made with europeana

Enriching Europeana with user transcriptions and annotations (EnrichEuropeana)

Milestone 2: First version of automatic enrichment services including NER and translation

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	30.06.2019
Author(s)	Sergiu Gordea, Denis Katic, Giulia Benotto
Dissemination Level	public

Revision History

Revision No.	Date	Author	Organization	Description
0.1	7.06.2019	Sergiu Gordea, Denis Katic	AIT	first contributions to all sections of the document
0.2	14.06.2019	Sergiu Gordea, Denis Katic	AIT	enhancing description of Architecture Overview and Service Description Sections
0.3	17.06.2019	Giulia Benotto	Net7	Updated experimental evaluation
0.4	19.06.2019	Luca de Santis	Net7	Document review
0.5	28.06.2019	Hugo Manguinhas, Antoine Isaac	EF	Document review
1.0	30.06.2019	Sergiu Gordea	AIT	Final version

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This report contains original unpublished work except where clearly indicated otherwise.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

Introduction

This document presents the technical documentation for the automatic enrichment services developed within the scope of **Activity 2: Semantic Enrichment and Quality Control Services.**

Manual transcription of text materials, as enabled by the Transcribathon platform, generates high-accuracy resources, supports a better understanding of materials and allows further reuse. The goal of the newly developed platform is to enhance the accessibility and understandability of material related to overarching themes which hold relevance to important events that contributed to the development of modern Europe, namely World War I, Migration or Revolutions of 1989. New technologies for the automated analysis of transcribed text are developed to achieve this goal. Machine-translation services are integrated in order to make the content accessible and searchable for users in their preferred languages. Natural language processing technology is employed to automatically extract named entities and to match them against semantic web repositories like Europeana Entity Collection, DbPedia and/or Wikidata. Translations are also used as an enabler for integrating advanced solutions for extraction of contextual information from historical documents.

The work presented in this document includes the progress achieved so far within the following Tasks of the Action:

Task 2.1. Controlled vocabularies for transcription and metadata enrichments
The semantic enrichment of textual resources and metadata is based on usage of linked data resources, which will be accessed through a unique endpoint represented by Europeana Entity API. For this purpose, relevant named entities from open linked data repositories (e.g. Wikidata, DbPedia, Geonames, etc.) will be integrated within the Europeana Entity Collection (EC). The EC is the underlying knowledge graph developed by Europeana by combining statements from various linked data repositories. It has the goal to centralise information about the contextual entities related to the cultural heritage objects. New named entities identified by Task 2.3 will be prepared and proposed for ingestion in EC which can then be made available through the Europeana Entity API. This way, the new entities will be available to be used for semantic enrichment of cultural heritage objects, displayed in Europeana as Entity Pages, and used for auto-suggestion and search, both in Europeana Collections and EnrichEuropeana platform.

Task 2.2. Pilot on automatic translation of transcribed text
The textual resources generated through the transcription campaigns will be translated into all official European languages by using the translation services provided by the eTranslation DSI and Google Translate. The automatic translations will be stored in a database and in a multilingual search index (i.e. using Apache Solr), and made available to the application developers through the new developed Translation API. This way, the API will provide support for the retrieval of cultural heritage objects (i.e. transcribed manuscripts) by using search terms in any European language. The translations of the transcribed text will be used as input for the Natural Language Processing services developed in Task 2.3.

Task 2.3. Automatic analysis and Named Entity Recognition
Natural language processing algorithms will be used to analyse the German and English transcriptions and translations (i.e. Natural Language Processing models are language specific). By employing Named Entity Recognizers such as Stanford NER (See https://nlp.stanford.edu/software/CRF-NER.html), this task will discover named entities within the transcribed text or in their translations. In the following step, the identified entities will be searched within the Europeana Entity Collection using the search and auto-suggestion functionality of the Entity API. The new entities that are not yet available in the Entity Collection, will be searched within the Linked Data repositories, such as Wikidata, DBPedia, Geonames, and proposed for ingestion within the scope of Task 2.1. All resolvable named entities will be used for semantic enrichment of cultural heritage objects by using their Europeana URIs or their Linked Data URIs.

Task 2.4. Quality control and validation
The accuracy of semantic enrichments and translations will be enhanced through manual verification and crowdsourcing. Currently, the moderation functionality in the Europeana Annotations API provides support for disapproving and automatic disabling of semantic enrichments (i.e. automatic or user generated annotations). This functionality will be enhanced to support the approval of automatic enrichments by regular end users or application moderators. Starting from the automatic translations of transcribed text, end users will have the possibility to provide improved text translations and to validate them by using the same approval process. Only validated semantic enrichments and translations (i.e. those that are confirmed of being correct) will be visible for the public information consumers.

The automatic enrichment services complement and support the existing functionality in the Transcribathon platform that allows end users to manually provide contextual information on the Europeana records. The main improvements and benefits of these services include:

- Enhancing accessibility to historical documents for the large public by machine translation, as the language used in most of the original documents is not familiar to individuals
- Enhancing the understandability of historical documents by integrating relevant information from Linked Open Data resources (i.e. Europeana Entity Collection, Wikidata, DbPedia)
- Improving discoverability of materials in Transcribathon platform by enabling search in different languages, interlinking trough semantic enrichments or enhanced browsing capabilities (e.g. based on newly identified locations, persons or geo-locations)
- Better support for manual enrichments through automatic suggestions, autocompletion or entity disambiguation

Architecture and Design

The aim of the service is to analyse the descriptions and transcriptions of historical documents in order to enrich them with contextual information and enhance their understandability for public users. This is achieved by identifying references to known entities available in semantic repositories including Europeana Entity Collection, Wikidata, DBpedia. The semantic enrichments will be made available

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

in a standardized format using the Web Annotations Data Model, which is compatible with the specifications of the Europeana Annotations API.

The service integrates mature natural language processing and named entity recognition solutions in order to provide a good quality of service. When available and offering good performance level, the models tailored for the original document will be used (e.g. for English, German, eventually Italian languages), otherwise the English translations of the text documents will be used for computing the enrichments (as English models are the most advanced ones).

The goal is to develop a generic service that can be used even beyond the scope of the EnrichEuropeana project.

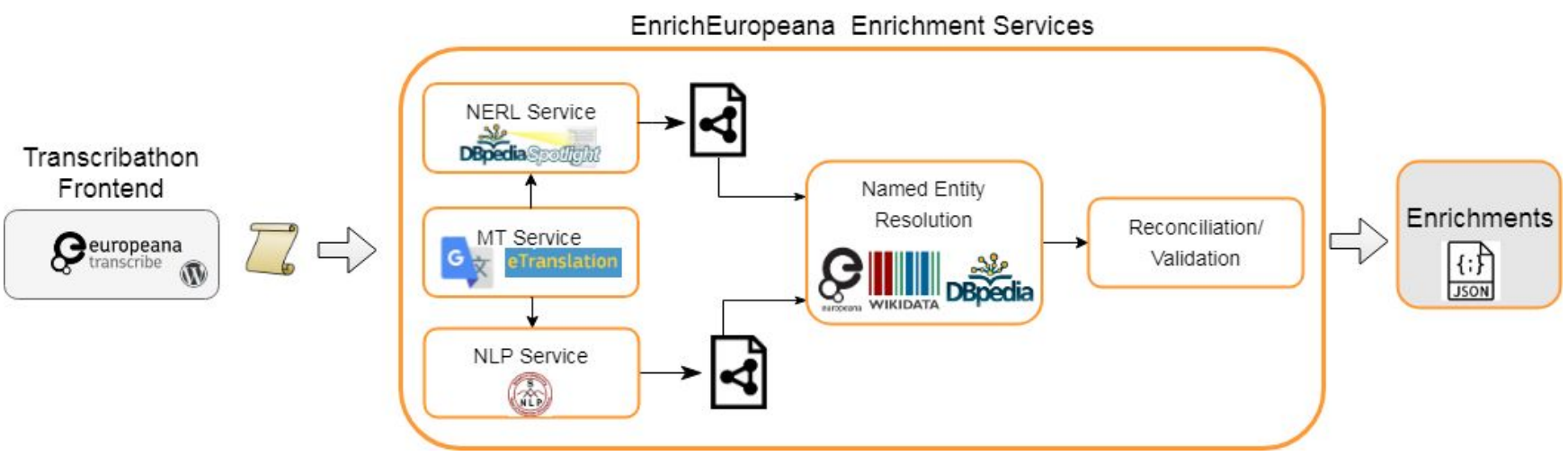


Figure 1 Architectural overview of Semantic Enrichment Services

Figure 1 presents the architectural overview of the Semantic Enrichment Services developed within the EnrichEuropeana project, including the information flow and the tools integrated to compute the automatic semantic enrichments. The input for these services is represented by the transcription and metadata (description and summary) of documents acquired through crowdsourcing activities within the frontend application of the Transcribathon platform (further details will be available in the technical documentation of the following milestone M4: Launch of EnrichEuropeana Platform). The service uses two processing pipelines to analyze textual information, identify named entities and to link these entities with the entries in the targeted linked data repositories:

- One pipeline is based on tools like DBpedia Spotlight which provides an integrated solution for named entity recognition and linking (NERL Service). The DBpedia resources are linked with Europeana Entity Collection and Wikidata entries within the Named Entity Resolution module.
- The second pipeline is based on natural language processing (NLP Service) tools that offer support for named entity recognition and classification like Stanford NER, in which case the linking is performing by matching identified entities with the labels of data items in Wikidata and Entity Collection.

The results of the two pipelines are combined within the Reconciliation/Validation module, which performs the disambiguation of identified named entities by cross-validating the results of the two pipelines. Finally, the semantic enrichments are exposed using the Web Annotation Data Model as implemented within the Europeana Annotations API.

Preliminary investigation and experimental evaluations confirm that Spotlight and Stanford NER provide good accuracy for person and location recognition, by using native models for English and German languages (see Section Experimental Results). They also confirmed that native models for other languages are not mature enough or they are not available at all. Consequently the textual information available in all other languages should first be translated into English.

The integration with the Europeana Core Services Platform is materialized through the usage of Europeana Entity and Annotations APIs.

Enrichment Workflow

The complete processing workflow implemented within the Enrichment Services is described in Figure 2. It describes the whole process starting with the import of documents from the Transcribathon Frontend and finishing with the generation of semantic enrichments and previews of named entities based on the information available in Wikidata repository. The generated entity previews include the same information and are serialized to the JSON-LD format that is used in the *suggest* method of the Europeana Entity API.

