Enriching Europeana with user transcriptions and annotations (EnrichEuropeana)

Milestone 6: Europeana Collections Integration

1. INFORMATION ON THE ACTION

Grant Agreement Nº	INEA/CEF/ICT/A2017/1568419
Action Title (Art. 1 of G.A.)	Enriching Europeana with user transcriptions and annotations (EnrichEuropeana)
Action number (Art. 1 of the G.A.)	2017-EU-IA-0142

Editorial Information

Revision	1.0
Date of submission	
Author(s)	Hugo Manguinhas, Sergiu Gordea, Marcin Helinski
Dissemination Level	public

Revision History

Revision No.	Date	Author	Organization	Description
0.1	17.12.2019	Sergiu Gordea	AIT	Template and table of contents
0.2	30.12.2019	Marcin Heliński	PSNC	Integration with Europeana CSP
0.3	17.01.2020	Hugo Manguinhas	EF	Integration in Europeana Collections
1.0	17.01.2020	Sergiu Gordea	AIT	review and final editing

Co-financed by the Connecting Europe Facility of the European Union

Statement of originality

This report contains original unpublished work except where clearly indicated otherwise.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

The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the opinion of the European Union.

Table of contents

Co-financed by the Connecting Europe
Facility of the European Union

1. Introduction

This document presents a summary of the work carried out for the integration of enrichments created within the Transcribathon Platform into Europeana Collections. This work is related to the **Activity 1: Platform for Transcription and Enrichment of EnrichEuropeana** Action and its subtasks:

> Task 1.1. Graphic design and user experience
>
> This task will focus on the front-end of the web platform, in order to optimise the design and user operability of the platform. It will develop an attractive, interactive and user- friendly interface for end users to participate in transcription and enrichment activities that will be organised during Activity 4. The website will be designed with modular styles to allow adaptability to different sources and topics. It will use methods of User-Centered Design (UCD) / User Experience (UX).
>
> Task 1.2. Enhanced end user functionalities
>
> This task will incorporate the development of new and innovative user tools and functionalities including enhanced search capabilities. The controlled vocabularies (developed under the Task 2.1) will be made available for creation of user annotations. Additional toolsets will be offered to the users within the Graphical User Interface of Enriching Europeana Platform including those for enhanced visualization and interaction with the transcriptions, assisted geo tagging, map-based visualisation and browsing. The improvement of the enrichment abilities will allow users to engage in greater citizen-science activities.

It presents the evaluation of the enrichment use cases and their relevance for the regular users of Europeana Collections portal. It also describes the technical infrastructure that was developed to automate the enrichment transfer process which is based on the Data Exchange Infrastructure (see also MS3), the Annotations API and the Full-text Indexing Tool developed as part of Europeana Core Services.

2. Assessment of the proposed use cases for enrichment

The evaluation of the enrichments contributed within the scope of the action are documented within the document “Supporting user contributed transcriptions and enrichments coming from Transcribathon platform”. Within this evaluation, the following use cases were identified as possible enrichments that could be offered to the user:

- A transcription for each page in the item (users can additionally set the language, styling and alignment of the text)
- A description of what the item is about and describe in detail images and objects that might appear in the item (e.g. description of the image on a postal stamp, or individual images in a photo album)
- A date that reflects the content of the item
- A classification of the item (either Letter, Diary, Postcard or Picture)
- A Person somewhat associated to or mentioned by the item (e.g. creator, subject) further detailing it with its name, place and date of birth and death
- A precise (a geo-coordinate or a street address) or named location (ie. Place)
- A link to an external webpage with information pertaining to the item’s content
- A free-text keyword pertaining to the topic and content of the item

From this list of use cases identified for the Transcribathon platform, the UX team at Europeana has evaluated their usefulness and concluded that only the transcriptions, semantic and keyword enrichment ought to be considered for integration in Europeana due to either uncertainty on the quality of resulting enrichments made by users and their added value to the already available metadata provided by the data provider. Other use cases may be considered in the future for integration if they are found to contribute positively to the user experience on Europeana.

