

# DELIVERABLE

**Project Acronym:** Europeana Newspapers  
**Grant Agreement number:** 297380  
**Project Title:** A Gateway to European Newspapers Online

---

## D4.5 Report on newspapers data aggregated by The European Library

---

**Revision:** 3.0

**Authors:** Alastair Dunning (The European Library / Europeana Foundation)  
Alena Fedesenka (The European Library / Europeana Foundation)  
Anastasia Gasia (The European Library / Europeana Foundation)  
Markus Muhr (The European Library / Europeana Foundation)

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

## Revision History

Revision	Date	Author	Organisation	Description
1.0	28-02-2014	Alastair Dunning	TEL / EF	Draft
1.1	29-07-2014	Alastair Dunning	TEL / EF	Draft
1.2	30-07-2014	Clemens Neudecker, Sandra Kobel	SBB	Revised Draft
1.3	31-07-2014	Alastair Dunning	TEL / EF	Updated
1.4	31-07-2014	Clemens Neudecker, Hans-Jörg Lieder	SBB	Internal review
2.0	31-07-2014	Sandra Kobel	SBB	Final version
2.1	16-03-2015	Alastair Dunning Alena Fedesenka Anastasia Gasia Markus Muhr	TEL TEL TEL TEL	Updated draft version
2.2	22-03-2015	Neil Fitzgerald Adam Sofronijevic	BL UB	Internal review
2.3	30-03-2015	Clemens Neudecker Hans-Jörg Lieder Sandra Kobel	SBB SBB SBB	Internal review
2.4	13-04-2015	Alastair Dunning	TEL	Following feedback from reviewers
2.5	29-04-2015	Alastair Dunning	TEL	Following request from Commission to present D4.6 and D4.7 as separate from D4.5.
3.0	30-04-2015	Clemens Neudecker Sandra Kobel	SBB SBB	Internal review and final version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

<b>1. Executive Summary .....</b>	<b>4</b>
<b>2. Some Headlines .....</b>	<b>5</b>
2.1 Largest Aggregating Project to Europeana.....	5
2.2 Project has succeed in reaching targets.....	5
2.3 Largest amount of data in Europeana Data Model .....	6
2.4 First project to represent hierarchical data in Europeana .....	6
2.5 Encouraging new strategic directions for Europeana: storing and delivering content for Europeana .....	7
2.6 New technologies for sharing images.....	8
<b>3. The Aggregated Content .....</b>	<b>9</b>
<b>4. The Aggregation Workflow.....</b>	<b>11</b>
4.1 Full-Text.....	11
4.2 Images.....	12
4.3 Metadata.....	12
4.4 Seeing the Metadata Formats .....	13
<b>5. The Resulting Corpus of Data .....</b>	<b>15</b>
<b>6. Conclusion .....</b>	<b>15</b>
<b>Appendices .....</b>	<b>16</b>
I - Content aggregated to The European Library .....	16
II - Content aggregated to Europeana.....	19
III - Breakdown of issues by library .....	21
IV - Licensing Conditions .....	23
V - Components and Data Formats in Project.....	24

# 1. Executive Summary

The online presentation of historic newspapers has been one of libraries' success stories in the era of digitisation. Newspapers are at the centre of the user experience for Trove, the National Library of Australia's digital library.<sup>1</sup> The demand for the British Library's newspapers has been so considerable that they have been able to agree that commercial providers can scan over 40 million pages from their collection.<sup>2</sup> Scholarly demand for the newspapers dataset of the National Library of the Netherlands far outstrips demand for other datasets.<sup>3</sup>

Not all newspapers are easy to digitise but their serial nature and the incredible range of content that they provide for a variety of different audiences (school teachers, family historians, genealogists, historians, linguists etc.) mean that they have been at the forefront of many libraries' digitisation strategies. Various articles have looked at challenges that have accompanied this work. Of particular interest have been the quality of digitisation (examining the quality of Optical Character Recognition and the article extraction) and the task of crowdsourcing, where users contribute correction to automatically transcribed text.<sup>4</sup>

This deliverable addresses a new challenge: That of aggregating such digitised material at a continent wide level and then developing an interface to search through that aggregated data. Aggregation has taken place at a national level, for instance the Chronicling America programme in the United States, Trove as part of the National Library of Australia, Gallica for French newspapers, Delpher in the Netherlands, or the British Newspaper Archive in the UK.

However, the Europeana Newspapers project is the first large scale attempt within the public sector to draw together a critical mass of newspapers across national boundaries. Funded by the European Commission, the project had many related goals.<sup>5</sup> Other issues - such as the OCR (Optical Character Recognition) processing and the development of a METS-ALTO (Metadata Encoding and Transmission Standard - Analysed Layout and Text Object) profile for newspapers are documented in other deliverables in the project.<sup>6</sup>

This deliverable first looks at some of the headline successes of the project. It then continues by addressing the specific challenge of aggregating metadata, full-text and images from over 20 contributing libraries. It is based on the work of The European Library (TEL), a cultural organisation based in the National Library of the Netherlands in The Hague. Within the project, TEL had the role of aggregating the data from the various content providers involved and then developing a newspaper browser for end users. TEL also forwarded the metadata to Europeana, the EU-funded metadata platform for European cultural heritage.

---

<sup>1</sup> Tim Sherratt, Digitised newspapers and the varieties of value, <http://www.slideshare.net/wragge/digitised-newspapers-and-the-varieties-of-value>.

<sup>2</sup> [http://news.bbc.co.uk/2/hi/uk\\_news/england/london/8690919.stm](http://news.bbc.co.uk/2/hi/uk_news/england/london/8690919.stm).

<sup>3</sup> Interview with Steven Claeysens, Data Services Manager, National Library of Netherlands, 20 April 2015. The popularity of newspapers was explored in more depth at the event Symposium: Digitale historische kranten als 'big data' <https://www.kb.nl/nieuws/2015/symposium-digitale-historische-kranten-als-big-data>.

<sup>4</sup> For instance, Tanner et al (2009), Measuring Mass Text Digitization Quality and Usefulness, <http://www.dlib.org/dlib/july09/munoz/07munoz.html> or Holley (2009), How Good Can It Get? <http://www.dlib.org/dlib/march09/holley/03holley.html>.

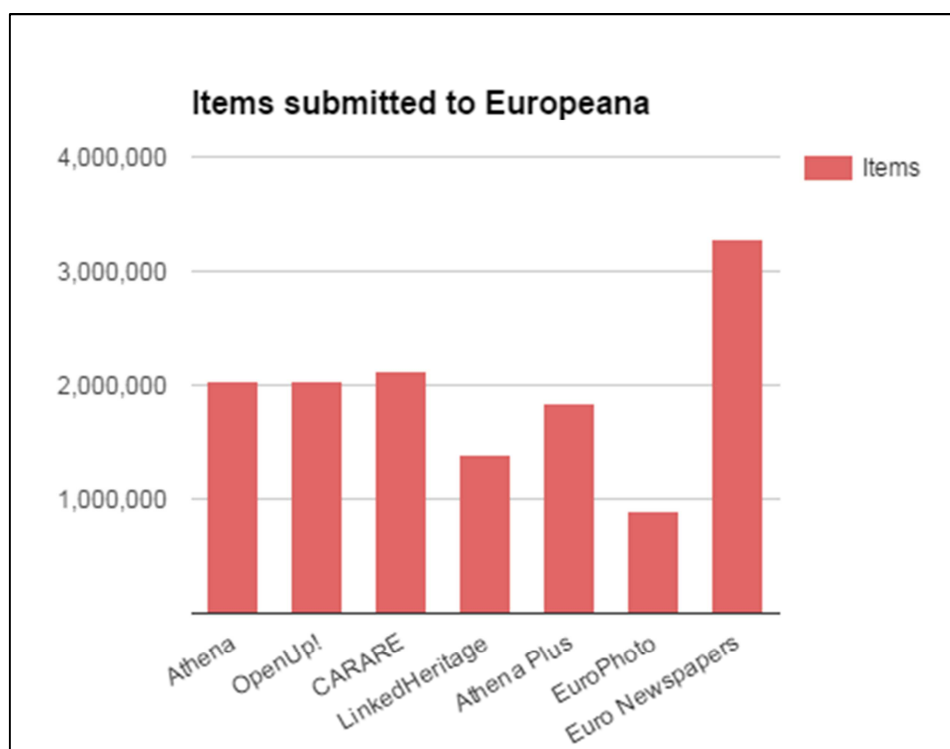
<sup>5</sup> See <http://www.europeana-newspapers.eu>.