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

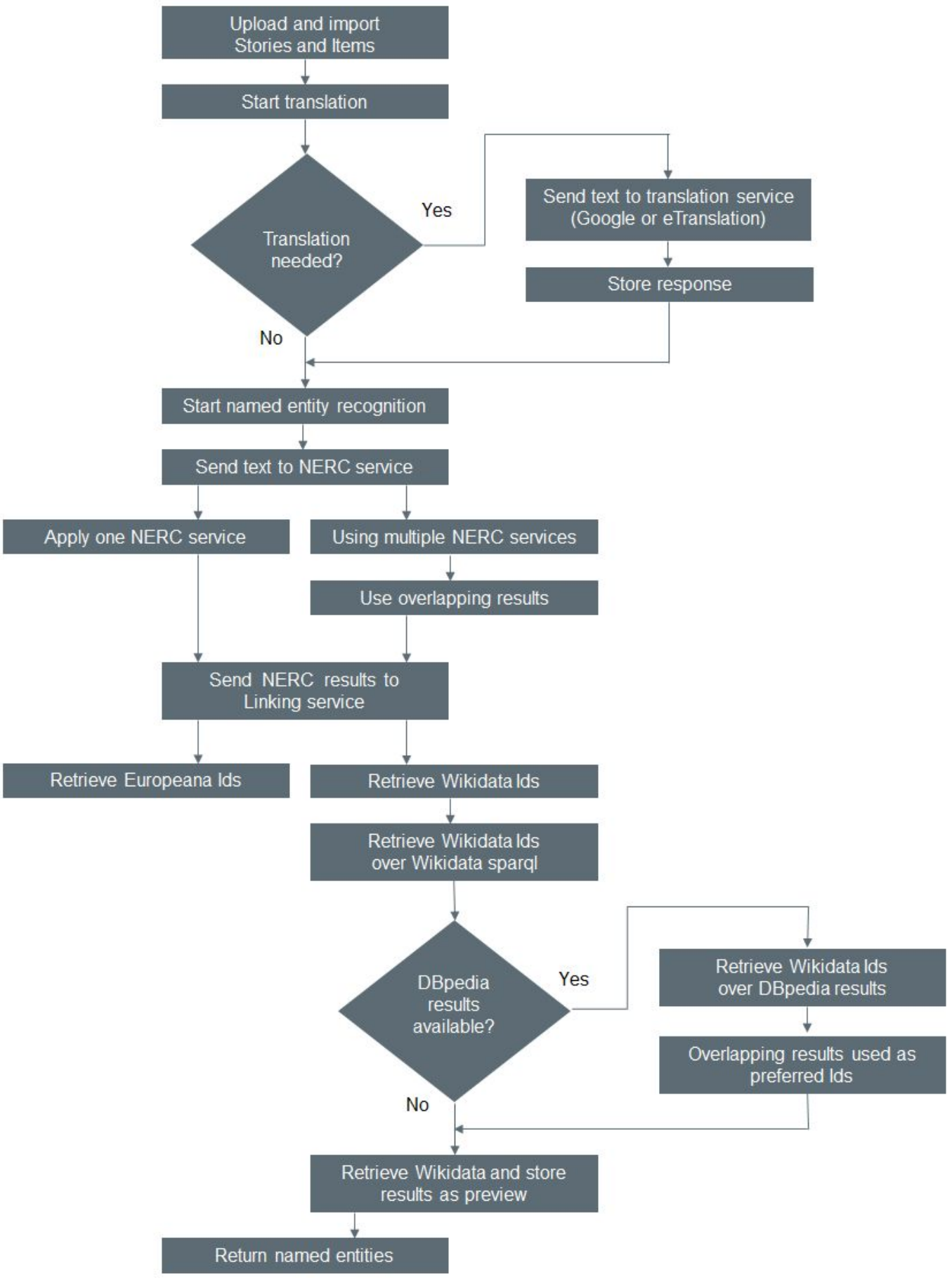


Figure 2: Enrichment Workflow

The enrichment workflow includes the following data processing steps:

- **Upload and import Stories and Items**: The descriptions and transcriptions of Transcribathon stories and items are imported in the database of Enrichment Services. The import is currently based on a database export, the import using the new developed APIs will be implemented for the next version of the services. The documents are also indexed into a solr server at ingestion time. Additionally, the transcription of the story is processed for eliminating the HTML markup used in the Transcribathon platform. This plain text representation can then be used by enrichment services.
- **Start Translation:** The descriptions and translations of Transcribathon stories and items are translated to English in order to make them available for named entity recognition processing, to enhance the accessibility and to improve discoverability through the search functionality. Google Translate and eTranslation services are available to be used for translating documents, and the translations are stored into the database and solr index as well. However, the experimental evaluations indicate that the output of eTranslation is altering the labels of named entities and, consequently, these translations are not used for the named entity recognition currently applied.
- **Start named entity recognition**: The invocation of named entity recognition and classification (NERC) processing can be run separately for the individual services available (i.e. apply one NERC service), or multiple services can be run with one

command (i.e. using multiple NERC services). In the second case, the results provided by individual services will be cross-validated and only the non-ambiguous results will be used in the following processing steps. The output of this processing step includes the labels of the recognized named entities and their categorization. While more classes (i.e., categories) of entities are identified by the employed tools, only Persons and Locations are considered to be relevant for the enrichment of historical documents.

- **Send NERC results to Linking service**: The linking service is used to resolve entities against items in semantic repositories by using their labels. While DBpedia Spotlight resolves itself entities against language-specific DBpedia versions (e.g. English or German DBpedia), the service fetches the data from the repository to retrieve the identifier (URI) of the Wikidata resource, that is considered to be the primary source of information. In contrast, the NLP based pipeline is using the recognized entity name, its category and the languages of the text to search for entries in the Wikidata repository. The *suggest* method of the Europeana Entity API is used to retrieve the URIs for the entities of the Europeana Entity Collection.
- **Retrieve Wikidata resource descriptions:** The description of linked resources is retrieved from Wikidata in order to generate entity previews. The format used for representing the previews is based on the information included in the *suggest* method of Europeana Entity API. These previews may be used by end users to verify the correctness of enrichments that are automatically generated.
- **Generate Semantic Enrichments:** The final step of the processing workflow consists in the serialization of automatic enrichments using Web Annotation Data Model, following the representation supported by the Europeana Annotations API. Currently, only enrichments with higher confidence (acquired through the cross-checking results of the two processing pipelines) are exposed.

The individual steps from the enrichment process can be performed individually by using the REST API described in Section Service Description.

After this process, we indent to display the automatically generated enrichments for manual verification within the Transcribathon Frontend.

Summary of requirements

Semantic Enrichments Services include Machine Translation and Named Entity Recognition. The development of these services is driven by the identification of the functional and non-functional requirements, which are presented in the following:

Translation Service

- Must have
  - translate text from any other language to English
  - translate text from English into the user's preferred language
  - perform translations using Google Translate or eTranslation
  - serialize transcriptions and translations using the Web Annotation Data Model, in compliance with Europeana Annotations API
- Nice to have
  - automatically detect incomplete/faulty translations (e.g. identify transcriptions that contain fragments in different languages, automatically detect language of transcription if not indicated by end users)
  - detect abbreviations in the text (they create issues for automatic translations)
  - text pre-processing (e.g. eliminate html markup, merge words split between lines, remove line endings in the middle of the sentence)

Named Entity Recognition Service

- Must have
  - Entity recognition on original text (description and transcription) for English and German
  - Entity recognition on translated text (description and transcription) for languages other than English and German
  - Linking recognized entities to semantic web repositories including Wikidata, Europeana Entity Collection, DBpedia
  - Cross-validation of semantic enrichments computed by NERL and NLP processing pipelines
  - Represent text positions for recognized named entities (directly based on NER output when NER is applied on the original text and using extra text alignment when NER is applied on translations)
  - Serialization of semantic enrichments using the Web Annotation Data Model, in compliance with Europeana Annotations API
  - Generation of entity previews based on Wikidata information
  - Retrieving of all enrichments for a given entity
- Nice to have:
  - Create statistics on enrichment process
  - Store and Index recognized but not linked named entities
  - Suggest named entities for keywords provided by end users
  - Entity recognition on Italian transcriptions
  - Topic recognition (possibly as a separate service, as named entities and topics can be treated separately)
  - Clustering documents by topics (possibly as a separate service)

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

Service Description and Developer Console

The service functionality is exposed through REST interfaces in order to simplify integration within the Transcribathon frontend and third party applications (e.g. national aggregator or content provider websites). The specifications are following the conventions and the technology stack used for development of Europeana Core Services APIs. Consequently, the services are empowered with a Swagger console, which enables application developers to build and test correct HTTP requests for individual functionalities (see Subsection Developer Console).

The examples used for showcasing the functionality of the enrichment service use the description and transcription of Dumitru Nistor's Journal. The loading of documents through the administration console is a prerequisite for the invocation of translation and named entity recognition services. The client applications must authenticate themselves for each request through the submission of the "wskey" parameter, which is consistently used throughout all REST API methods.

Translation service

The translation service is used to translate the transcriptions into English, descriptions and summaries of documents from the Transcribathon Platform.

Translation request

The translation service expects an authentication parameter "wskey" and the parameters required to identify the text information submitted to external machine translation services. The document is identified by its identifier in the Transcribathon platform ("storyId"), the textual type of the document is indicated through the parameter "type" and the machine translation engine to be used via the "translationTool" parameter. For the "type" parameter, the user can choose between "summary", "description" and "transcription", whereby "description" is selected as the default value. The supported machine translation engines include "Google" and "eTranslation".

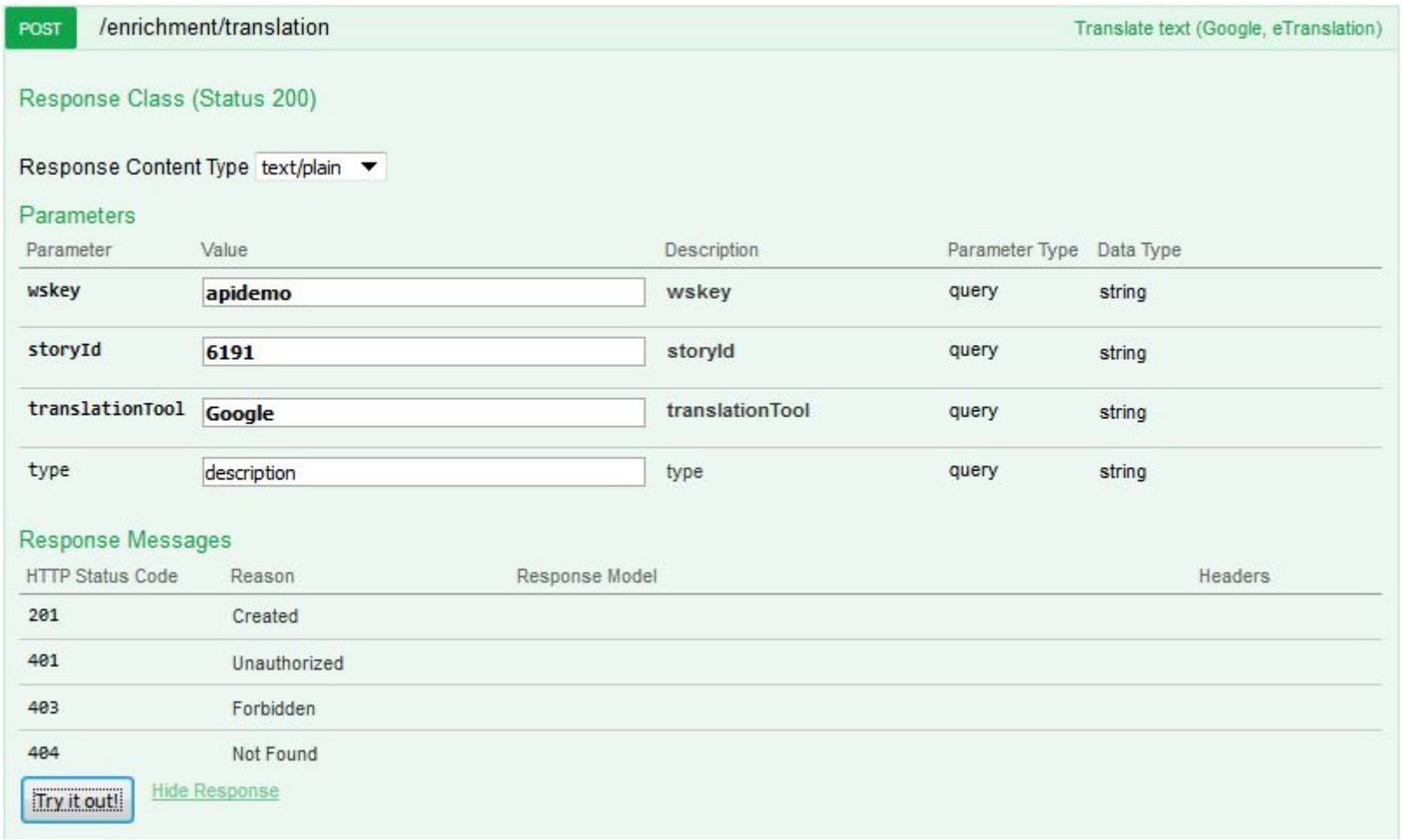


Figure 3: Example of translation request

The response of the translation requests depends on the selected translation tool, as the "Google" API processed translation request in real time, while the "eTranslation" API is processing requests in asynchronous way. Therefore, the text translated by Google or the unique identifier of eTranslation will be integrated into the response. In the case of eTranslation, the user must use the "getTranslation" service to retrieve the translation of the text.
Figure 4 shows the response based on the translation request of Figure 3, where the request for a translation of the Dumitru Nistor's Journal description has been submitted. The response body is retrieved as plain text.