2.1. Compliance and adjustments to the EDM Annotations Profile

The EDM Annotations Profile (see Part I and II) has been reviewed in the light of the use cases that were selected for integration. It was found that the current model was able to cope with the selected use cases, with the exception of the user contributed transcriptions. This was mainly due to the fact that the actual transcription text was not accommodated in the use case (only a link to an external resource) and also it was not considering the need to indicate the license of the transcription which was identified as a requirement in the guidance policy document. As a result, the EDM Annotations Profile was updated to cope with the new transcription and is now modelled as presented in Fig. 2.

```
{
  "motivation" : "transcribing" ,
  "body" : {
    "type" : "FullTextResource" ,
    "language" : "de" ,
    "value" : "BÜNDNIS  90\n\n\nBürger für Bürger\n\n\nInitiative Frieden\nund  Menschenrechte\n" ,
    "format" : "text/plain" ,
    "edmRights" : "http://creativecommons.org/publicdomain/zero/1.0/"
  },
  "target" : {
    "scope" : "http://data.europeana.eu/item/135/_nnVvTdx" ,
    "source" :
"rhus-209.man.poznan.pl/fcgi-bin/iipsrv.fcgi?IIIF=1//135/_nnVvTdx/2_DSC_0214_crop_web.tif/full/full/0/default.jpg"
  }
}
```

Figure 2. The representation in JSON-LD of a transcription annotation.

In case of a transcription, the following fields can be supplied according to the following rules:

- motivation - for transcriptions always *“transcribing”*
- body.type - for transcriptions always *“FullTextResource”*
- body.language - when available this value is taken from TP response with transcription
- body.value - content of the transcription
- body.format - for the time being always *“text/plain”*
- body.edmRights - for the time being always *“http://creativecommons.org/publicdomain/zero/1.0/”*
- target.scope - URL to record in Europeana
- target.source - URL to the specific image for which transcription was prepared

Given that transcriptions refer to a specific image (media resource) within an item, besides referring to the source item on Europeana, the annotations must also refer to it using its URL. This is indicated in the “source” field of the annotation alongside the “scope” field which indicates the source item.

3. Transcribathon integration with the Annotations API

Materials used in Transcribathon Platform are prepared to be transferred to Europeana CSP using the Data Exchange Infrastructure (DEI). During physical or online Transcribathon events, the participants create transcriptions for the documents available in the Transcribathon Platform. The transcriptions and other types of enrichments are verified and validated by experienced users and the DEI is informed about the complete validation of documents using the REST API. In the following step, DEI submits the transcriptions to Europeana CSP via the Annotations API, by using the EDM Annotations Profile (a standardized representation compliant with W3C Web Annotations standard). The following diagram sketches the communication flow between the main components of the Enriching Europeana infrastructure.

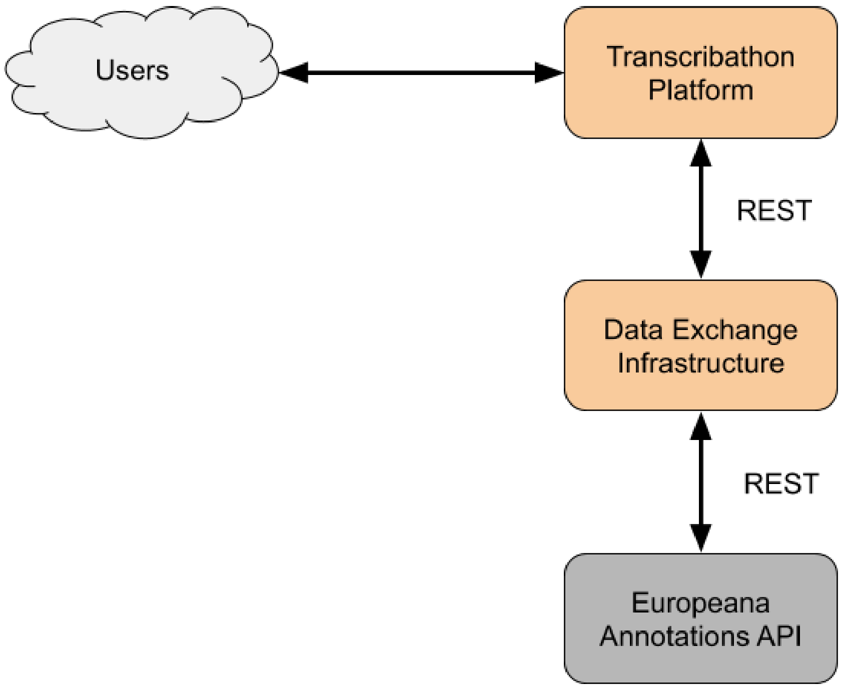


Figure 1. Communication between main components in transcribing process.