<sup>6</sup> Deliverables are available at <http://www.europeana-newspapers.eu/public-materials/deliverables/>.

## 2. Some Headlines

### 2.1 Largest Aggregating Project to Europeana

Europeana Newspapers has provided the largest quantity of metadata records to Europeana for any one single project. The January publication of the Europeana portal included 2.5 million metadata records to Europeana.<sup>7</sup> By March this was above 3 million (full details of this and all statistics related to the project are available in Appendix 1).



*Metadata records submitted to Europeana by project (March 2015)*

### 2.2 Project has succeed in reaching targets

**The project has surpassed the targets set out in the Description of Work.**

Number of Issues (i.e. Metadata Records) - **3,572,194 (estimate based on DoW: 1,874,641)**

Number of Full Text Pages - **11,202,411 (target: 10m)**

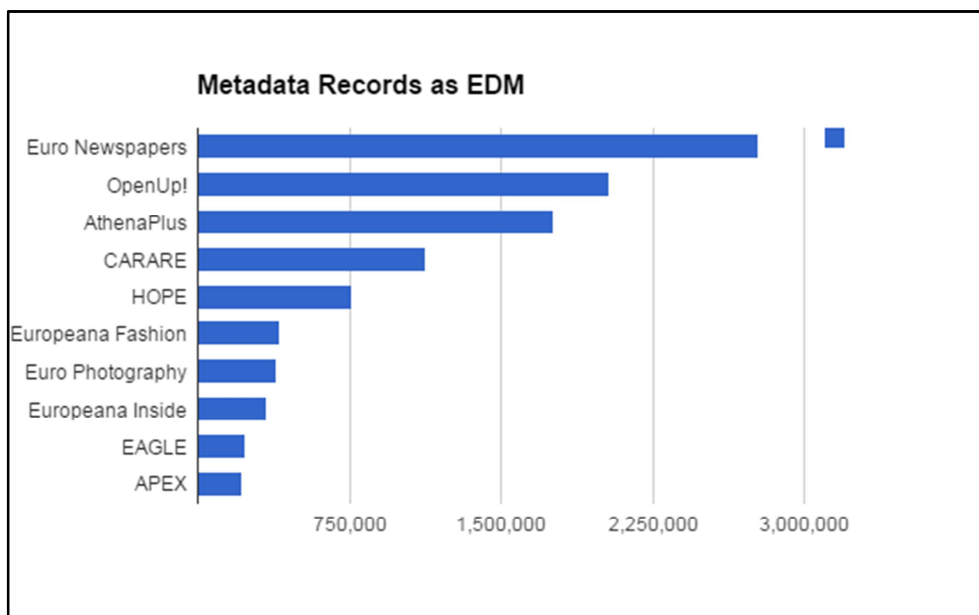
Number of Images - **11,202,411** (includes images harvested on the fly)

Number of Titles - **991**

<sup>7</sup> Figures derived from monthly internal spreadsheet created by Europeana ingestion team.

## 2.3 Largest amount of data in Europeana Data Model

All data aggregated via The European Library has been mapped from the native metadata format to the Europeana Data Model. (The Europeana Data Model is the updated metadata model required by Europeana, replacing the earlier Europeana Semantic Elements)<sup>8</sup>.



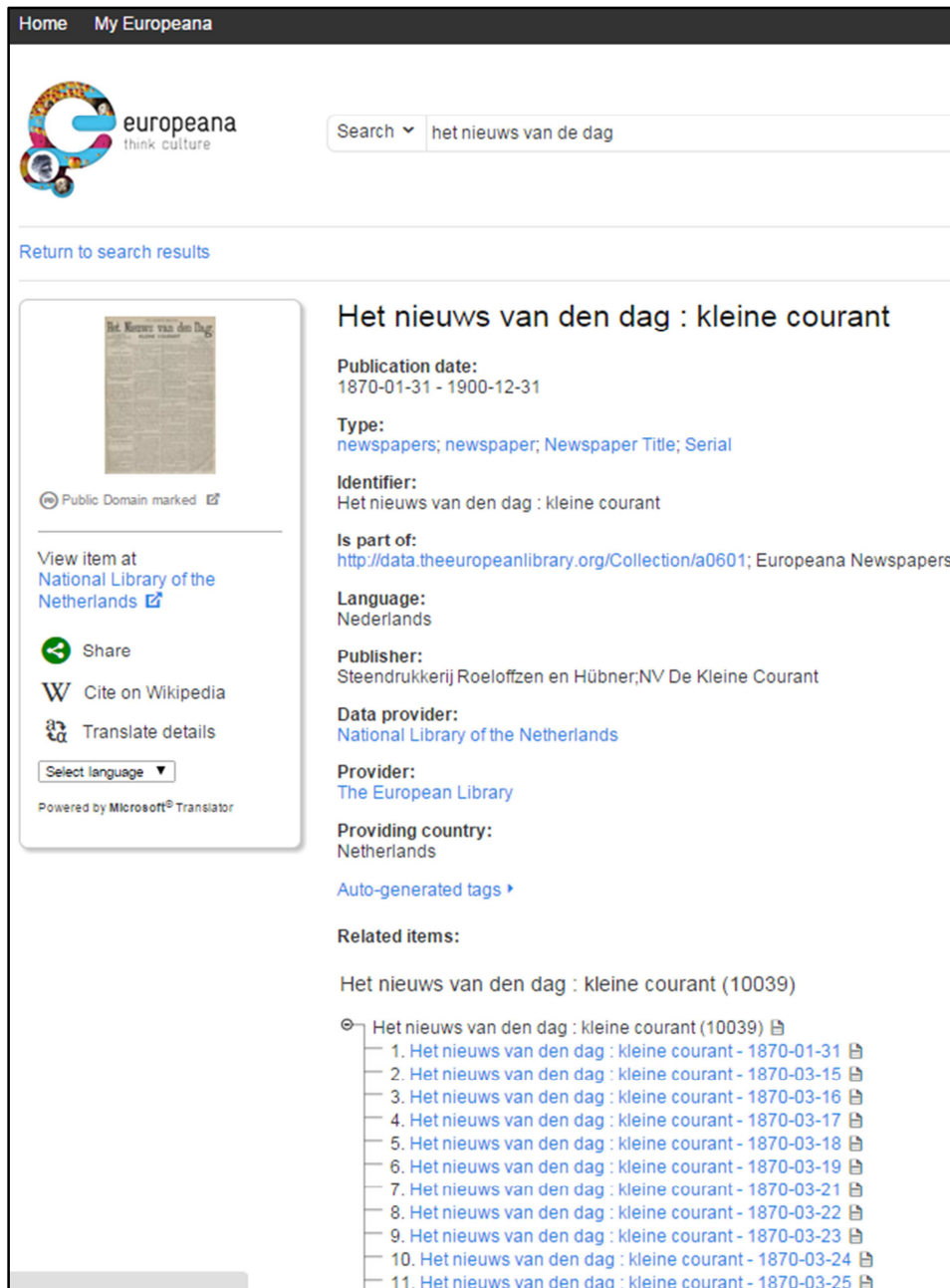
*Metadata records submitted to Europeana in EDM format by project (January 2015)*

## 2.4 First project to represent hierarchical data in Europeana

The sustained use of the Europeana Data Model in the project means that Europeana can **now represent hierarchies between different objects in their portal.**

In the example below, the different issues of one newspaper are represented near the bottom of the record page

<sup>8</sup> <http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation>



Home My Europeana

Search

Return to search results

**Het nieuws van den dag : kleine courant**

Publication date:  
1870-01-31 - 1900-12-31

Type:  
[newspapers](#); [newspaper](#); [Newspaper Title](#); [Serial](#)