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

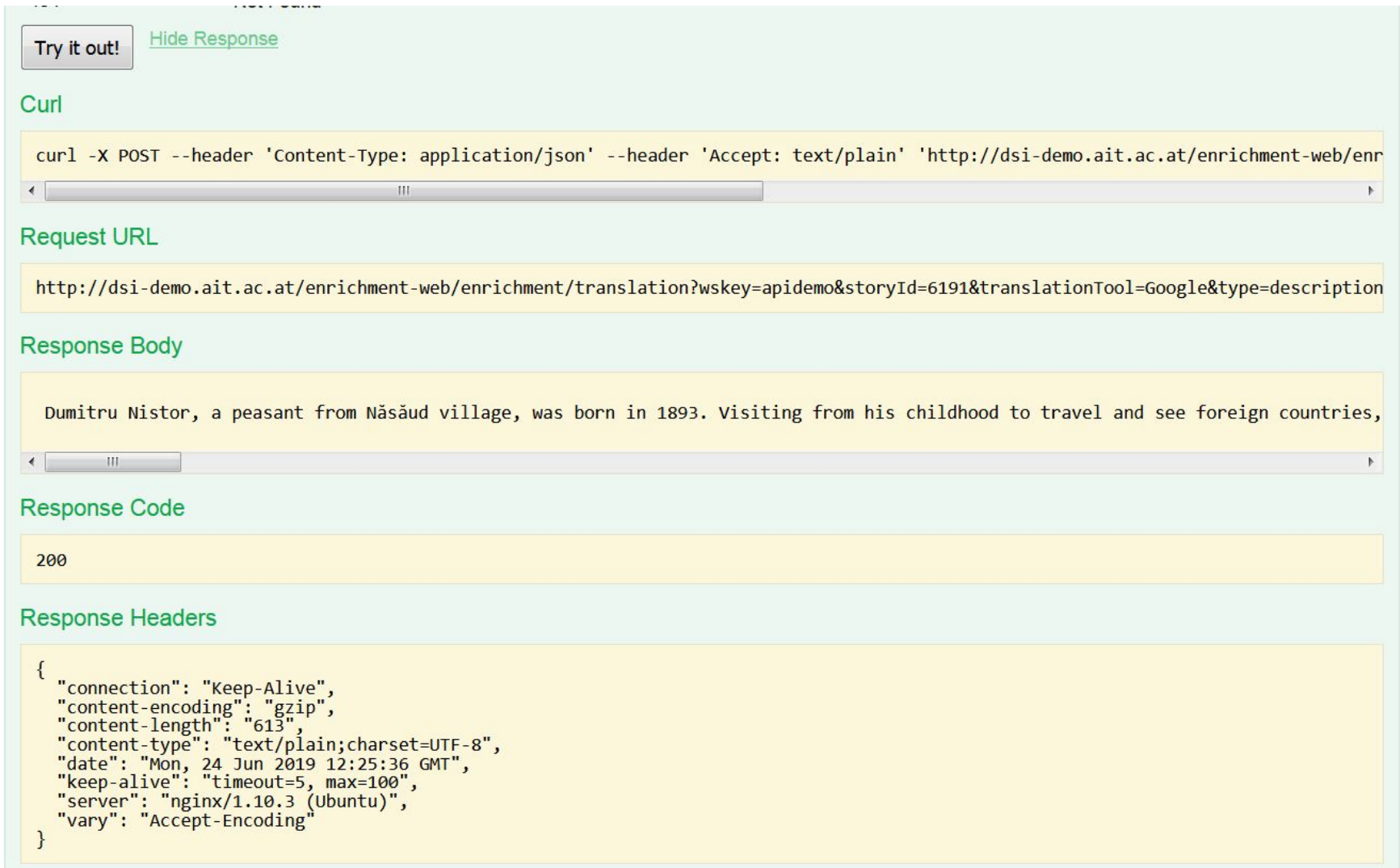


Figure 4: Example of translation response

Retrieve translation request

The retrieve translation method is used to fetch the translation of textual information for Transcribathon documents. This request doesn't perform the translation of the text, but it retrieves it from the database in the case that the translation was already performed using the translation request. "storyId" represents the identifier of the document in Transcribathon platform. The "property" parameter indicates which textual information will be translated. Supported values are "summary", "description" and "transcription". The "translationTool" parameter indicates which machine translation tool will be used for performing the translation, with supported values: "Google", "eTranslation". In the case that the translation was not yet performed, an error message is returned indicating that the translation request needs to be issued first.

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

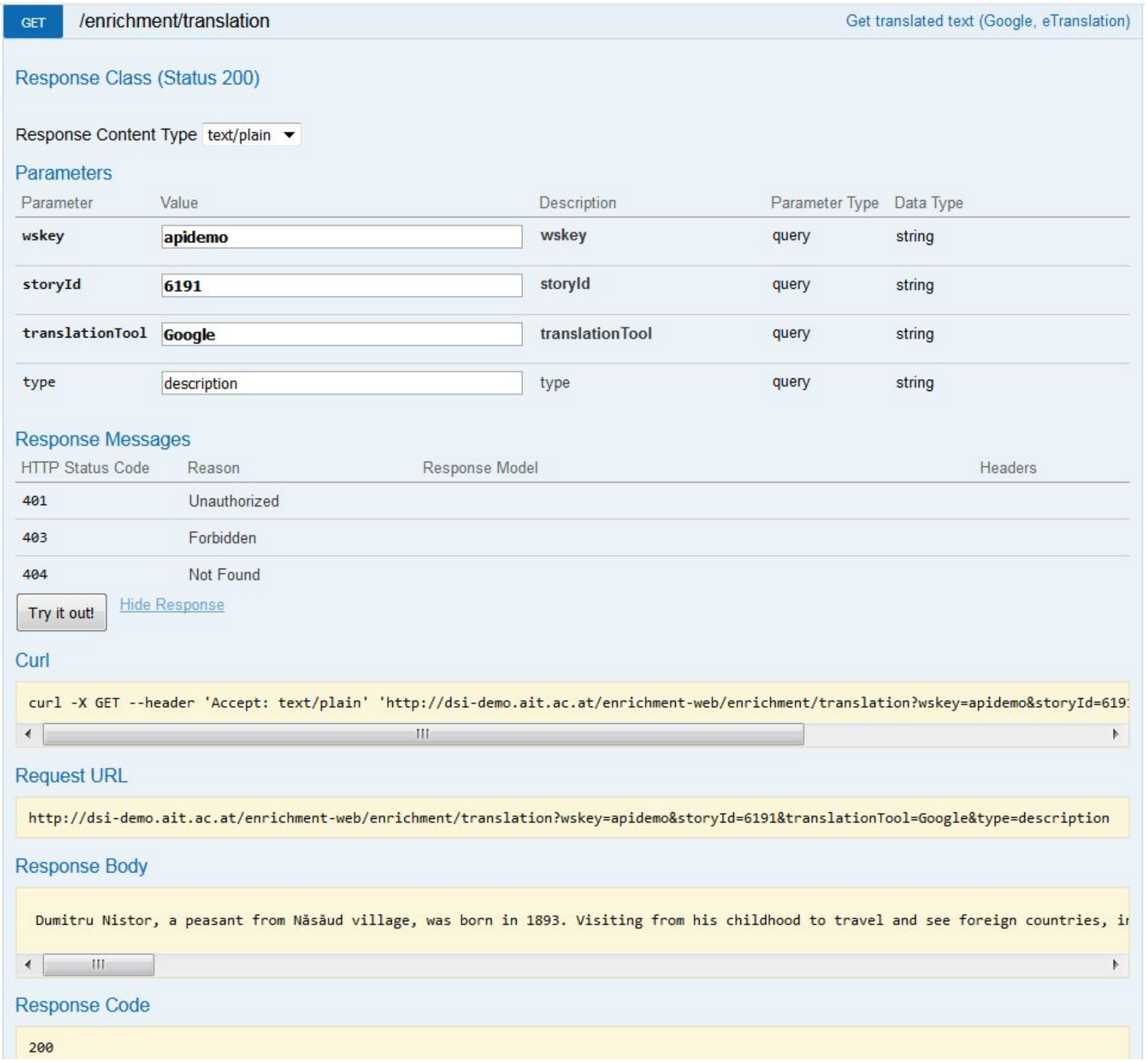


Figure 5: Example of retrieve translation request

Enrichment Service

The enrichment service implements the functionality for named entity recognition, classification and linking. Additionally the enrichments are retrieved in JSON-LD format using the W3C Web Annotation standard. The document descriptions, summaries, transcriptions and their translations are used to identify references to named entities. The service functionality is presented in the following using the translated description of Dumitru Nistor's Journal as input.

Retrieve enrichments request

The semantic enrichments computed with the named entity recognition and linking services are stored in the database and can be accessed through the retrieve enrichments method. Similar to the translation service, the document for which the enrichments are retrieved is indicated through its identifier within the Transcribathon Platform (i.e. "storyId" parameter[1]). All enrichments for the given document are included in the response within a collection of annotations (i.e. type of JSON-LD response is AnnotationCollection). The individual enrichment are serialized according to the semantic tagging specifications within the Europeana Annotation API. Within this version, the target resource is the resource URI of the document within the Transcribathon Platform. After integration with the Data Exchange Infrastructure it will be possible to use the Europeana record identifiers within the target. The body of the annotation includes the Wikidata entity identifier.

[1] In the next rounds of implementation we will include another parameter that would allow to invoke the service at the level of individual items.