When a transcription is ready in the Transcribathon Platform, it informs DEI by sending POST request to `/api/transcription` endpoint and supplying record identifier.

`POST /api/transcription`	
Header:	- Accept: application/json - Content-Type: application/json
Parameters:	- recordId (mandatory): record identifier
Processing	
- Check credentials, otherwise respond with HTTP 401. - Check if requested format is supported, otherwise respond with HTTP 406; - Check if the request is properly specified (including the record identifier validation), otherwise respond with HTTP 400. - Find Record in database and change its state to waiting for enrichment. Return HTTP 404 in case the record was not found. - Return response with HTTP 200	
Response	
Success:	An HTTP 200 response is returned with empty body
Headers:	The following headers are returned in response: `Content-Type: application/ld+json; charset=utf-8`
Errors:	HTTP 400, 401, 404, 406 see section on error handling.

When handling the above request, DEI enqueues the task to perform further actions that will lead to deliver a prepared transcription to Europeana.

In order to retrieve transcriptions/enrichments for a certain item, it is necessary to specify the id of the record and the purpose of the enrichment (e.g. transcribing, tagging). For transcriptions that were sent to Europeana previously, there is also the possibility to specify its annotation identifier.

When transcriptions are retrieved from TP and stored temporarily in DEI they can be sent to Europeana Annotations API. For each transcription DEI prepares a POST request to `/annotation` endpoint where the text of the transcription is submitted inside the request body. Since this operation is a *"write"* operation it requires certain authorization. Therefore, only a client with certain privileges can be used for that. The privileges are confirmed by supplying the JWT token in the Authorization header of the request, which needs to be acquired in advance by using the Europeana Authorization Server.

Upon receiving a successful response from the Annotations API confirming that the annotation has been created, the DEI will submit the identifier received from Europeana to TP using the POST request to /enrichments/transcription endpoint. In case a transcription is updated after publication in Europeana, the identifier received from Europeana will be used to perform an update to the version on Europeana.

4. Improvements to the user experience on Europeana Collections

Bringing transcriptions into Europeana will improve the experience and accessibility for users when reading items (specially manuscripts) but also allow users to find items more easily items by searching on the transcriptions. Semantic or metadata-like enrichments are expected to improve the overall description and multilinguality of the item while contributing to the browsing experience on Collections, as well as, allowing items to be found more easily using contextual information.

4.1. Searching on transcriptions

While Annotations API provides its own search functionality, this doesn't fully support expectations of the Europeana Collections users, which are interested to have a simple and seamless interface that allows one to search both in metadata and the content of the documents for a given search query. Consequently, the searching of the transcriptions is implemented to follow the solution that was developed to search in the Newspapers Collection, by indexing transcriptions in the so called Full Text Index.

So that metadata and full-text can be searched together, the fulltext index needs to be regularly updated with the metadata coming from the main index (and future updates) and with new contributions made via the Annotations API. The following use cases need to be taken into account for keeping in-sync the search indices or databases used by Europeana CSP:

- Updates in Search API (main index): update or removal of an existing dataset/record. The addition is not necessary since no transcription would be expected to be in the Annotations API. For the update, it should only cover the case that the record is updated using the same identifier.
- Updates in Annotations API: addition, update and removal of a transcription.