Identifier:  
Het nieuws van den dag : kleine courant

Is part of:  
<http://data.theeuropeanlibrary.org/Collection/a0601>; Europeana Newspapers

Language:  
Nederlands

Publisher:  
Steendrukkerij Roeloffzen en Hübner; NV De Kleine Courant

Data provider:  
[National Library of the Netherlands](#)

Provider:  
[The European Library](#)

Providing country:  
Netherlands

[Auto-generated tags](#)

Related items:

Het nieuws van den dag : kleine courant (10039)

- Het nieuws van den dag : kleine courant (10039)
  - 1. [Het nieuws van den dag : kleine courant - 1870-01-31](#)
  - 2. [Het nieuws van den dag : kleine courant - 1870-03-15](#)
  - 3. [Het nieuws van den dag : kleine courant - 1870-03-16](#)
  - 4. [Het nieuws van den dag : kleine courant - 1870-03-17](#)
  - 5. [Het nieuws van den dag : kleine courant - 1870-03-18](#)
  - 6. [Het nieuws van den dag : kleine courant - 1870-03-19](#)
  - 7. [Het nieuws van den dag : kleine courant - 1870-03-21](#)
  - 8. [Het nieuws van den dag : kleine courant - 1870-03-22](#)
  - 9. [Het nieuws van den dag : kleine courant - 1870-03-23](#)
  - 10. [Het nieuws van den dag : kleine courant - 1870-03-24](#)
  - 11. [Het nieuws van den dag : kleine courant - 1870-03-25](#)

Screenshot of 'Het nieuws van de dag'

[http://www.europeana.eu/portal/record/9200359/BibliographicResource\\_3000112763320.html](http://www.europeana.eu/portal/record/9200359/BibliographicResource_3000112763320.html)

## 2.5 Encouraging new strategic directions for Europeana: storing and delivering content for Europeana

This is the **first large-scale project where Europeana has shown images**. Up to now, the Europeana portal has only shown thumbnails of images. Each image aggregated by TEL or a library is available for viewing on the Europeana website, as is the full text.



Screenshot showing full text and image browser in Europeana based on

[http://www.europeana.eu/portal/record/9200359/BibliographicResource\\_3000112763320.html](http://www.europeana.eu/portal/record/9200359/BibliographicResource_3000112763320.html)

## 2.6 New technologies for sharing images

By making use of library partners' own image servers, TEL was able to save on the costs of hosting images of three large libraries (National Libraries of France, Austria and the Netherlands). This technique of requesting images from a third party allows libraries much greater control of their images and allows aggregators to concentrate on design and interface issues rather than spend time on the costly and time consuming transfer of data. This technique, when standardised by the International Image Interoperability Framework, provides a significantly enhanced way of managing and presenting images.<sup>9</sup> **Europeana Newspapers has been one of the first to test (and succeed) in this approach at scale.**

<sup>9</sup> More on the IIIF is available here: <http://iiif.io/>



### 3. The Aggregated Content

Moving now to the details of the project - 23 libraries were featured in Europeana Newspapers; 12 libraries were considered full partners and were sharing their full text, images and metadata with TEL. The planned amounts of data to be shared at the start of the project (January 2012) are listed in the tables below. The first column shows the number of pages to be linked to via Europeana; of these 18 million pages, 10 million would undergo the OCR and OLR (Optical Layout Recognition) processes and, because they were being made available in the public domain, be available for full-text searching.

#### Full Partners

Library	Pages to be linked to from Europeana in DoW	Titles in DoW
National Library of Austria	7,057,772	204
National Library of France	3,000,000	1,200
National Library of Netherlands	2,000,000	not stated
Dr. Friedrich Teßmann State Library	840,082	15
National Library of Estonia	594,657	34
National Library of Latvia	475,000	117
National Library of Finland	114,000	11
State and University Library of Hamburg	1,545,000	6
State Library of Berlin	1,710,000	4
University Library of Belgrade	387,700	13
National Library of Poland	99,106	117
National Library of Turkey	418,345	847
<b>Total</b>	<b>18,241,662</b>	<b>2,671</b>

Quantities of content from original DoW

## Associate Partners

During the lifetime of the project, associate partners joined the project network. 11 of these were gathered during an outreach campaign done jointly by WP4 and WP6 earlier in the project; a further two were suggested by the EU reviewers for the project (the national libraries of Slovakia and Serbia).<sup>10</sup>

Originally, these partners were only joining the project as network partners (i.e. to attend meetings and share best practices), but TEL was also keen to expose their content. It was therefore agreed that these two would also share their metadata with TEL. There were no funds for their content to undergo the OCR or OLR processes, or for the cleaning up of their metadata for smoother integration into TEL and Europeana.<sup>11</sup>

The figures below were the libraries' calculations as to how many newspaper pages they could share. TEL would then aggregate as much metadata as possible related to these pages.

Library	Number of pages digitised and potentially available for aggregation
National and University Library of Slovenia	(not stated but interest shown)
National and University Library of Iceland	4,112,602
Royal Library of Belgium	3,500,000
National and University Library of Zagreb	300,000
St. Cyril and Methodius National Library of Bulgaria	12,000
National Library of Romania	442,398
National Library of Luxembourg	620,000
National Library of the Czech Republic	2,760,000
National Library of Spain	5,831,826
National Library of Portugal	400,000

<sup>10</sup> See the Deliverable 4.1: European Newspaper Survey Report at <http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf>.

<sup>11</sup> Furthermore, during the course of the project, interest was also shown by the national libraries of Denmark, Norway and the Faroese Islands.

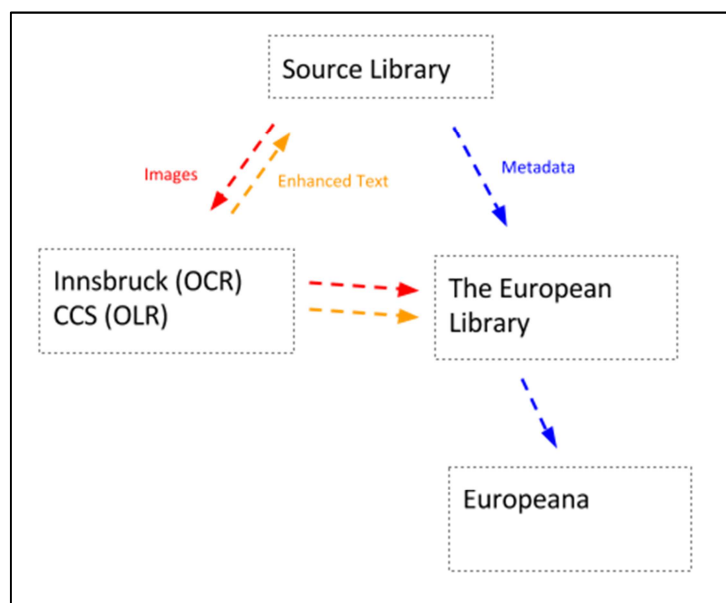
National Library of Wales	1,100,000
National Library of Serbia	(not stated but interest shown)
National and University Library of Slovakia	(not stated but interest shown)

Additional metadata to be made available by Associate Partners

## 4. The Aggregation Workflow

Once the content numbers had been agreed the aggregation workflow could get underway.

As had been agreed at the outset of the Europeana Newspapers project, both OCR (Optical Character Recognition) and OLR (Optical Layout Recognition) workflows were taking place. OCR was taking place on 8 million pages of newspapers of full partners by the University of Innsbruck, while commercial partners CCS (Content Conversion Specialists) were responsible for the Optical Layout Recognition of 2 million pages, i.e. segmenting the digitised pages into their distinct articles. This meant that full text, images and full metadata would often arrive at TEL from different destinations before being all put together again.



A simplified version of the aggregation workflow in Europeana Newspapers

### 4.1 Full-Text

Except for the National Library of Netherlands, which had already undertaken their own newspapers digitisation and refinement (and could therefore provide their data directly to TEL), all the other full partners provided the OCRed full text pages encoded in ENMAP, the Europeana

Newspapers recommended METS/ALTO metadata format, via data sent from the University of Innsbruck and CCS. For the most case, images were also sent to TEL by the same route - some advances in technologies allowed this to be modified later in the project.