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

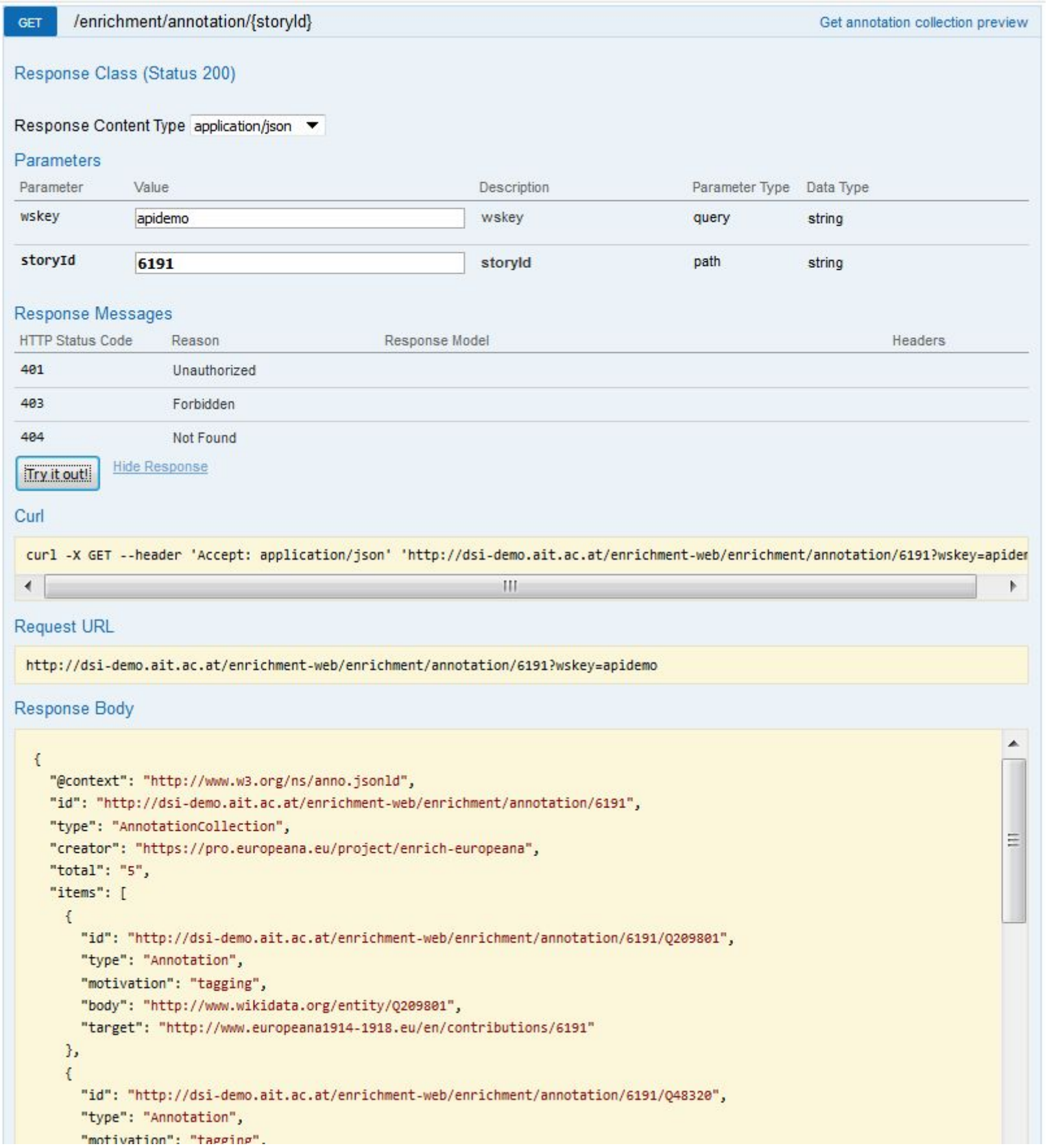


Fig 6: Example of retrieve enrichments request

Retrieve individual semantic enrichments request

Individual semantic enrichments are retrieved by their URIs, which are built from base URL of the service, the identifier of Transcribathon document (i.e. storyId) and the identifier of the entity used for enrichment (i.e. wikidataIdentifier). The retrieval of an individual semantic enrichment is showcased in Figure 7. An extended representation of the annotation will be implemented within the next version of semantic enrichment services, which may include the entity preview as well (see Subsection **Retrieve entity preview request**).

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

GET /enrichment/annotation/{storyId}/{wikidataIdentifier} Get annotation preview

Response Class (Status 200)

Response Content Type application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
wskey	apidemo	wskey	query	string
storyId	6191	storyId	path	string
wikidataIdentifier	Q651322	wikidataIdentifier	path	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out! Hide Response

Curl

```
curl -X GET --header 'Accept: application/json' 'http://dsi-demo.ait.ac.at/enrichment-web/enrichment/annotation/6191/Q651322?wske
```

Request URL

```
http://dsi-demo.ait.ac.at/enrichment-web/enrichment/annotation/6191/Q651322?wskey=apidemo
```

Response Body


```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://dsi-demo.ait.ac.at/enrichment-web/enrichment/annotation/6191/Q651322",
  "type": "Annotation",
  "motivation": "tagging",
  "body": "http://www.wikidata.org/entity/Q651322",
  "target": "http://www.europeana1914-1918.eu/en/contributions/6191"
}
```


Response Code

```
200
```

Fig 7: Example of retrieve individual semantic enrichment request

Retrieve entity preview request

With the usage of resolve method, the end users or applications are able to retrieve the preview of named entities included in the semantic enrichments. Figure 8 showcases the retrieval of a preview for the wikidata entity representing the Năsăud city (i.e. first enrichment in Figure 6). The method retrieves the description from Wikidata for the resource indicated through the “wikidataId” parameter, and saves it in the Solr based storage of enrichment services. Subsequent requests are displaying the data from the local storage, without issuing request to Wikidata. The serialization of the entity preview follows the specifications for the output of the *suggest* method in the Europeana Entity API. Additionally to the fields of previews in the Europeana Entity API, the description, wikipedia links and entity type specific fields are included in the response. For locations the geo-coordinates and for persons the country of citizenship are included when available in Wikidata.

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

GET /enrichment/resolve Get entity preview

Response Class (Status 200)

Response Content Type application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
wskey	apidemo	wskey	query	string
wikidataId	http://www.wikidata.org/entity/Q651322	wikidataId	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out! Hide Response

Curl

```
curl -X GET --header 'Accept: application/json' 'http://dsi-demo.ait.ac.at/enrichment-web/enrichment/resolve?wskey=apidemo&wikida
```

Request URL

```
http://dsi-demo.ait.ac.at/enrichment-web/enrichment/resolve?wskey=apidemo&wikidataId=http%3A%2F%2Fwww.wikidata.org%2Fentity%2FQ65
```

Response Body


```
{
  "@context": "http://www.europeana.eu/schemas/context/entity.jsonld",
  "id": "http://www.wikidata.org/entity/Q651322",
  "type": "place",
  "prefLabel": {
    "de": "Nussdorf",
    "ceb": "Năsăud",
    "ru": "Нэсэуд",
    "pt": "Năsăud",
    "fr": "Năsăud",
    "hu": "Naszód",
    "yi": "נאסאד",
    "min": "Năsăud",
    "sh": "Opština Năsăud, Bistrița-Năsăud",
    "uk": "Неceуд",
    "id": "Năsăud",
```


Fig 8: Example of retrieve entity preview request

Named entity recognition and linking

The named entity recognition and linking functionality is the core of Enrichment Services. It is developed using Stanford NER and DBpedia Spotlight tools, which are used to recognize persons and locations within the description, summary, transcription of Transcribathon documents, or on their translation when NLP models are not available for the language used in the original text. In addition to the recognition of named entities, this module performs also the linking with resources from semantic web repositories.

Named entity recognition and linking request

Figure 9 presents the swagger console used to build and test requests sent for performing named entity recognition and linking of textual information available in Transcribathon documents. The text used for performing entity recognition is identified though the “storyId” parameter representing the identifier within the Transcribathon platform and by the “property” parameter indicating if the description, summary or transcription should be used. Additionally, the “original” parameter specifies if the processing will be applied on the text in the original language (i.e. only available for English, German) or on the text translation. In the latter case, the “translationTool” parameter must be submitted indicating if the translation with Google or eTranslation should be used. The tools used for named entity recognition are indicated through the “nerTools” parameter, for which the “Stanford_NER” and “DBpedia_Spotlight” option can be used, together or individually. When the tools are applied together, the results will include only the cross validated named entities, that were identified by both tools. The linking to DBpedia resources is performed automatically by Spotlight, through the “linking” parameter, the user has the possibility to resolve found entities against Wikidata and Europeana repositories (i.e. Wikidata and Europeana are the valid options for “linking” parameter).

POST /enrichment/entities — Get named entities

Response Class (Status 200)

Response Content Type application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
wskey	apidemo	wskey	query	string
stroyId	6191	stroyId	query	string
translationTool	Google	translationTool	query	string
property	description	property	query	string
linking	Wikidata	linking	query	string
nerTools	DBpedia_Spotlight,Stanford_NER	nerTools	query	string
original	false	original	query	boolean

Response Messages

HTTP Status Code	Reason	Response Model	Headers
201	Created		
401	Unauthorized		
403	Forbidden		
404	Not Found		

Try it out! Hide Response

Fig 9: Example of named entity recognition and linking request

Named entity recognition and linking response

Figure 10 presents the response of the named entity recognition and linking operation, when the request presented in Figure 9 is submitted. The response indicates the named entities identified by Spotlight (i.e. “DBpediaIds” field) and the resources found in Wikidata for named entities recognized by Stanford_NER (i.e. “WikidataIds”). The disambiguation of linking results is based on identified DBpedia resource and the corresponding Wikidata entity is returned within the “preferredWikidataIds” field. The position of the named entity within the textual information is indicated within the “positionEntities” field. In the given example, the Năsăud city was found in the english translation of the document description starting at the 31st character.

Request URL

```
http://dsi-demo.ait.ac.at/enrichment-web/enrichment/entities?wskey=apidemo&stroyId=6191&translationTool=Google&property=descriptio
```

Response Body


```
{
  "place": [
    {
      "positionEntities": [
        {
          "storyId": "6191",
          "storyFieldUsedForNER": "description",
          "offsetsTranslatedText": [
            31
          ]
        }
      ],
      "preferredWikidataIds": [
        "http://www.wikidata.org/entity/Q651322"
      ],
      "DBpediaIds": [
        "http://dbpedia.org/resource/Năsăud"
      ],
      "wikidataIds": [
        "http://www.wikidata.org/entity/Q651322",
        "http://www.wikidata.org/entity/Q16898128"
```


Response Code

```
200
```

Fig 10: Example of named entity recognition and linking response

Named entity recognition and linking retrieval request

After performing the named entity recognition and linking on a given Transcribathon document, the result of the operation can be retrieved from the enrichment database without running the text processing again. Figure 11 shows the swagger console building the request for retrieval of the named entity recognition and linking results. The parameters and the response are the same as in the case of the request for performing the entity recognition indicated in Figure 9 and Figure 10. The only differences are the usage of the HTTP GET method, and the user notification in the case that the entity recognition was not yet performed.