The following requirements were identified to be implemented to support the full-text search using transcriptions:

- the fulltext index must store the lastUpdate/modification date within the fulltext index. (eventually we could differentiate if the modification comes from the record metadata or from annotation data)
- The modification of the Annotations must be identifiable through the annotation API (including deletions)
- The modification of Europeana records must be identifiable through the record metadata index, eventually with additional information from fulltext index (including deletions)
- The import tool must be able to identify annotation that are newly added, updated, deleted
- The import tool must be able to identify Europeana records that are updated (when preserving their identifier) or deleted
- The tool needs to store Europeana Record metadata and annotation text (e.g. transcriptions) in the fulltext index

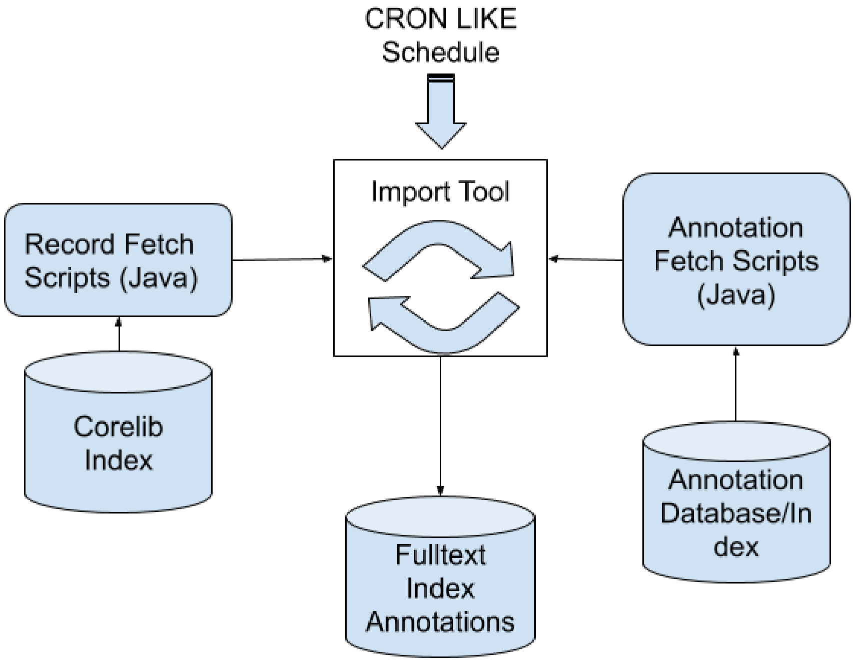
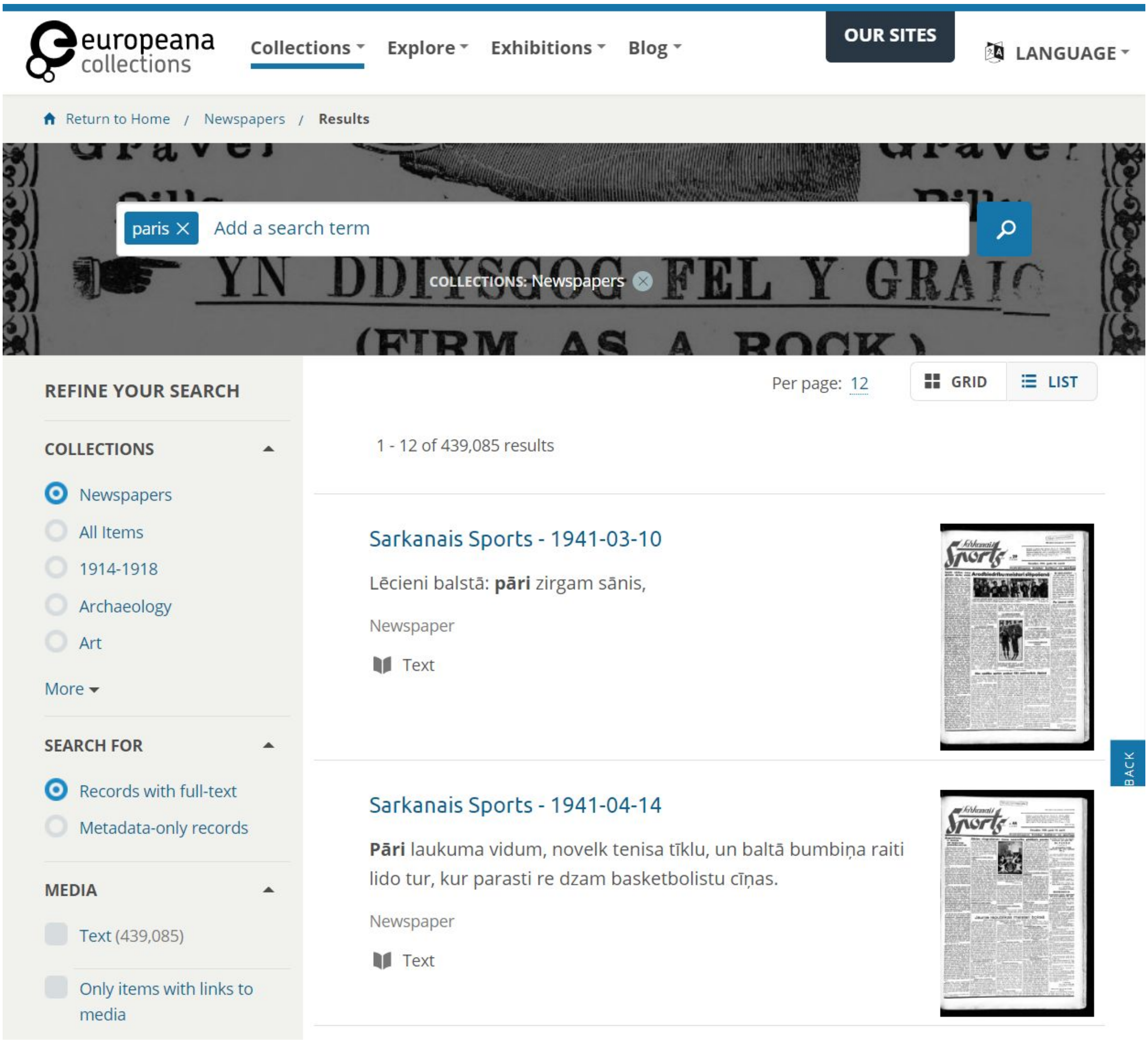