Thus, the aggregation plan was dictated by the progress of work at the University of Innsbruck and CCS. Given the quantity of data being transferred every time, the most efficient way of sharing the data was via hard drive - each hard drive contained between 1 and 2 TB of data.

Since the metadata in the METS/ALTO files only included basic information sufficient for post capture processing, additional essential information, such as the links to the original newspapers on the library's own website needed to be added. It was agreed that extra metadata would be harvested individually from each partner (for this process see below).

## 4.2 Images

The digitised images of the newspapers were received along with the METS/ALTO on hard drives. At first TEL received images as TIFF and JPEG and these were converted to JPEG2000. During the project it was then agreed that CCS and UIBK would perform the JPEG2000 conversion (working to an agreed profile) before forwarding data to TEL.

The images that are made available for the viewer are served from the open source software IIP Image Server<sup>12</sup>. It supports multiple protocols for access to images such as IIP and IIIF.

As the project advanced, other opportunities regarding changes in technology allowed TEL to gain access to the images. Three libraries (Austria, France and the Netherlands) decided they could make their image servers available to TEL. Thus the images would not need to be transferred to the TEL servers; rather the newspapers interface would dynamically request each individual image when a user made the appropriate search on the website. As explained elsewhere, this method of showing images brings significant advantage to both the owner and the consumer of the digital image.

## 4.3 Metadata

For metadata being harvested from both Full and Associate Partners, the standard aggregation workflow of TEL was usually followed. Sometimes, due to the hierarchy of the data again it needed to be adjusted in a way that both title and issue-level metadata could be ingested (e.g. for the National Library of Slovenia, 28 newspaper titles were harvested separately and merged afterwards once they had arrived at TEL).

The aggregation of metadata in TEL comprised three steps: the preparation, the ingestion and the publication phase. Within the preparation phase the nature of the content is checked, contact information with the source library is exchanged and if necessary sample data is processed.

The ingestion phase follows. This facilitates the processing and enriching of the metadata, and involves three separate parts: harvesting, mapping, and finally normalisation and enrichment.

---

<sup>12</sup> <http://iipimage.sourceforge.net/>.

Making use of the harvesting tool Repox, the metadata stored on the provider's side is imported to TEL.<sup>13</sup> The preferred transferring method for TEL is via OAI-PMH but other transport mechanisms are supported as well, such as FTP and HTTP, and in general all means of data exchange.<sup>14</sup> Any library format can be ingested.

Once it arrives the data is loaded into the TEL data repository, the location to store unmodified data arriving from data providers. When the data is mapped, structural information is transformed from one metadata format into an internal representation.

This allows for the last step of the ingestion phase to take place, that of normalisation and enrichment, where the data is aligned with common authority files and is enriched with links to web resources, rights information, etc.

The data is then converted into the Europeana Data Model (EDM) for publication into the Europeana portal.

There were some cases though (e.g. Estonia, South-Tyrol, Berlin, Poland) where extra metadata was not harvested from the data providers. This was not preferred but if the partner was not able to provide extra metadata (e.g. the newspapers were not part yet of their digital collections or the digital systems were being upgraded), the METS/ALTO files were used in order to extract the necessary title and issue-level metadata. Anyhow, in such cases and in order for the newspapers to be published to TEL and delivered in EDM to Europeana, the collections were enriched with mandatory metadata, such as links to web resources and rights information, generated or even merged (e.g. BnF with Gallica) from the TEL side.

For full Partners if extra metadata was harvested, it needed to be aligned with the data received from the UIBK and CCS, and so both title and issue-level records were necessary.

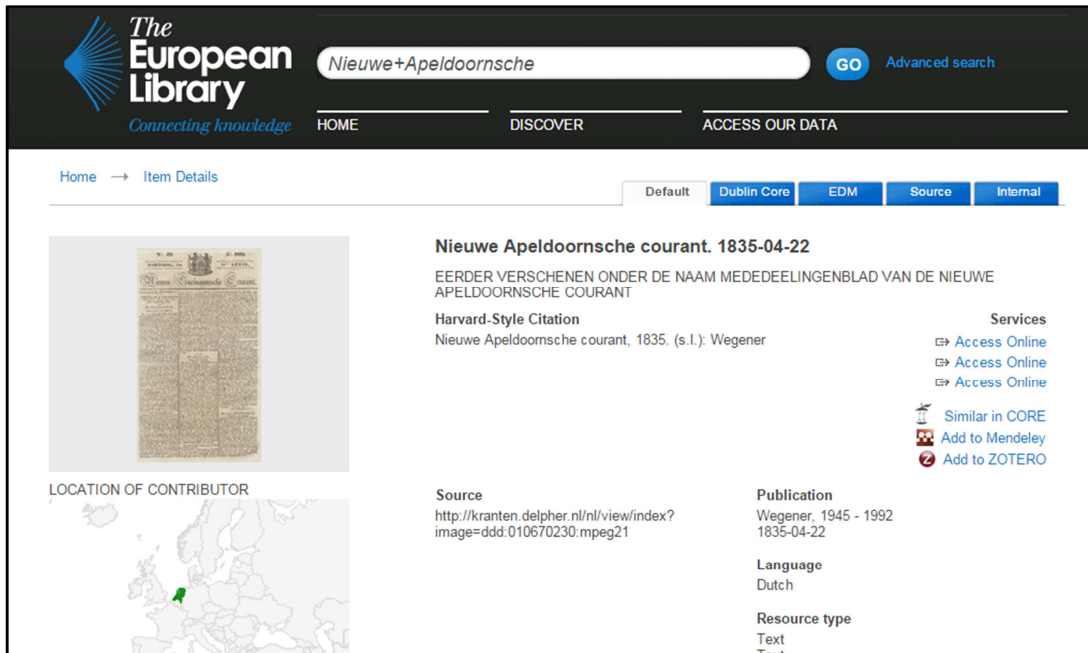
#### 4.4 Seeing the Metadata Formats

Logging into The European Library allows users to see all these different metadata formats. (Login is available at the top right of the homepage).

---

<sup>13</sup> The tool Repox was originally brought into The European Library workflow via the Europeana Libraries project. Earlier versions of Repox are available here - <http://rebox.sysresearch.org/>.

<sup>14</sup> <https://www.openarchives.org/pmh/> Open Archives Initiative Protocol for Metadata Harvesting. FTP (File Transfer Protocol) and HTTP (Hypertext Transfer Protocol) are standard web protocols for the exchange of data.