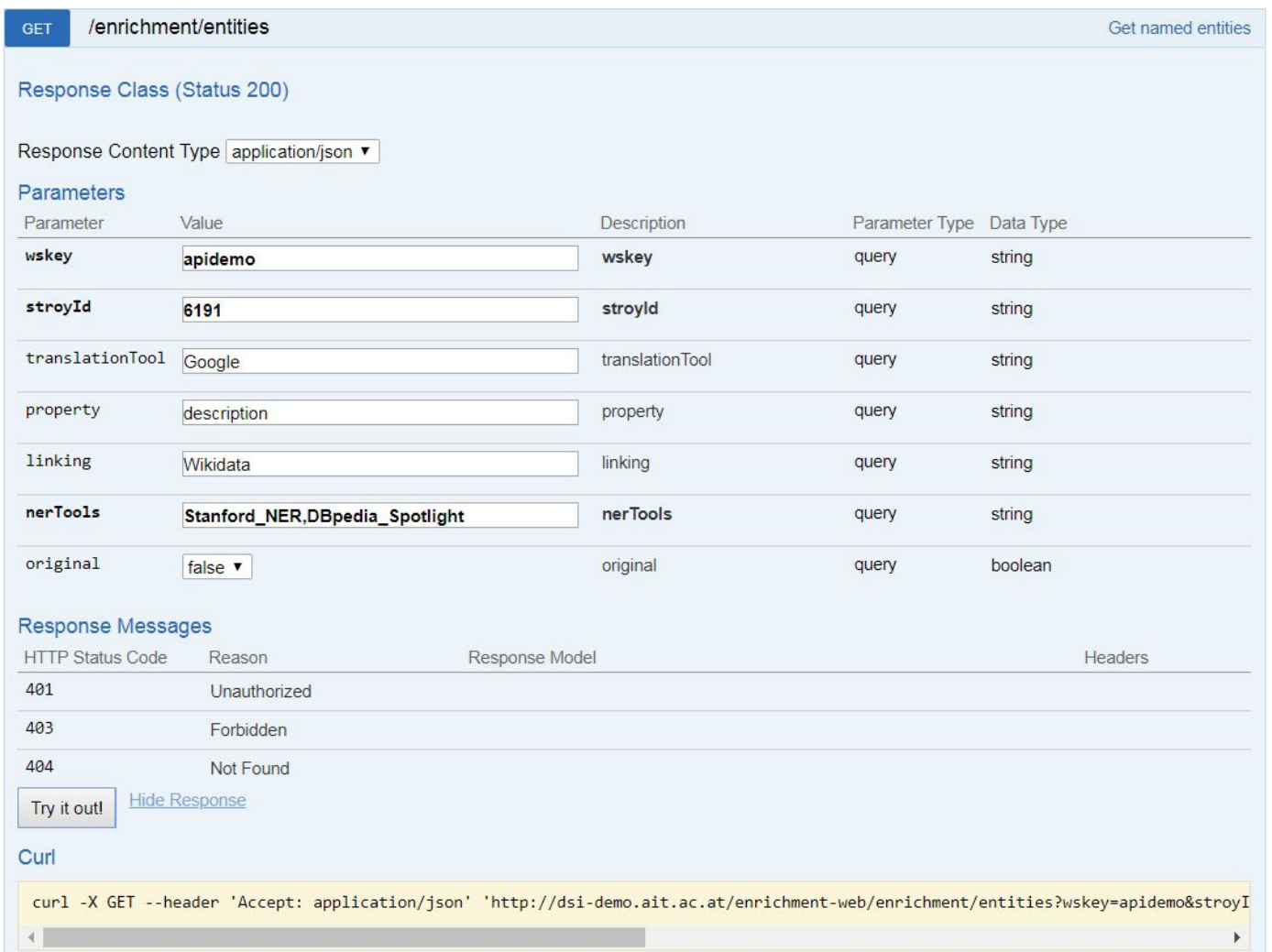


Fig 11: Example of named entity recognition and linking retrieval request

Admin Console (Swagger)

Upload service

The Upload Service is used to import Transcribathon stories and items into the enrichment services database. This is a prerequisite for invocation of translation and semantic enrichment services. The upload service provides two data import options. The first option allows the system administrators to import database dumps from the Transcribathon backend (i.e. “uploadStoriesAndItemsFromJson”). The second method supports the direct upload of individual stories and items in the semantic enrichment database (i.e. “uploadStories” and “uploadItems” methods).

Import database dumps

The files containing the database dumps of Transcribathon stories and items have considerable size. Consequently, they are not well suited for being submitted through Http requests. For importing these files into enrichment database the import request presented in Figure 12 is used by the system administrator. The database dump files need to be physically uploaded on the enrichment services server as their locations are used as input parameters for the import method. At import time, the full transcription for stories is generated through the concatenation transcriptions of individual story items. Additionally, the transcription of the story is processed for eliminating the HTML markup used in the Transcribathon platform. Consequently, the plain text representation is generated as it is used by enrichment services.

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

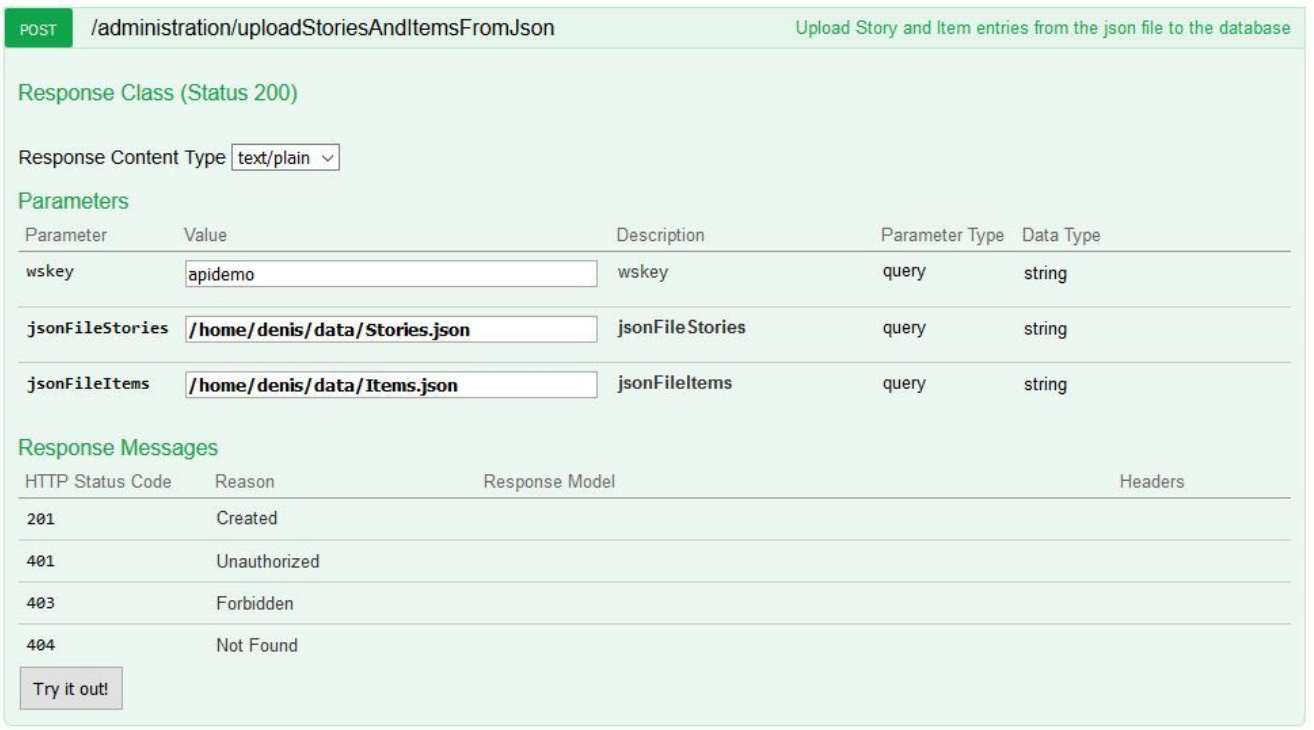


Fig 12: Example of request to import stories and items from database dump files

Upload stories service

The upload stories service can import or update provided stories into the enrichment database. The request body data format is a list of JSON objects whose properties cover the key features such as the storyId, title, source (e.g. http://www.europeana1914-1918.eu/en/contributions/6191), description, summary, language and transcription of a story.

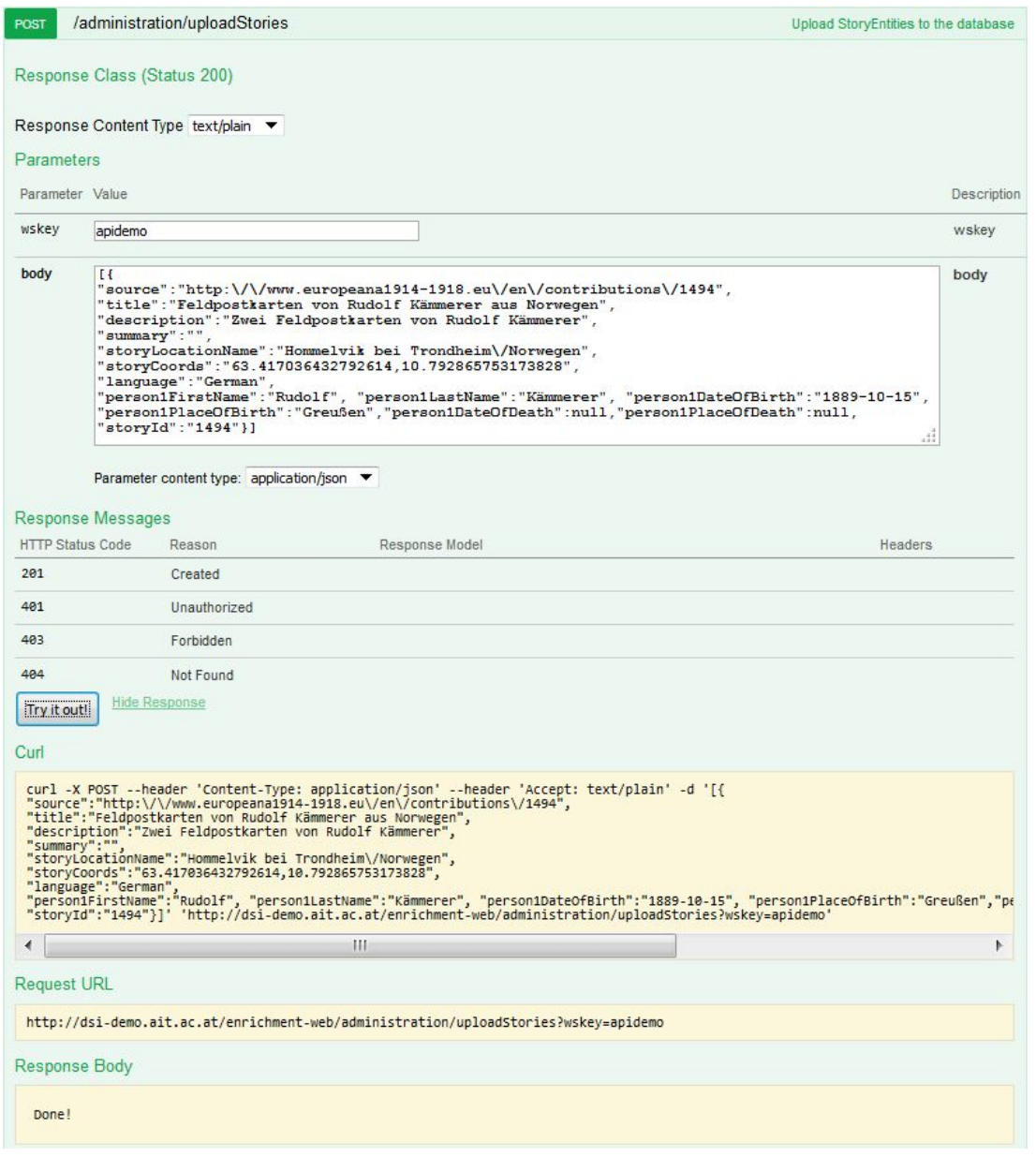


Fig 13: Example of upload stories request

Upload items service

The upload items service is used to import or update story items into enrichment database. The request body contains a list of JSON objects representing the item properties such as storyId, title, type (e.g. "letter" or "diary" or "picture"), language and transcription. A hash value for the transcription text is generated in order to easy verify when updates to item transcription are submitted to enrichment services.