CRON LIKE Schedule
Import Tool
Record Fetch Scripts (Java)
Annotation Fetch Scripts (Java)
Corelib Index
Annotation Database/In dex
Fulltext Index Annotations

Fig. 3: Data Flow Diagram

The following changes were made in the Annotation API to meet the requirements listed before:

- The modified data is stored both in the index and the mongo database, but the deleted items are available only in the mongo database where the disabled field is set to true.
- The search method in Annotation API is used to get the object that are updated. The field list must be used in the search for performance considerations. The results must be ordered ascending by modified date.
- A new API method would be needed to get the list of identifiers of Annotations deleted after a given date. This method should take a date as input to filter results by modified field and disabled status. It must return the ids of the disabled annotations, ordered ascending by modified date.
- On the Search API, the updated records are discovered using the last modified solr field.
- The deleted items in the Record API can be identified, one by one though the Search API. A separate functionality to identify all records deleted in record metadata index can be implemented, but such information will be expensive to obtain.

The search of transcriptions in Europeana Collections will follow roughly the same experience as exists now for Newspapers, which will be preserved in the new version of the Collections portal but adjusted to fit the new way of interacting with the search box.



Return to Home / Newspapers / Results

paris ✕ Add a search term

COLLECTIONS: Newspapers

REFINE YOUR SEARCH

COLLECTIONS

● Newspapers
○ All Items
○ 1914-1918
○ Archaeology
○ Art

More

SEARCH FOR

● Records with full-text
○ Metadata-only records

MEDIA

□ Text (439,085)

□ Only items with links to media

Per page: 12 GRID LIST

1 - 12 of 439,085 results

Sarkanais Sports - 1941-03-10

Lēcieni balstā: **pāri** zirgam sānis,

Newspaper

Text

Sarkanais Sports - 1941-04-14

Pāri laukuma vidum, novelk tenisa tīklu, un baltā bumbiņa raiti lido tur, kur parasti re dzam basketbolistu cīņas.