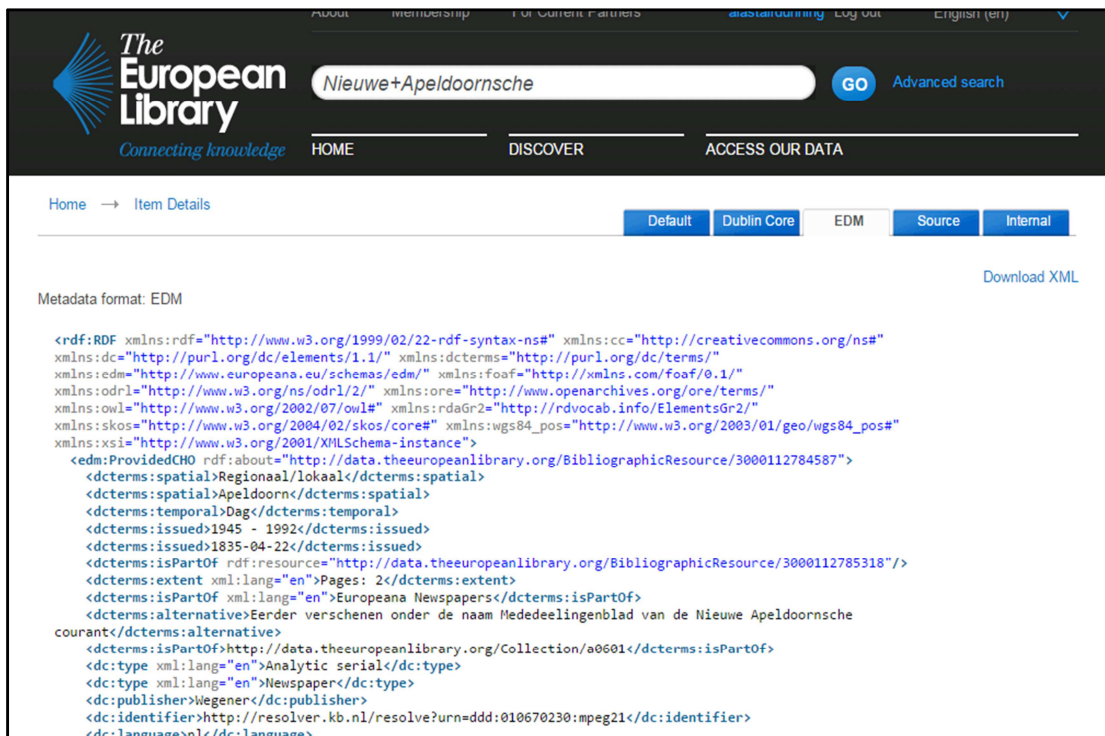


The screenshot shows the 'Item Details' page for 'Nieuwe Apeldoornsche courant, 1835-04-22'. The metadata is displayed in a clean, user-friendly format. Key elements include a thumbnail of the newspaper, a map showing the location of the contributor, and a list of services such as 'Access Online' and 'Add to Mendeley'. The 'Source' field provides a URL to the original document, and the 'Publication' field indicates the date and publisher.

The default metadata view for the *Nieuwe Apeldoornsche courant*, 22 April 1835

<http://www.theeuropeanlibrary.org/tel4/record/3000112784587>

(Note: this functionality is only available to users who have logged in)



The screenshot shows the 'EDM' (Export Metadata) view of the same record. It displays the raw XML metadata, which is structured according to the Dublin Core (DC) and EDM standards. The XML includes various metadata elements such as 'dc:issued', 'dc:isPartOf', 'dc:language', and 'dc:identifier', providing a detailed view of the underlying data structure.

The EDM representations for the *Nieuwe Apeldoornsche courant*, 22 April 1835

<http://www.theeuropeanlibrary.org/tel4/record/3000112784587>

(Note: this functionality is only available to users who have logged in)

## 5. The Resulting Corpus of Data

At the time of writing, the newspapers interface serves 3.5 million issues with over 11 million pages of full-text indexed. There are around 5 million images of newspapers images served via the TEL IIP image server; the other images are served via the image servers at the respective libraries. The size of the full text repository is 900 GB and the metadata repository including all catalogue records is 850 GB. The newspaper index is 330 GB. The images are around 8 TB.

Full details of which library presented which quantity of data are in Appendix 1.

## 6. Conclusion

### **The aggregation process worked well.**

In collecting and distributing over 3.5 million metadata records and over 11 million pages of full text, the project has significantly surpassed the high targets it set for itself.

However, it was a complex process with metadata, full text and images all travelling in different directions and at different speeds. To get it right required much discussion between relevant project partners. If there were delays in any point earlier on in the chain, it caused further delays for TEL.

Maintaining that level of aggregation on a service (rather than project) level will be difficult. Rather than a serial workflow, where one partner passes data from their servers in location A to another in location B and then to a third in location C, making use of a shared technical environment might be more efficient in the future; this would allow different partners to perform different operations on the same data in the same place.

It is interesting to note that even in the time of greatly increased internet speeds, hard drives are still required to pass large quantities of data. While this is a robust way of sending images, it is slow and not overly cheap. Therefore the incorporation of dynamic harvesting of images is of extreme importance, negating the need to pass large libraries of digitised images between different partners. TEL has also started experimenting with harvesting full text via OAI and it is hoped that full text from the national libraries of Belgium and Iceland will be harvested in this fashion.

## Appendices

### I - Content aggregated to The European Library

Totals (as of March 2015):

Number of Issues (i.e. Metadata Records) - **3,572,194 (estimate based on DoW, 1,874,641)**

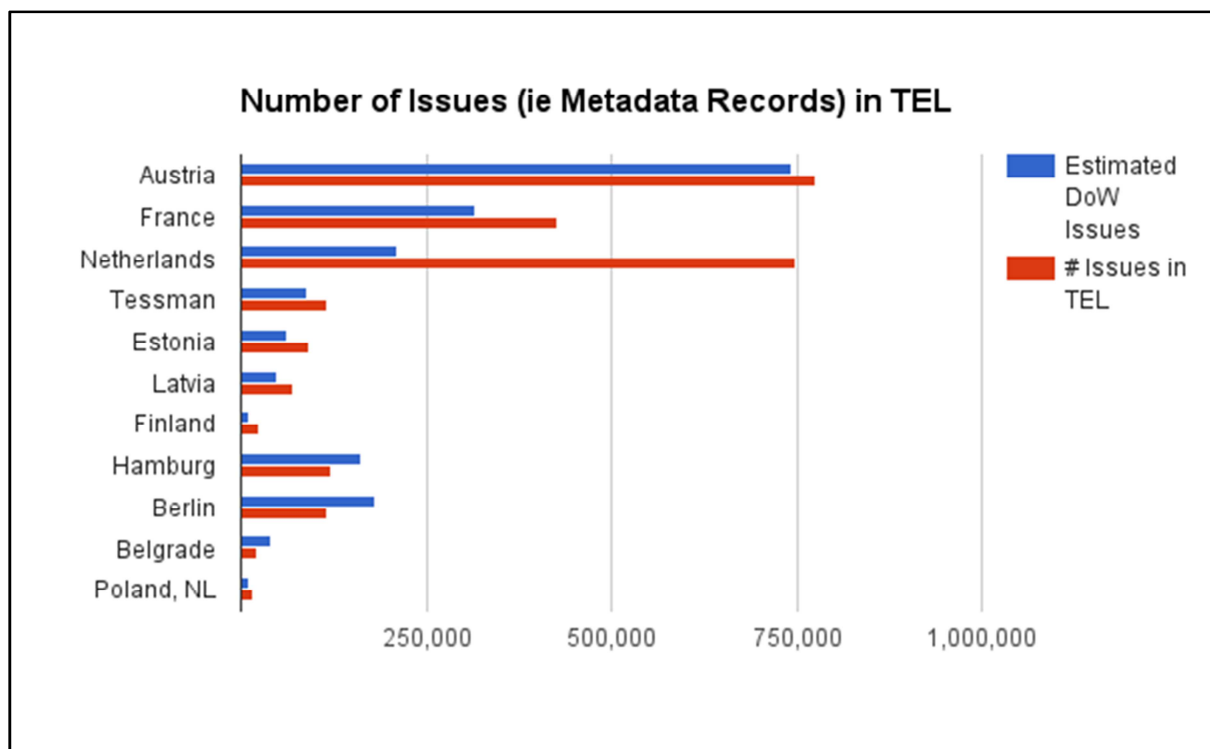
Number of Full Text Pages - **11,202,411 (target: 10m)**

Number of Images - **11,202,411** (includes images harvested on the fly)