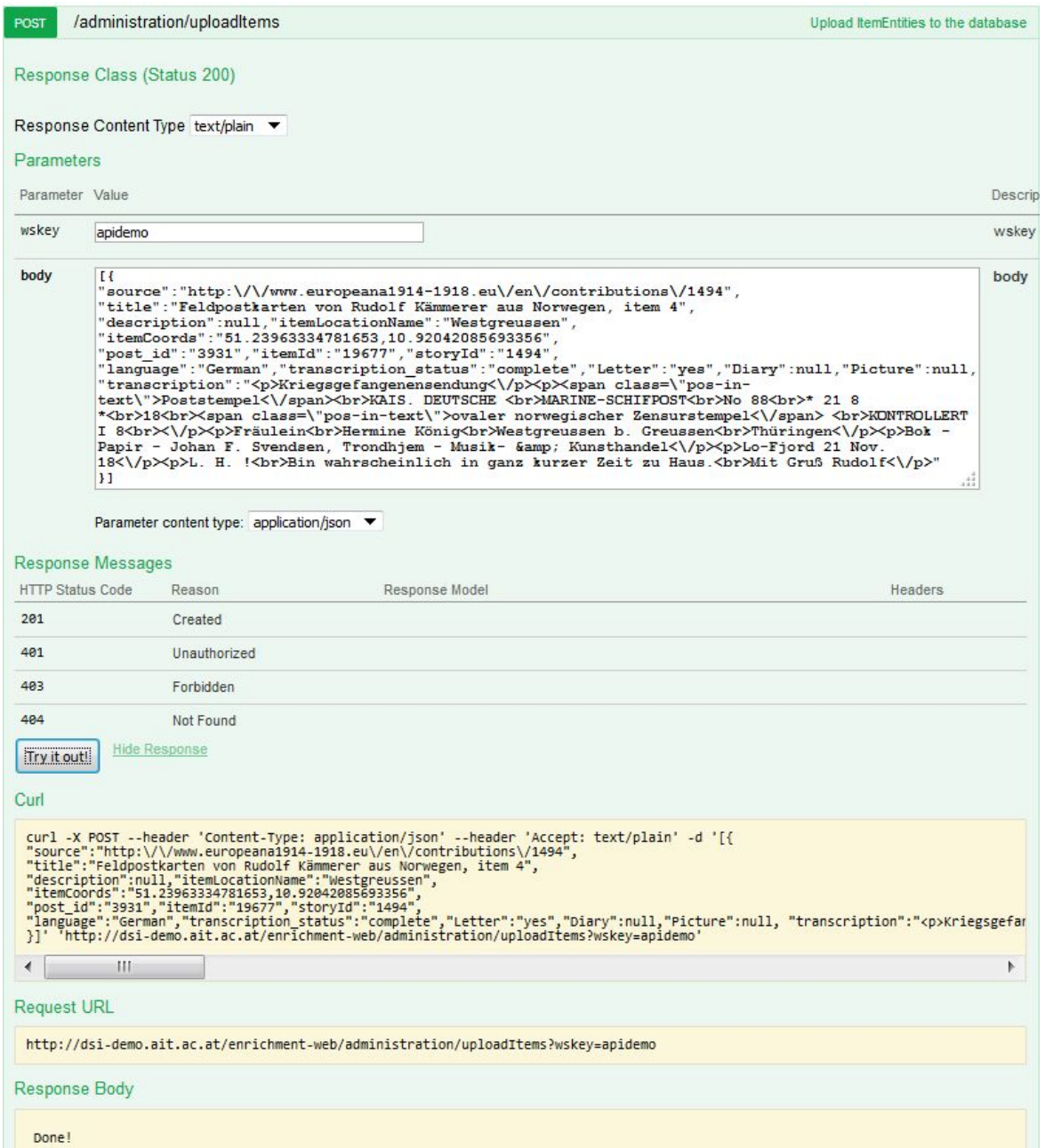


Fig 14: Example of upload story items request

Experimental Evaluations

The semantic enrichment workflow involves usage of machine translation (MT) and named entity recognition (NER) technologies. The quality of results provided by these tools is highly related to the quality of the provided text. The models used by these services were built by taking into account the grammar and the vocabulary of modern languages. Given the fact that the documents used by EnrichEuropeana are historical documents, we performed preliminary evaluations to assess the quality of MT and NER using 3 different documents available through the Transcribathon platform (in German, Italian and Romanian). The conclusions of the experiments indicate that these technologies can be successfully used to perform semantic enrichment of transcribed documents from the Europeana 1914-1918 Collection. This section presents our experimental evaluation and its conclusions, which were used to finalize the design and the architecture of the semantic enrichment services.

Experiment description

Three war diaries were selected for the experimental evaluation, namely the journals of Dumitru Nistor (in Romanian), Eduard Scheer (in German) and Bruno Celestino (in Italian). As all historical documents, these journals are written in a manner that doesn't comply with the rules of modern languages, which are the basis for training machine translation and natural language processing models. We assessed the impact on the quality of service by comparing the results of the original text transcriptions and a version that was manually corrected by native speakers.

The first goal of this experiment was to compare the performance of different named entity recognition tools on texts containing references to well known places and persons. The second goal of this experiment was to link the named entities with existing resources from Europeana Entity Collection, Wikidata and DBpedia, thus building a small sample of correct semantic enrichments (ground truth).

Two different approaches for entity recognition and linking were investigated. DBpedia Spotlight offers an integrated solution for entity detection and name resolution. However, the name resolution is performed against the version of DBpedia version specific for given

language, which extracts information from Wikipedia articles of that language. This solution can therefore suffer from the low language coverage of Wikipedia articles in that language[2]. The alternative solution is using state of the art tools for named entity recognition and classification like Stanford NER library and performing the entity resolution against Wikidata which has is a larger repository of data items with better language coverage. The Wikidata knowledge base is actively enhanced through systematic data integration activities but also through crowdsourcing activities performed on site. The main drawback of this solution is the need to perform an additional disambiguation of entities when matching the labels of identified entities.

The multilingualism of the data corpus represents a great challenge for semantic enrichment services, as mature natural language processing models are available only for a few of the most used languages. We therefore use machine translation services are used to translate the transcription texts into English. For the journal of Dumitru Nistor, the named entity recognition was evaluated only on the English translation, while for Eduard Scheer and Bruno Celestino diaries the performance of the native models was compared with the one of the English translations.

The evaluated documents were annotated by native speakers in order to create a ground truth with respect to the identification of named entities. The overview of manually identified named entities is presented in Table 1.

Table 1: Overview of ground truth

Document	Locations	Persons
Dumitru Nistor's Journal	41	44
Eduard Scheer's Journal	25	7
Bruno Celestino's Journal	14	18

Preliminary conclusions

The quality of the machine translation was evaluated with regard to the quality of the output text and the preservation of original named entities. Both eTranslation and Google Translate services provide good quality translations for the texts describing the items. However, the transcribed texts contain a large number of misspelled words or incorrect formulations with respect to the current writing standards for the given languages, and also use inconsistent punctuation. This is a great challenge for machine translation services. Many sentences in the translated text are understandable and preserve the original meaning with respect to the original message. However, there is still a large part of the text for which the translation is unacceptable for the human user. Therefore, in order to publish the translations for public access, a manual correction of the translations would be required. Google Translate tends to be more robust wrt. misspelled words than eTranslation: it is able to recognize/translate correctly some misspelled words, while eTranslation typically includes them in the translation as they are written in the original text. Additionally, eTranslation is not able to preserve the original wording for named entities, as the services tend to translate and build derivatives of those words by applying the grammar rules of the targeted language. Consequently, only translations provided by Google Translate service are appropriate for applying named entity recognition.

The quality of named entity recognition and classification of depends on the quality of the analyzed text, while the name resolution depends also on the popularity of those entities (i.e. their existence in the linked data repositories). Several open source libraries implementing name entity recognition were evaluated. Stanford NER provided the best performance for the evaluated documents.

For the analysis of Dumitru Nistor journal, the English translation of the improved transcription was used (i.e. the text being collected from the printed book). The experimental results indicated the best performance on this document. The true positive rate for location is higher than 90% in case of locations in both approaches, using Stanford NER and DBpedia Spotlight. Most of the referenced locations are well known, therefore, the name resolution was successfully performed. Similar results are found for the person recognition, however, most of those entities are not available in Wikipedia, consequently, the true positive rate is very low when using DBpedia Spotlight.

For the Eduard Scheer journal, the original German text and the English translation were evaluated, as native models are available in both languages for the Stanford NER and DBpedia Spotlight. This allows us to compare the performance of NER services using the original language and the machine translations, and to estimate the performance when machine translations are used instead of the original text. The true positive rate for entity recognition is high on this document as well, however the machine translation generated a positive effect. The classification of named entities performed better, seemingly because the machine translation produces a text that is closer to the modern language, especially wrt. the use of prepositions and exprimation of subject-object interactions. The recognition of persons using Stanford NER delivered also a high true positive rate, however the style used by the author to write this document resulted in a high false positive rate, as a large number of places were categorized as persons. Many locations referenced in the document were written with special characters specific to foreign languages (e.g. polish language), which resulted in a lower

[2] see https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages/Inter_language_approach

performance for linking through DBpedia Spotlight. Also, a low rate for name resolution for persons was encountered, which is also related to the fact that most of the persons are not present in linked data repositories.

The diary of Bruno Celestino, a simple soldier, is written in non-typical Italian, the text of the document contains many misspelled words and some grammatical errors. Consequently, the performance of named entity recognition and classification is lower in the case of this document.

