporary until The European Library will use a new ingestion pipeline completely based on Europeana Cloud. Since this would be heavily aligned with Europeana developments (METIS) currently underway, we cannot make a specific comment on when and how this will be possible.

Table 1 Ingestion plan for Q1/2015

Provider	**Country**	**Collections**	**Records**	**Digital Objects**	**DoW**
National Library of Wales	United Kingdom	NLW: Europeana Newspaper		Circa 2 million pages of full text	Loaded as a0644 - **Not in DoW**
National Library of Luxembourg	Luxemburg	BNL: Europeana Newspaper		Circa 620,000k pages of full text available	a0639 - **Not in DoW**
National Library of Spain	Spain	Digital Periodicals Library		Circa 4.5m pages of full text available	a0597 - **Not in DoW**
National Library of Belgium	Belgium	KBR: Europeana Newspaper		Circa 700,000 pages of full text available	a0640 - **Not in DoW**
National Library of Iceland	Iceland	Newspapers and Journals - Digitized newspapers and journals from Iceland, the Faroe Islands and Greenland (1773-2001)		Circa 2.8m pages of full text available	a0277 - **Not in DoW**
		Total estimate		**To be confirmed**	

A note on the numbers for newspapers. The European Library and Europeana have both the metadata for the newspapers listed above. This metadata is at the issue-level; ie there is one record for each daily newspaper held at the libraries. The number of pages is calculated by multiplying the number of issue-level records currently held by The European Library by a figure of 9.5. This figure was derived from a sample analysed within the Europeana

Newspapers project, that calculated there was on average 9.5 pages within each historic newspaper aggregated in the project.

Table 2a Ingestion plan for Q2/2015

Provider	Country	Collections	Records	Digital Objects	DoW
OAPEN Foundation	Netherlands	OAPEN		2,200 PDFs	Loaded as a1139
National Library of Technology, Prague	Czech Republic	Historical Monographs Historical Maps Historical Periodicals		119 digitised historical monographs 103 digitised historical maps and atlases, 6 digitised periodicals	Loaded as a1147 a1148 a1149
UCL	United Kingdom	Bentham Project		To be confirmed	a1198
VU Amsterdam	Netherlands	Letters Historical maps Manuscripts Portraits Photographs	292 items 1,456 items 18 items 18 items 7 items 1,574 items 749 items	1,369 images 5,493 images expected 3.223 images 3.107 images 10.678 images 1,628 images 749 images	a1157 a1154 a1150 a1151 a1311 a1155 a1152
NL Scotland	United Kingdom	Early Gaelic Book Collection	1,700 items	450,000 transcriptions	a1302, and http://digital.nls.uk/early-gaelic-book-collections/

					pageturner.cfm?id=75733573
NL Scotland	United Kingdom	Word on the Street Broadside collection	1,770 titles	1,770 full text transcriptions	a1305
		Total estimate		**477,047 digital objects**	

Table 2b Direct uploading, using the Europeana Cloud API

Provider	Country	Collections	Records	Digital Objects	DoW
Swedish National Heritage Board	Sweden	To be confirmed		1.2 – 2.9 million objects are made available	**Not in DoW**
Istituto Luce – Cinecitta	Italy	Newsreels of Italian everyday life in the fifties and sixties Italian a1350		100 film reels[1]	Project Partner

Table 3 Ingestion plan for Q3/2015 **- Buffer for TEL team**

Provider	Country	Collections	Records	Digital Objects	DoW

[1] Subject to file sizes, and ease of transport between server locations

Table 4 Ingestion plan for Q4 /2015 **- Buffer for TEL team**

Provider	Country	Collections	Records	Digital Objects	DoW