Number of Titles - **991**

#### Full Partners

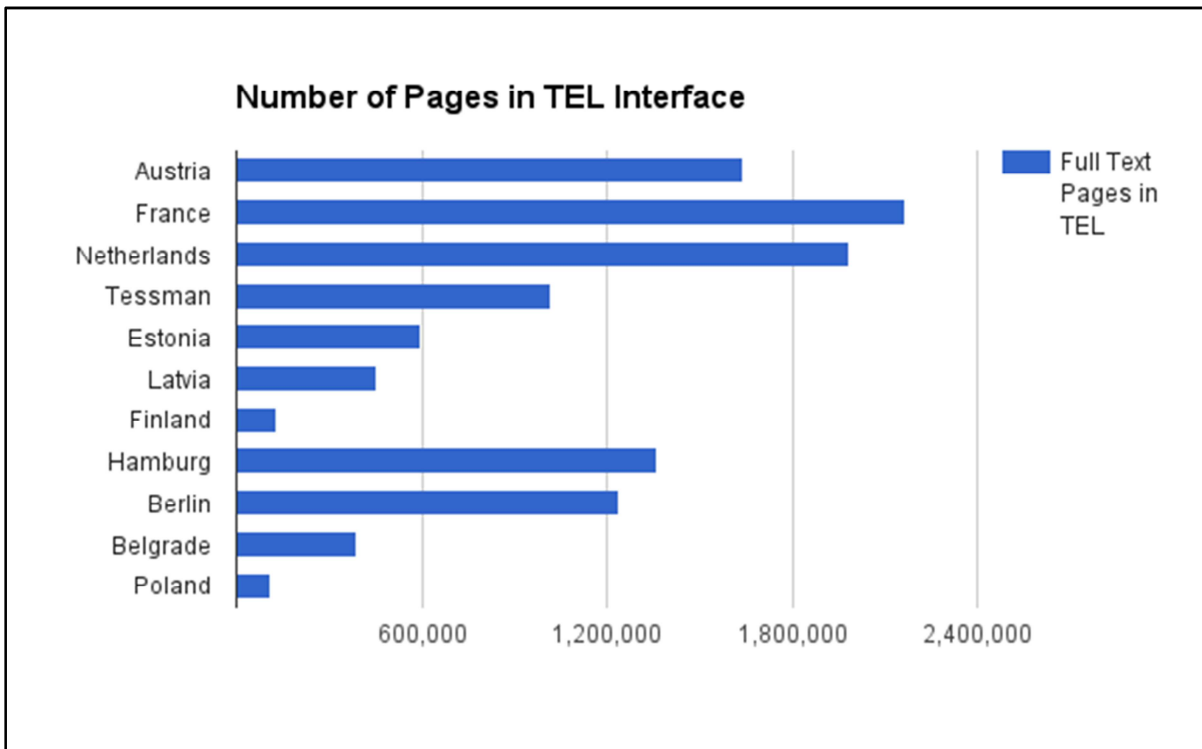
Number of Issues (i.e. Metadata Records)



*Number of metadata records in The European Library per provider (red). The figure in blue is the original estimate of issues to be provided based on a sampled average of 9.5 pages per issue*

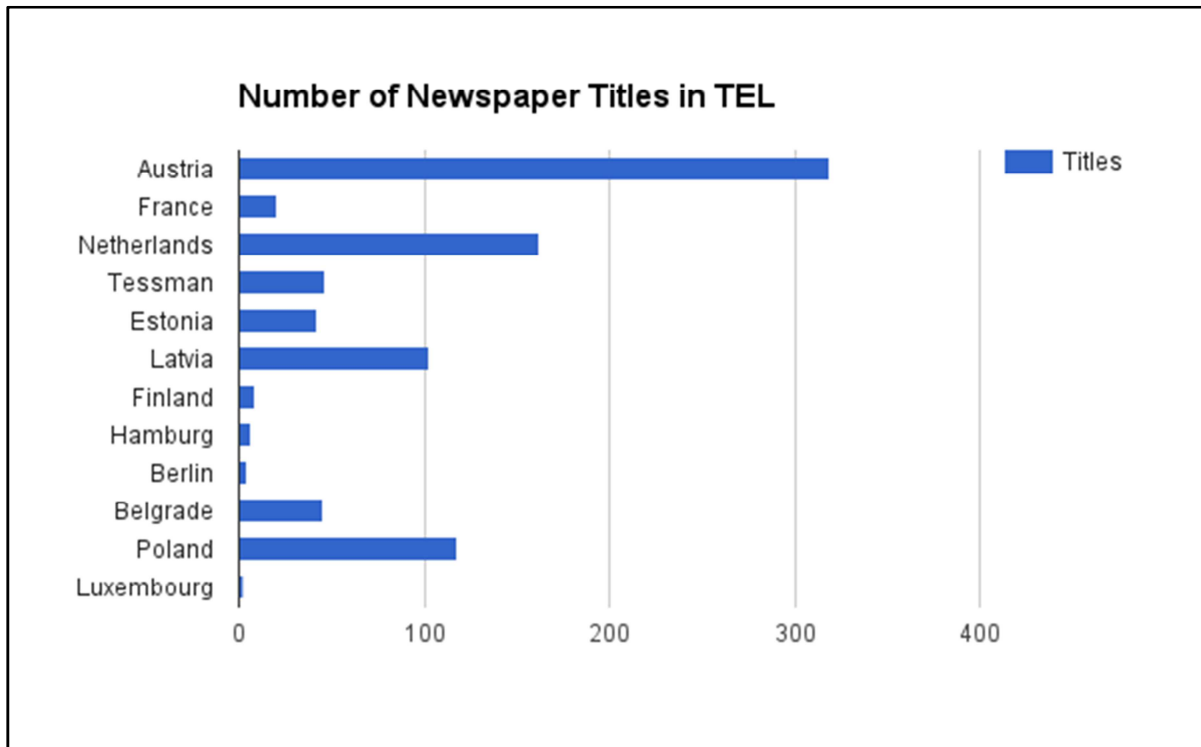


## Number of Full Text Pages



*The number of full text pages (and images) from each Full Partner, available to search and browse on the TEL website.*

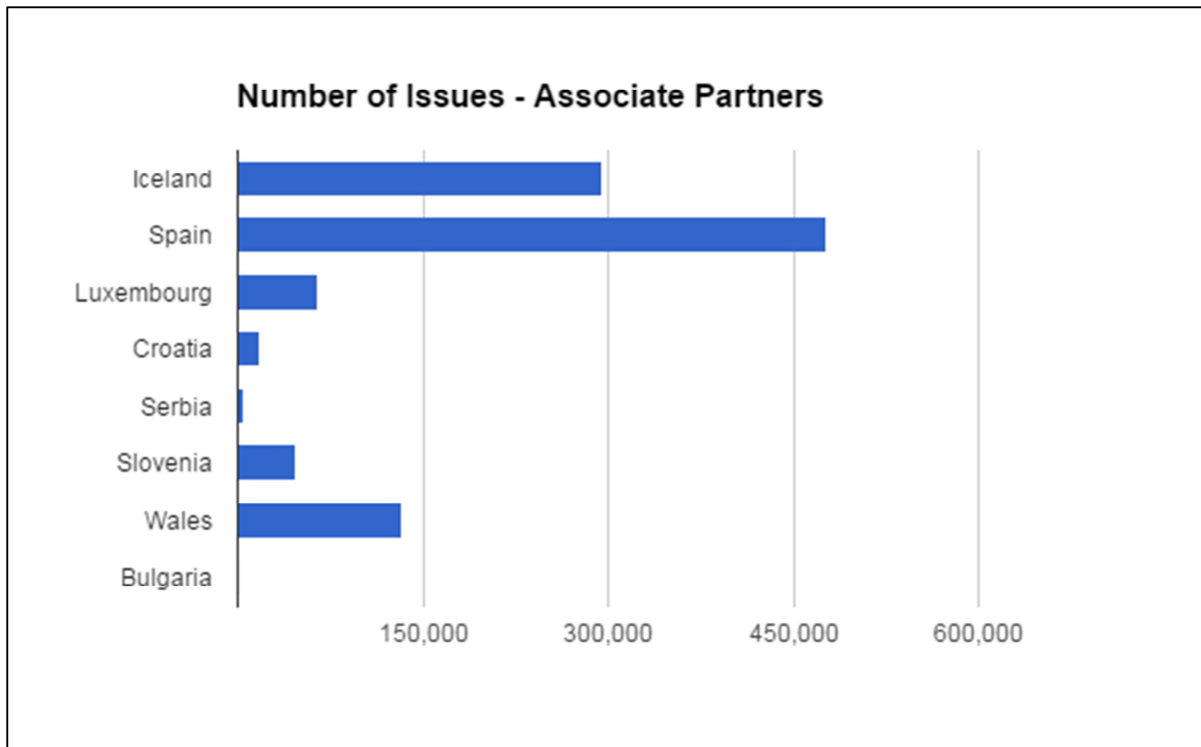
## Number of Titles



*The number of newspaper titles from each Full Partner, available on the TEL website*