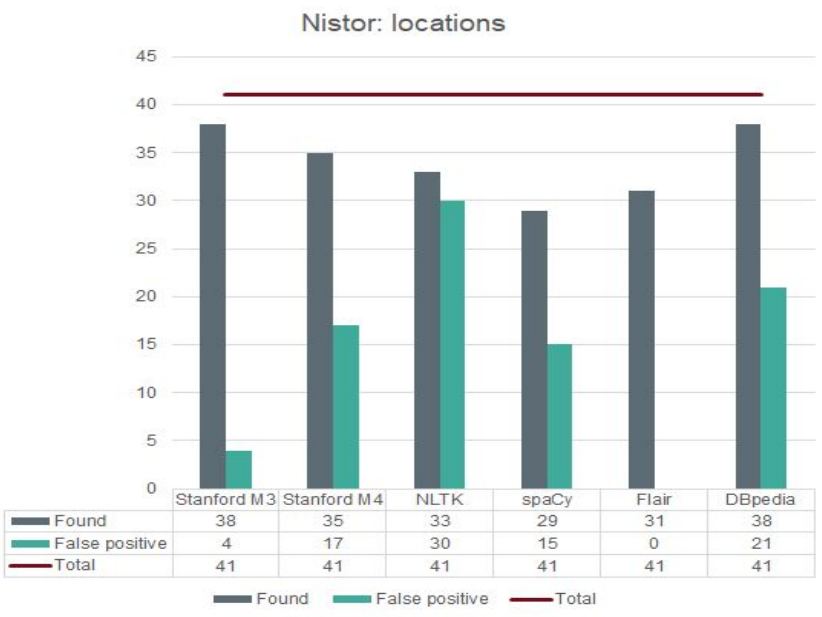
Experimental Results

Named Entity Recognition and Classification

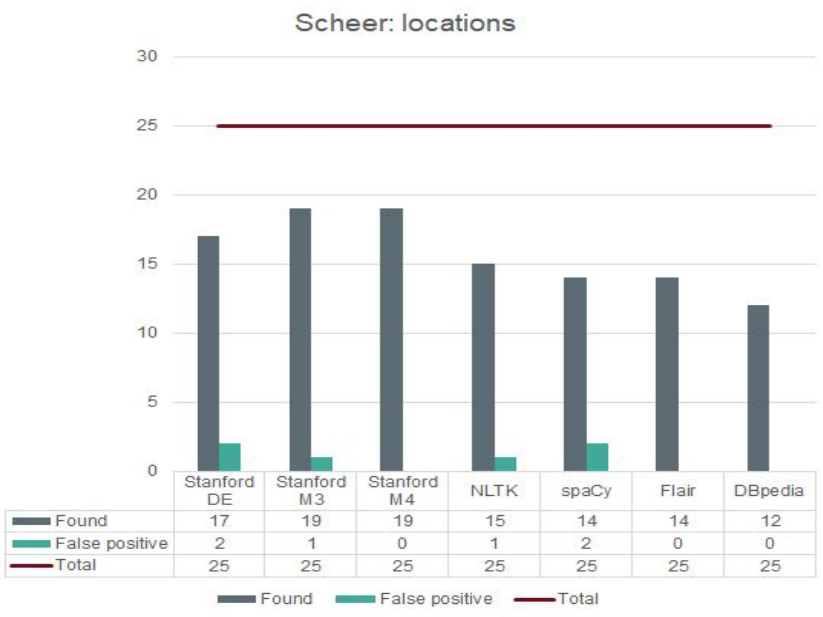
Figure 15 presents the experimental results on detection and classification of locations within the Dumitru Nistor and Eduard Scheer journals. For this experiment we included the evaluation of several open source named entity recognition libraries including: Stanford Model 3, Stanford Model 4, NLTK, Spacy, Flair, DBpedia Spotlight. A short description of these libraries is available in Annex 1.

In the case of Dumitru Nistor journal, Stanford NER library provides the best experimental results, having a true positive rate higher than 90%, for both configurations with Model 3 and Model 4. Model 3 configuration has also the lowest false positive rate, below 10%. The results of NLTK tool indicates also a high true positive rate, but the recognition precision is affected by the high false positive rate. spaCy provides decent results, but the precision is significantly lower than for Stanford, while the current version of Flair has a non deterministic behaviour (i.e. different runs with the same settings generate different results), for which no objective conclusion can be drawn with respect to its performance. The performance of DBpedia Spotlight is also very good, indicating the same true positive rate as Stanford Model 3 and the false positive rate slightly higher than Stanford Model 4. The relatively high false positive rate on this document is a consequence of the fact that it includes quotations in different foreign languages, which cannot be handled correctly by language specific natural language processing models.

The evaluation of recognizing locations on Eduard Scheer journal indicates the best performance when using the Stanford NER library, with a good true positive rate around 70% with all three models, German (DE), Model 3 and Model 4 (English version) and a low false positive rate. However, a couple of recognized entities were wrongly categorized as Persons, which is a consequence of the style used by the author when writing the text. The way of writing the text creates an issue for correct detection of locations for all tools. The lower performance of DBpedia Spotlight is related to the name of locations that are often written in foreign languages and using characters that are not available in german or english label.



	Stanford M3	Stanford M4	NLTK	spaCy	Flair	DBpedia
Found	38	35	33	29	31	38
False positive	4	17	30	15	0	21
Total	41	41	41	41	41	41



	Stanford DE	Stanford M3	Stanford M4	NLTK	spaCy	Flair	DBpedia
Found	17	19	19	15	14	14	12
False positive	2	1	0	1	2	0	0
Total	25	25	25	25	25	25	25

Figure 15: Location recognition on Dumitru Nistor and Eduard Scheer Journals

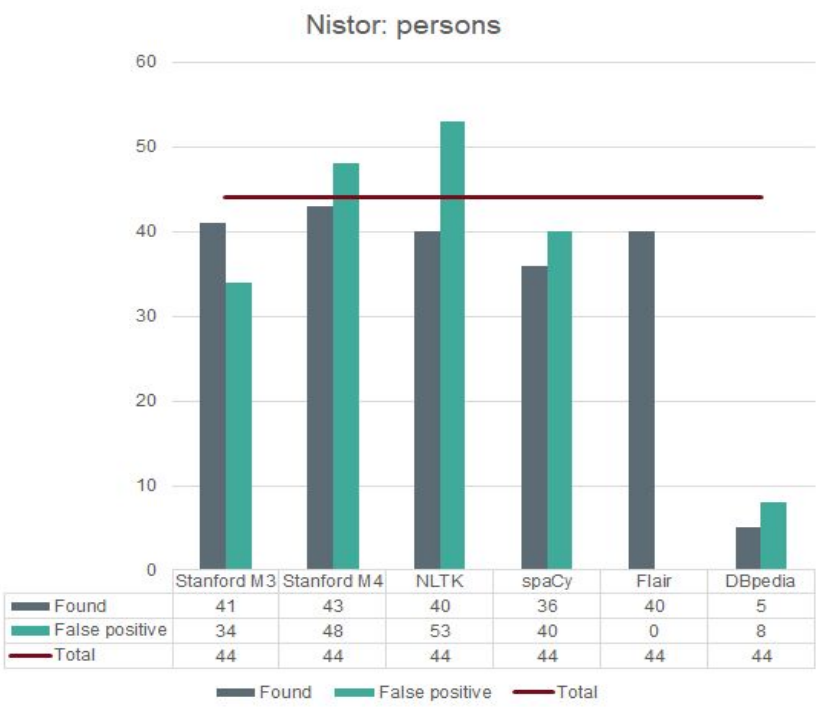
Figure 16 presents the experimental results on detection and classification of persons within the Dumitru Nistor and Eduard Scheer journals. Stanford NER has the best performance also with respect to person recognition on these documents. However, the false positive rate for person detection raises significant issues. Especially, a large part of false positives is generated by the misclassification of locations as persons. When considering this false positive rate, one must be also aware that ethnic groups are also categorized as persons, which conflicts with the approach in the ground truth that takes into consideration only individual persons. DBpedia Spotlight is not able to link many detected person names to web resources (only 15%), as the referenced persons are rather unknown/unpopular individuals. Often, the automatically linked resources does not represent the same person as the Transcribathon documents.

Dumitru Nistor journal is rich in person names, which are well detected by all libraries, still the most of them are indicated only by first name and the area they are coming from. The linking to web resources is impossible to be successfully performed with an automatic

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

process, but they might still be useful to be used as keywords for browsing and searching functionality within the Transcribathon platform.

The ground truth of Eduard Scheer journal indicates only 7 entities, therefore it is hard to draw objective conclusions on person detection for this document. Still, on a relative scale, the performance of named entity recognition tools indicates a similar trend as in the case of Dumitru Nistor Journal.



	Stanford M3	Stanford M4	NLTK	spaCy	Flair	DBpedia
Found	41	43	40	36	40	5
False positive	34	48	53	40	0	8
Total	44	44	44	44	44	44

	Stanford DE	Stanford M3	Stanford M4	NLTK	spaCy	Flair	DBpedia
Found	6	6	7	5	6	6	1
False positive	4	5	13	8	5	0	0
Total	7	7	7	7	7	7	7

Figure 16: Person recognition on Dumitru Nistor and Eduard Scheer Journals

Named entity linking - further detail

The linking of recognized entities with semantic web repository is an important goal of the semantic enrichment services. This will allow end users to get better contextual information, to better understand and retrieve historical documents. Still the main prerequisite for performing the linking of named entities is their existence in potential target semantic web repositories, which is only the case of locations found in the evaluated documents. As Wikidata is the largest semantic repository with a strong support from large companies and volunteer communities, we aim at linking named entities with this repository, but also with Europeana Entity Collection which is already in use within the Europeana Core Services Platform.

Figure 17 presents the evaluation on linking locations from Dumitru Nistor and Eduard Scheer Journals. As expected, the linking to Wikidata resources is more successful, as this is a way larger repository than Europeana Entity Collection. This experimental evaluation is also a good indicator for locations that are relevant for Europeana Core Services Platform, but not yet available or used as such. Over 90% of locations found in Dumitru Nistor Journal were successfully and correctly linked with Wikidata, while only 57% of those were also found in Europeana Entity Collection. For Eduard Scheer only 56% were found in Wikidata (and 32% in the Europeana Entity Collection). Manual investigation indicates that Europeana Collection includes only countries and cities, but it lacks information about historical regions or geographical structures, like water areas, continents or mountains that are often used in historical documents.

Note that the percentage of found entities may significantly improve when taking into account alternative labels in different languages, an option we want to pursue in order to solve issues indicated in previous sections.