Newspaper

Text

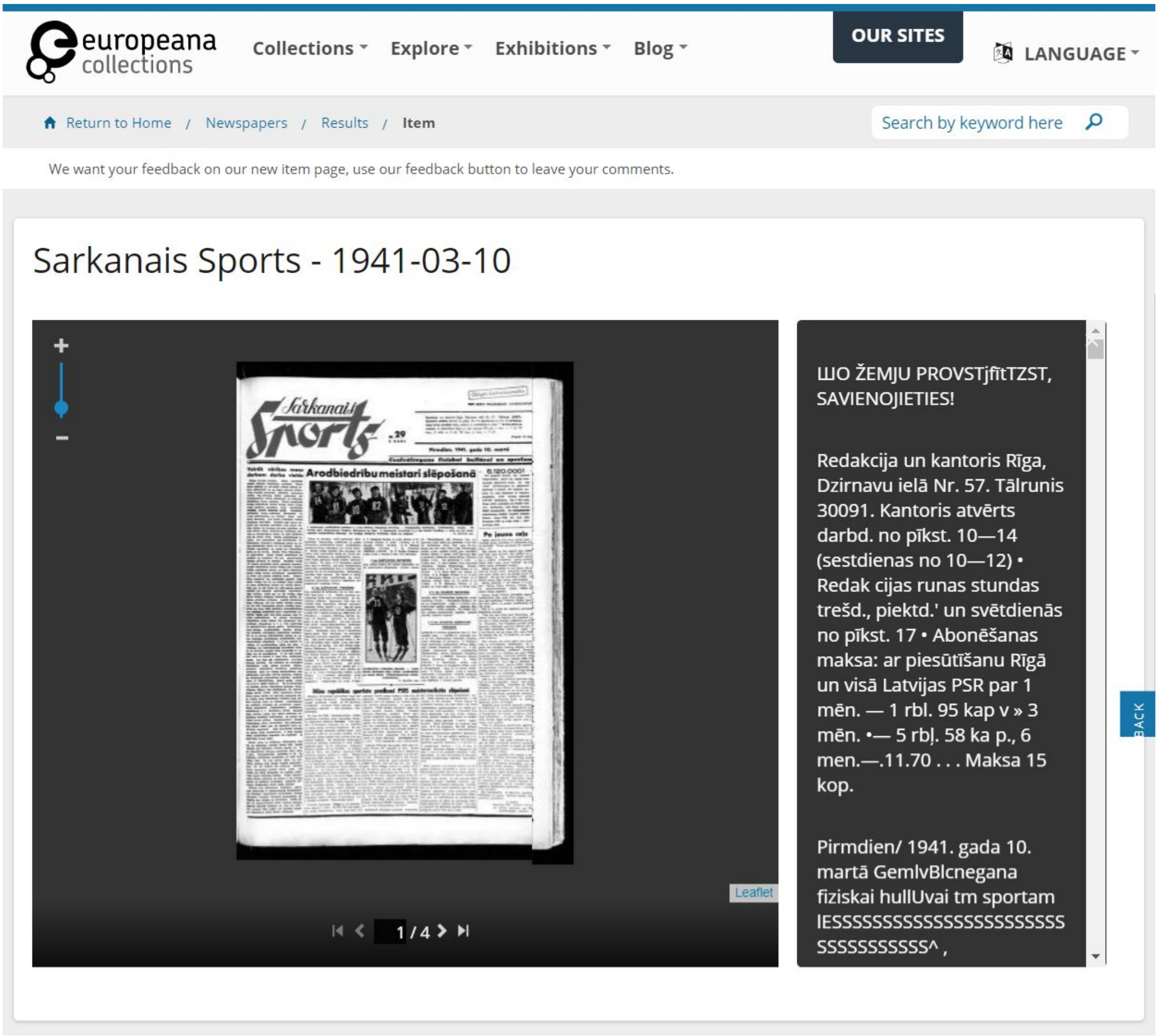
Fig. 4: View of the search results page for the Newspapers Thematic Collection in the current Collections portal.