## Associate Partners

### Number of Issues



*The number of issues (metadata records) from each Associate Partner, available on the TEL website*

A live version of this data is available at

[https://docs.google.com/spreadsheets/d/1CjVU85xcPo2F-OGMNw2wo\\_I18ElaGifu-EKoZmgNfF8/pubhtml](https://docs.google.com/spreadsheets/d/1CjVU85xcPo2F-OGMNw2wo_I18ElaGifu-EKoZmgNfF8/pubhtml)

## II - Content aggregated to Europeana

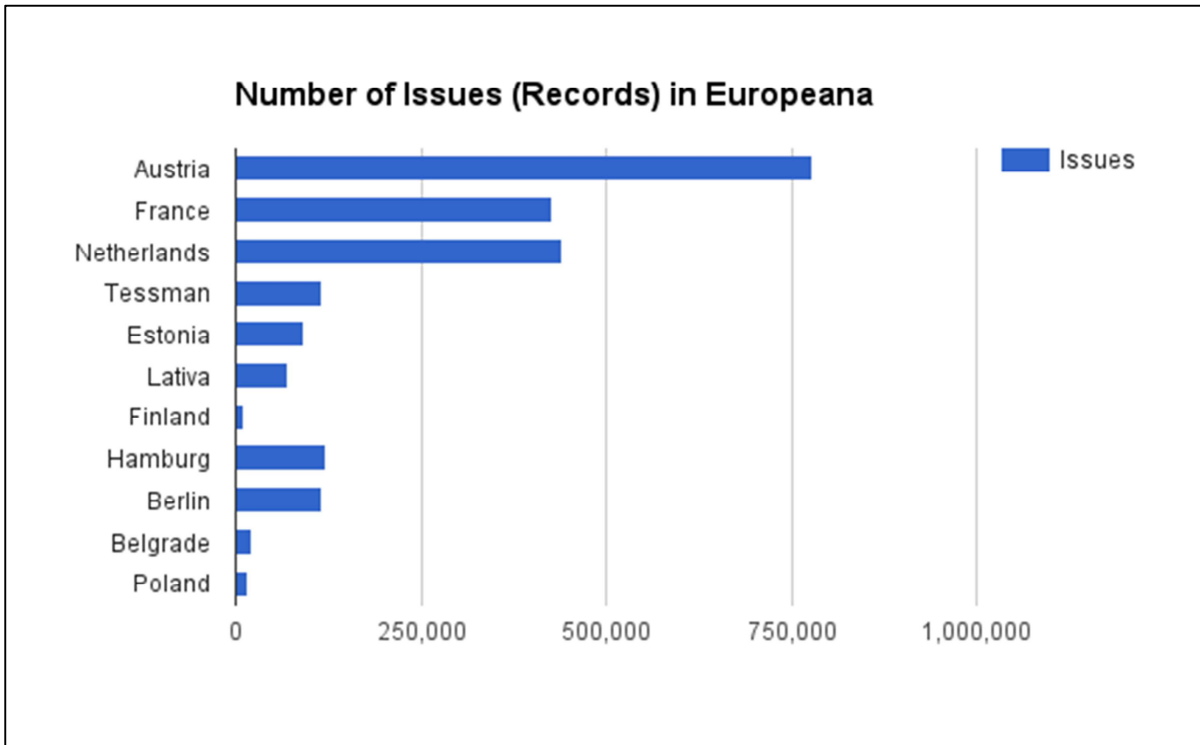
Totals as of March 16 2015:

Number of Issues (i.e. Metadata Records) - **3,281,123**

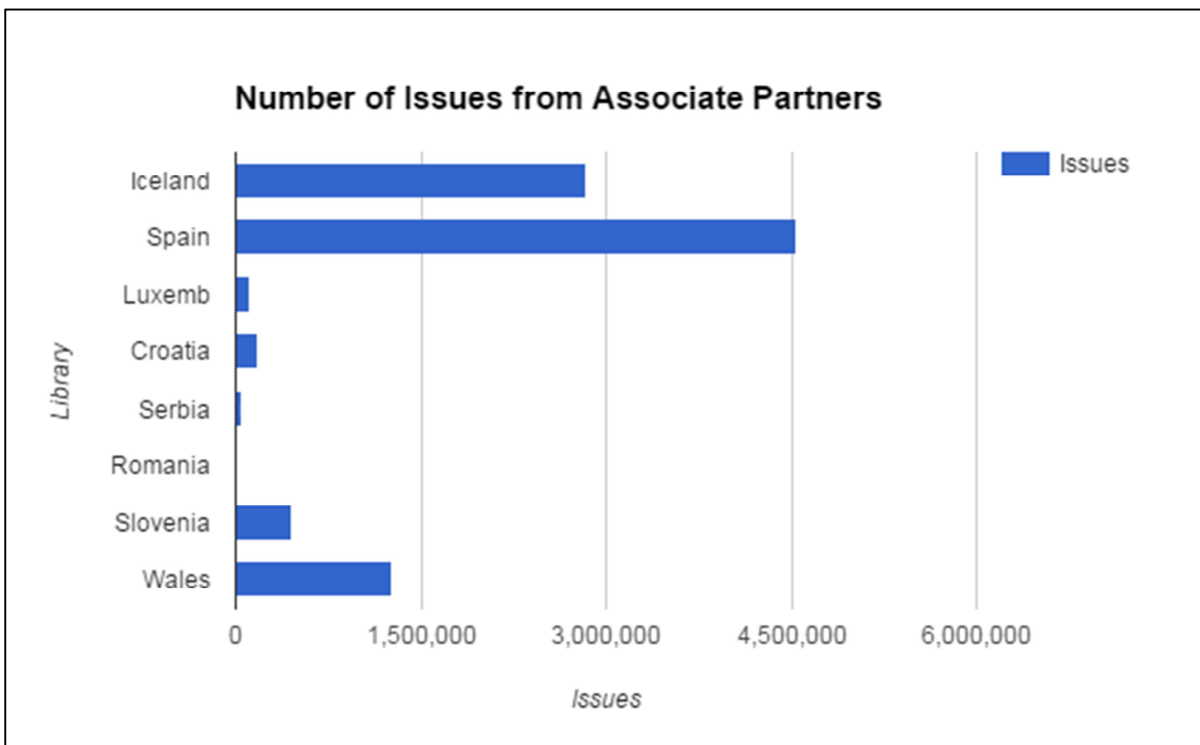
Number of Full Text Pages - **not applicable, as there is not full text in Europeana**

Number of Images - **11,202,411** (all images served via TEL or source library )

### Full Partners - Number of Issues



The number of issues (metadata records) from each Full Partner, available on Europeana portal



The number of issues (metadata records) from each Associate Partner, available on Europeana portal

A live version of this data is available at

[https://docs.google.com/spreadsheets/d/1CjVU85xcPo2F-OGMNw2wo\\_I18ElAGlfu-EKoZmgNfF8/pubhtml](https://docs.google.com/spreadsheets/d/1CjVU85xcPo2F-OGMNw2wo_I18ElAGlfu-EKoZmgNfF8/pubhtml)

### III - Breakdown of issues by library

#### Full Partners

Library	Any Issues
National Library of Austria (ONB)	All data scheduled to be released has been made available on TEL and Europeana. After receiving internal legal advice, the ONB could not make content from after 1876 available on TEL.
National Library of France (BnF)	Although TEL received the full text and images from UIBK and CCS in various batches (amounting to 2.1m pages) throughout the lifetime of the project, TEL did not receive the metadata for this data until early Spring 2015. Metadata for a further 600,000 pages during the project was received in late March.
National Library of Netherlands (KB)	All data received from the KB is available on TEL and Europeana. In addition to the DoW, the KB provided an extra 269,633 issue level records.
Dr. Friedrich Teßmann State Library (LFT)	All data received from LFT is available on TEL and Europeana.
National Library of Estonia (NLE)	All data received from NLE is available on TEL and Europeana. One early set of images from CCS arrived in a slightly different format profile and this caused display problems when TEL moved to IIP Server 7. TEL is currently attempting to rectify these problems.
National Library of Latvia (NLL)	All data received from NLL is available on TEL and Europeana.
National Library of Finland (NLF)	All data received from NLF is available on TEL and Europeana.
State and University Library of Hamburg (SUBHH)	The last delivery of content (related to 129,141 pages) arrived at TEL on 22 March. This has now been placed on the TEL servers and is awaiting integration into the TEL and Europeana websites.
Berlin State Library (SBB)	The last delivery of content (related to 162,149 pages) arrived at TEL on 22 March. This has now been placed on the TEL servers and is awaiting integration into the TEL and Europeana websites.
University Library of Belgrade (UB)	All data received from UB is available on TEL and Europeana.
National Library of Poland (NLP)	All data received from NLP is available on TEL and Europeana.