4.2. Display of transcriptions alongside the Items

Similarly to the search on transcriptions, their display will happen on the Item Page also following the same design as was done for Newspapers (see Fig. 5). This means that the transcription text will appear side by side with the image/page and as the user scrolls along the images, he/she will see the transcriptions when available. However, in case of transcriptions it will be indicated to the user that these are coming from the community as opposed to the providing institution where no such mention is displayed.

The display will use the Annotations API to pull all transcriptions associated to the item using its search method. This will be done selectively as the user scrolls along the page to optimize the bandwidth use and this way minimize waiting time for the user.

Co-financed by the Connecting Europe
Facility of the European Union



We want your feedback on our new item page, use our feedback button to leave your comments.

Sarkanais Sports - 1941-03-10

ШО ŽEMJU PROVSTjfïtTZST, SAVIENOJIETIES!

Redakcija un kantoris Rīga, Dzirnavu ielā Nr. 57. Tālrunis 30091. Kantoris atvērts darbd. no pīkst. 10—14 (sestdienas no 10—12) • Redak cijas runas stundas trešd., piektd.' un svētdienās no pīkst. 17 • Abonēšanas maksa: ar piesūtīšanu Rīgā un visā Latvijas PSR par 1 mēn. — 1 rbl. 95 kap v » 3 mēn. •— 5 rbļ. 58 ka p., 6 men.—.11.70 . . . Maksa 15 kop.

Pirmdien/ 1941. gada 10. martā GemlvBlcnegana fiziskai hullUvai tm sportam IESSSSSSSSSSSSSSSSSSSSSSS SSSSSSSSSSS^ ,

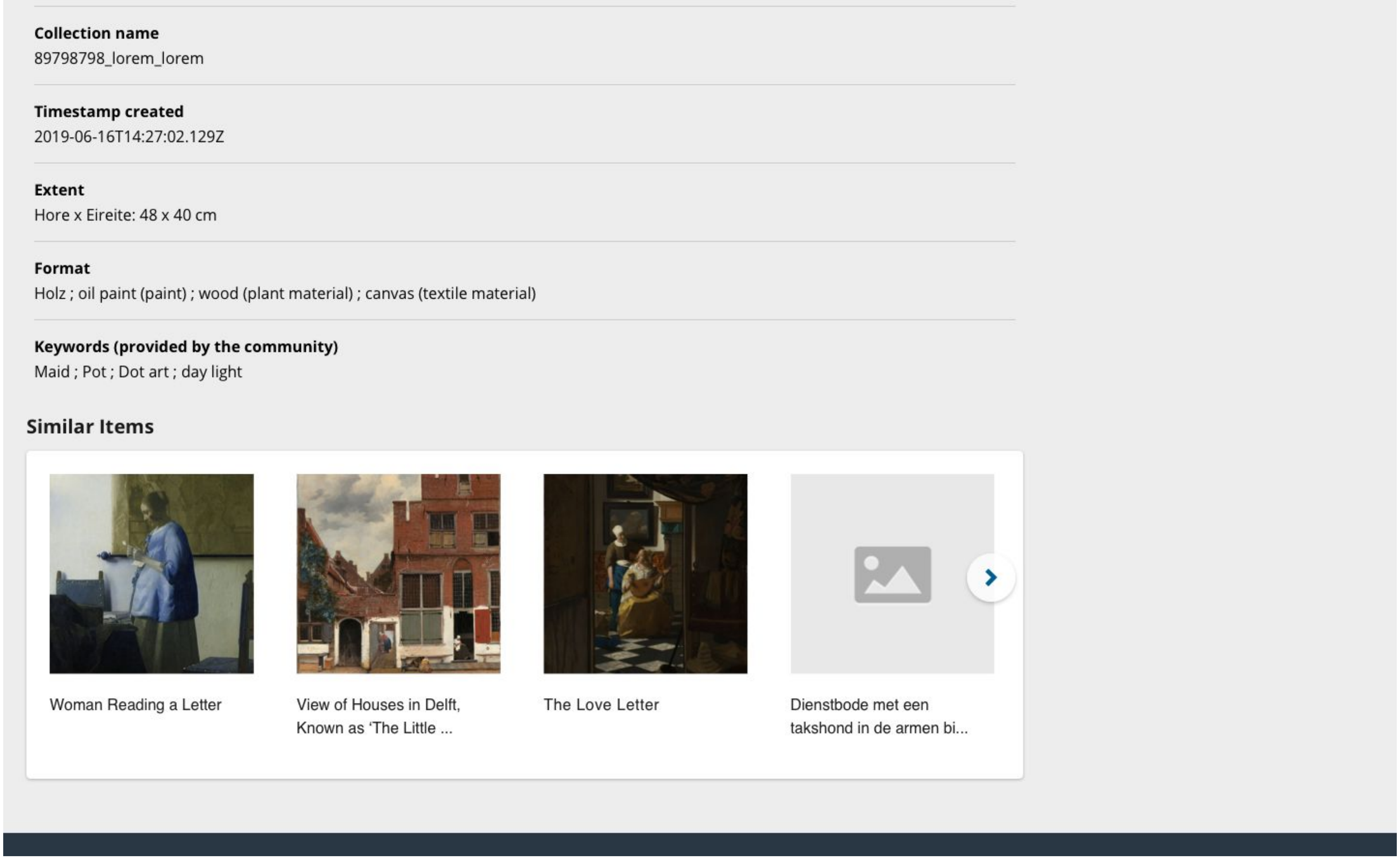
Fig. 5: View of the Item Page for a Newspaper issue in the current Collections portal.

4.3. Display of non-transcription enrichments

The display of other forms of enrichments besides transcriptions will happen on the Item page also using the Annotations API. It will use the search method of this API to find all annotations (except for transcriptions) that relate to the Item being display. To ease the display, the API was modified to dynamically obtain the labels for the semantic links by doing what is called “dereferencing”. This will use the service provided by Metis which is at present applied upon ingestion and will now also support directly the display.

With regards to the actual display and differently from the transcriptions, these enrichments will appear in a specific section under the “Extended Information” at the bottom of the metadata under a section named “Keywords (provided by the community)”. In case, the annotations are using semantic vocabularies, the keywords will adjust to the language of the user (when available) using the labels that are obtained via dereferencing. The figure below shows the designs for the upcoming display of enrichments.

Co-financed by the Connecting Europe
Facility of the European Union



Collection name
89798798_lorem_lorem

Timestamp created
2019-06-16T14:27:02.129Z

Extent
Hore x Eireite: 48 x 40 cm

Format
Holz ; oil paint (paint) ; wood (plant material) ; canvas (textile material)

Keywords (provided by the community)
Maid ; Pot ; Dot art ; day light

Similar Items

Woman Reading a Letter

View of Houses in Delft, Known as 'The Little ...

The Love Letter

Dienstbode met een takshond in de armen bi...

Fig. 6: A high fidelity mockup of the Item Page showing the display of enrichments.

5. Conclusions and further Development

As mentioned earlier in the document, other use cases may be considered in the future for integration on Europeana Collection if they are found to contribute positively to the user experience on Europeana. While not covered in this stage of the integration, the machine translations of the text transcriptions have the potential to significantly increase the search experience within Europeana Collections. This aspect will be taken in consideration to forster future opportunities for enhancing the use of transcriptions in Europeana Collections.

References

Europeana Annotation API: https://pro.europeana.eu/resources/apis/annotations

Europeana Entity API: https://pro.europeana.eu/resources/apis/entity

Web Annotation Data Model: https://www.w3.org/TR/annotation-model/

Annotation use cases submitted to Annotations Task Force:

https://docs.google.com/document/d/1af56Omq1GP1xLVvXHywxQXazWqEfhzRfMTbo6lwv5QU

Supporting user contributed transcriptions and enrichments coming from Transcribathon platform:

https://docs.google.com/presentation/d/1_wlkQr3NthfkmJg0iyR_CYfa4I84AR_nXa4qhwNPyyE

Requirements and specifications for Full Text Indexing:

https://docs.google.com/document/d/1jg9lSsRoQb97WcLdR2A_6JdmnK8voSt0eWDsLsmViHc

Co-financed by the Connecting Europe
Facility of the European Union