National Library of Turkey (NLT)	Initially it appeared that NLT could not share their data with Europeana as the content was hidden behind a paywall. A special API for sharing NLT data with TEL and Europeana was finally made available March 2015, but there was no more time during the project to integrate this data. This is foreseen for the latter half of 2015.
----------------------------------	---

## Associate Partners

Library	Any Issue
National and University Library of Slovenia (NUK)	All metadata received from NUK is available on TEL and Europeana.
National and University Library of Iceland (NLI)	All metadata received from NLI is available on TEL and Europeana.
Royal Library of Belgium (KBR)	TEL have agreed with the KBR that TEL will ingest both metadata and full-text content on a title by title base through the latter half of 2015.
National and University Library of Zagreb (NLZ)	All metadata received from NLZ is available on TEL. Still awaiting CC0 confirmation before sending to Europeana.
St. Cyril and Methodius National Library of Bulgaria (NLB)	All metadata received from NLB but still to be indexed by TEL. The metadata is included in Europeana.
National Library of Romania (NLR)	All metadata received. However, NLR had no permanent links to issues so title level data available only.
National Library of Luxembourg (BNL)	All metadata received from BNL is available on TEL and Europeana. Extra full text and images were also included.
National Library of the Czech Republic (NLCR)	All metadata received. However, the METS data required too much work to process fully and it has not been integrated at an item level yet. TEL has begun conversations with NLCR about title level metadata.
National Library of Spain (BNE)	All metadata received from BNE is available on TEL and Europeana.
National Library of Portugal (BNP)	All metadata received. However, BNP had no permanent links to issues so title level data available only.
National Library of Wales (NLW)	All metadata received from NLW is available on TEL and Europeana.

National Library of Serbia (NKS)	All metadata received from NKS is available on TEL and Europeana.
National and University Library of Slovakia (NLS)	All metadata received from NLS but not yet indexed and available on TEL or Europeana.

## IV - Licensing Conditions

The libraries sharing content in the project did so under the licensing conditions listed below.

### Full Partners

All full partners agreed to release their metadata as CC0 as part of the Consortium Agreement. For the content itself, all libraries agreed in the original Description of Work to marking their newspapers as public domain, except where national legal restrictions limited this. As part of the Consortium Agreement, UIBK and CCS waived any intellectual property that was created during the OCR and OLR process.

Library	Metadata	Copyright status of Full Text and Images
National Library of Austria	CC0	Mixed: Pre 1876 - The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a> 1876 and after - Unknown: <a href="http://www.europeana.eu/rights/unknown/">http://www.europeana.eu/rights/unknown/</a>
National Library of France	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
National Library of Netherlands	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
Dr. Friedrich Tessmann State Library	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
National Library of Estonia	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
National Library of Latvia	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
National Library of Finland	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>

State and University Library of Hamburg	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
State Library of Berlin	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
University Library of Belgrade	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
National Library of Poland	CC0	The Public Domain Mark (PDM): <a href="http://creativecommons.org/publicdomain/mark/1.0/">http://creativecommons.org/publicdomain/mark/1.0/</a>
National Library of Turkey	CC0	Unknown: <a href="http://www.europeana.eu/rights/unknown/">http://www.europeana.eu/rights/unknown/</a>

All associate partners have signed the TEL partnership agreement and indicated they are happy for their metadata to be released as CC0. The only exception to this is the National and University Library in Zagreb; they have been contacted in 2014 and 2015 about the issue but TEL has still not received a signed agreement.

## V - Components and Data Formats in Project

### Internal Object Model

TEL use a Java Object Model to represent data. It is completely customised, so that it can be easily adapted it to TEL needs, but it also doesn't follow any standard. TEL receives the METS/ALTO file and parse the METS for general metadata information like newspaper title, issue date, identifiers, etc. TEL then load the ALTO full text file into memory and also put it as an entry into the internal object model. The same is done with the images. The internal object model is treated as binary data internally, but there is a XML representation that is transformed for data cleaning and enrichment processes. For this reason, each dataset uses native Java parsers to read the data, but always have one or more XSLTs connected to provide dataset specific enrichments.

### UIM Workflows with Plugins

TEL supports multiple Unified Ingestion Manager (UIM) workflows for the newspapers ingestion. However, in general they work very similar. They read data from the file system or REPOX, parse the original format into the internal object model representation, optionally download the full text or just load it from the file system, optionally download or load the images, transform the images into JPEG2000 files if they are not already in this format, link issues to the same title, store the internal object model into the SOLR based repository and finally index the full text and metadata into a search index.



## IIP Image Server

The images for the viewer are served from the open source software IIP Image Server. It supports multiple protocols for access (IIP, IIIF, etc.), but the images to be served are limited to pyramidal TIFFs and JPEG2000 files. TEL provides a profile to create JPEG2000 images from TIFF and JPG files to the partners, using the Kakadu image library. If partners can't perform the JPEG2000 creation on their side, TEL can also do it on the fly for them.

## Solr Repository

The metadata and the full text are stored in a NoSQL database based on SOLR. There are multiple cores holding different kind of information. Authority, identifier, resource, sequence hold more provenance data. The metadata core holds metadata information and the full text core holds full text. The repository has a master/slave setup. On the master instance UIM is operating and the replicated slave serves end-user services.

## Solr Newspaper

The search index for newspapers is split into three cores. One holds issues called "full text", one holds title entries called "title" and one is dedicated for query suggestion called "suggestion". Again there is a master/slave setup that automatically replicates from the index master to the search slave. This setup is mostly done for performance reasons for end-users. Replication is done automatically every 24 hours.

## Data Formats

This section gives an overview of the different data formats used in the project.

### METS

The OCR and OLR partners in the project deliver data in a METS/ALTO combination (the ENMAP profile developed in WP5). METS provides metadata information like identifiers, titles, issue date and defines links to images and full text. There are already differences between the two partners in the format and other entities might again use it differently. One partner provides article information in the METS file by outlining which parts in the connected ALTO full text files make up an article. METS information must be mapped to EDM, so that EDM can be used as main data format. However, this means EDM needs to support full text links and article support. Otherwise, TEL would need to start processing METS natively.

### ALTO

Full-text is provided via the ALTO format. Currently, TEL gets a ZIP file consisting of the METS file with ALTO files and JPEG2000 files, so that nothing needs to be downloaded. However, for the National Library of the Netherlands TEL harvested the full text remotely after processing the metadata where the full text is referenced. In the future, TEL would need to agree on the way to retrieve full text. TEL may be able to harvest metadata via OAI-PMH and harvest full text by following embedded links.

## **JPG / TIFF / OTHERS**

TEL also sometimes retrieves images in formats like JPG or TIFF. These images are then transformed on the fly to JPEG2000 to be served from an image server.

## **JPEG 2000**

JPEG2000 is the format the IIP Image Server can read and provide tiles to a client. The Kakadu image library<sup>15</sup> is needed to create these images, as well as a library to build the IIP Image Server to support JPEG2000. However, the Kakadu library must be licensed.

## **The European Library Internal Object Model**

TEL use a Java Object Model to represent data internally.

## **Europeana Data Model**

EDM should be the main data format used everywhere in the systems. This means that the current functionality of the internal object model must be represented in EDM, so EDM should be adapted to reference full text objects as well as provide article support if possible. The advantage will be a standardised format and increased reusability, but there will be a loss in flexibility.

---

<sup>15</sup> <http://kakadusoftware.com/>