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

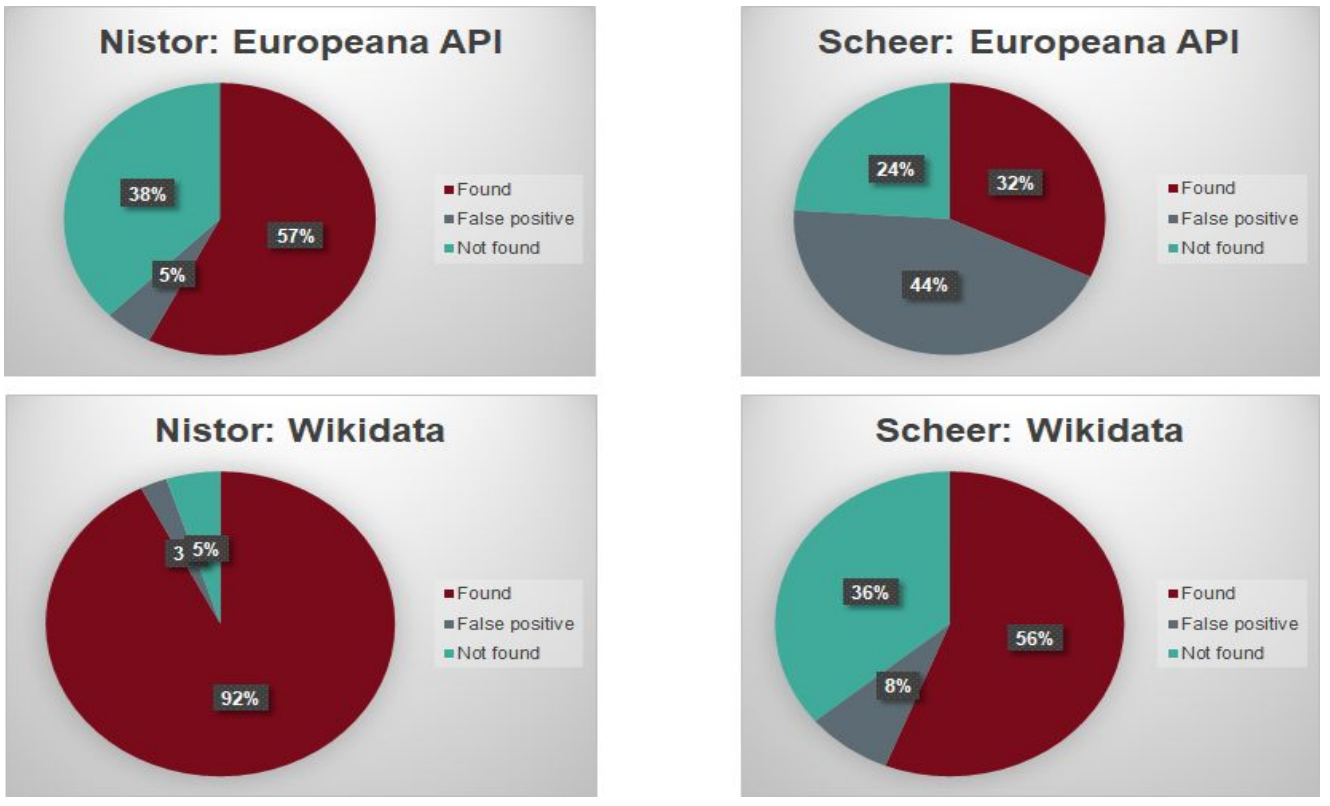


Figure 17: Entity linking on Dumitru Nistor and Eduard Scheer Journals

Named entity recognition and linking for italian language

A second evaluation was performed to assess the performance of native models (i.e. for Italian) for named entity recognition and linking. Latest research and development activities aim at extending mature solutions like Standford NER to provide native support for more languages. TINT NER is based on the classifier included in Stanford CoreNLP and used a model trained on the Italian Content Annotation Bank (I-CAB), containing articles taken from the Italian newspaper L'Adige. Dondelion, similar to Spotlight, provides and API for named entity extraction & linking. It currently works on texts in English, French, German, Italian, Portuguese, Russian, Spanish. With this API it is possible to analyze unstructured text and automatically enrich it with contextual information by linking with Wikipedia entities, semantic tagging and classification.

TINT NER was not able to present a good performance for recognition of locations (i.e. true positive rate of 35%), mainly because of misspellings and grammar mistakes that occur in the original text. DBpedia Spotlight and Dandelion reach a true positive rate higher than 70%, with Spotlight performing slightly better than Dandelion in terms of recognition and correct categorization of entities.

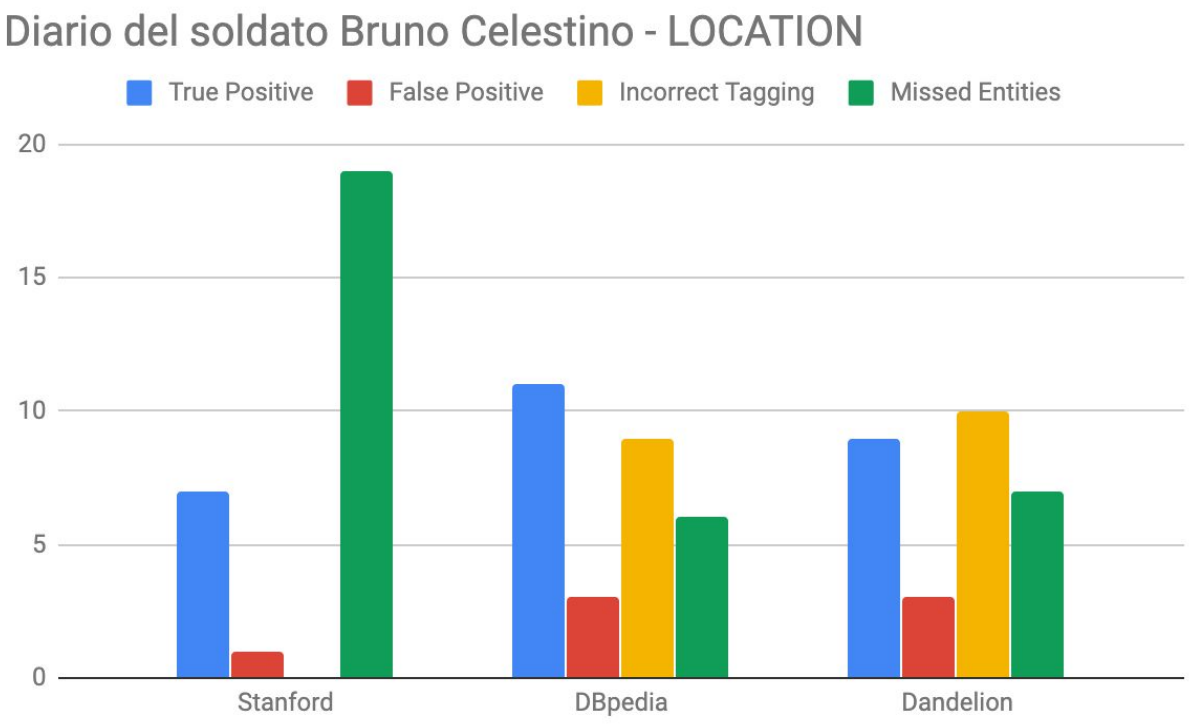


Figure 18: Location recognition on Bruno Celestino Journal

Table 2: Location recognition on Bruno Celestino Journal, results overview

	Stanford	DBpedia	Dandelion
True Positive	7	11	9
False Positive	1	3	3
Incorrect Tagging	0	9	10

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

Missed Entities	19	6	7
Total	26	26	26
PRECISION	0.875	0.7857142857	0.75
RECALL	1	0.25	0.2307692308

TINT NER recognized 4 out of the 6 total persons in the text, while DBpedia and Dandelion did not recognize any of them. Also, DBpedia and Dandelion both provide a high number of false positives. In both cases, this was due to particularities of the text, which references a traditional Neapolitan song, Santa Lucia. The song is meant to celebrate the picturesque waterfront district, Borgo Santa Lucia, in the Gulf of Naples, so the Santa Lucia in the text is primarily related to a location, but both DBPedia and Dandelion often tagged Santa Lucia as a person, hence the high number of false positives. Even if the name of the place itself is derived from the name of a person, the correct classification of this entity is *location*.
Also, Dandelion and DBpedia both have a significant number of missed entities (that is, persons that were not tagged at all). This probably relates to the fact that a lot of words, proper names included, are misspelled in the analyzed text and not available in Wikipedia.
The full evaluation results are indicated in the Figure 19 and Table 3.

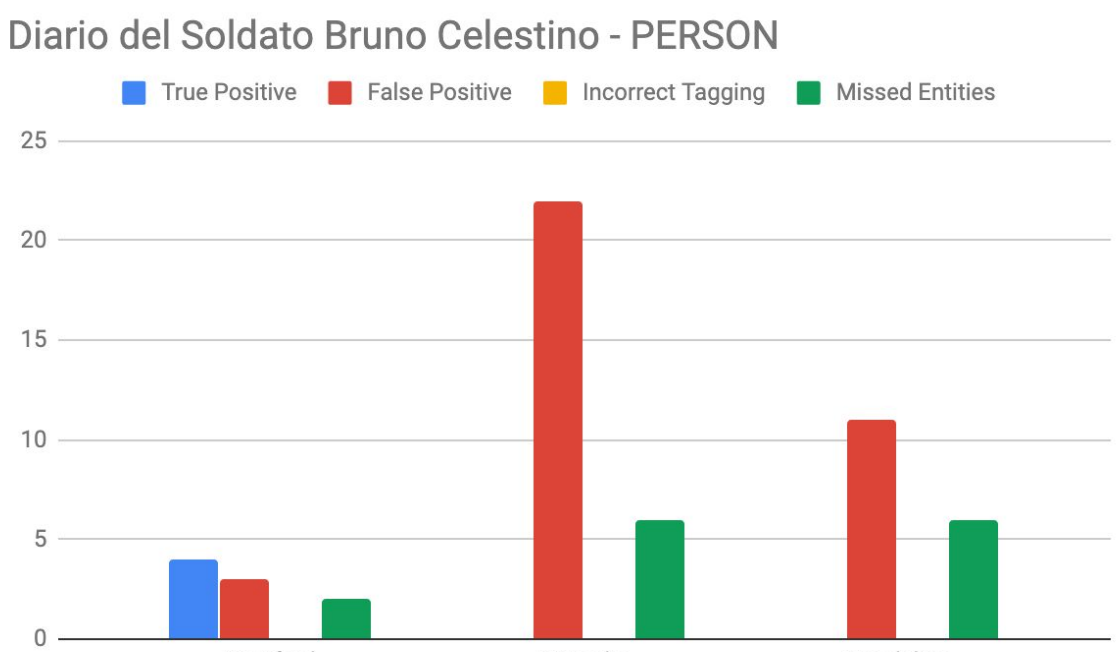


Figure 19: Person recognition and linking on Bruno Celestino journal

Table 3: Person recognition on Bruno Celestino Journal, results overview

	Stanford	DBpedia	Dandelion
True Positive	4	0	0
False Positive	3	22	11
Incorrect Tagging	0	0	0
Missed Entities	2	6	6
Total	6	6	6
PRECISION	0.5714285714	0	0
RECALL	1	0	0

Conclusions and further Development

This document presents the technical documentation for the first version of the semantic enrichment and quality control services. It presents the service architecture and design together with the preliminary experimental evaluation used to derive the specifications for the concrete service implementation and configuration. The quality of the enrichments are a major concern for this service. Consequently, the service is available to be used for documents written in the languages for which native models are available (i.e. Enrglish, German, Italian), or the quality of enrichment was assessed through the experimental evaluation (i.e. Romanian).

For the next version, we aim at supporting all languages used in the Transcribathon platform and to extend the evaluation for documents that are not based on latin script (e.g. Russian, Serbian, Greek). The standardized serialization of enrichments, that will include more information on the positioning of named entities in original text and/or machine translations, will be developed by using the model created by the EuropeanaTech Annotation task force.

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

The manual validation and improved usability within the Transcribathon are a second concern for the further development of the API. We will investigate appropriate approaches for computing a confidence score in order to provide to end users a ranked list of enrichments. This will improve the user experience for manual validation of enrichments, which is a prerequisite for integration with the Europeana Core Services Platform.

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/edit

Annex 1: Tools used for evaluation of named entity recognition, classification and linking.

Stanford NER (Version: 3.9.1)
Stanford NER was the first tool applied on the corpus, with two different standard English classifiers. One of these classifiers is based on a 4 class model (Location, Person, Organization and Misc) which were trained on the "CoNLL 2003 eng.train" dataset and the other one is based on a 3 class model (Location, Person and Organization) which also included the MUC 6 and MUC 7 training data set (for more information, see https://nlp.stanford.edu/software/CRF-NER.shtml). In addition to these two Stanford classifier models, the German model was also applied on Eduard Scheer's War diaries.

NLTK (Version: 3.3)
The Python-based named entity recognition tool NLTK (Natural Language Toolkit) was added as a comparison point for the Stanford NER classifiers. We used it with the default configuration and default classifier, which is also based on a 3 class model (Location, Person and Organization).

spaCy (Version: 2.0.16), Flair (Version: 0.4)
Flair and spaCy were applied on the corpus to serve as further reference to Stanford and NLTK. Both tools were used with the default settings and default classifiers.

DBpedia spotlight (Version: 0.7.1)
DBpedia spotlight provides an integrated solution for named entity recognition, classification and linking. Initially, the Candidates web interface of DBpedia spotlight was used to evaluate the tool performance on English text translations (i.e. using English version of DBpedia spotlight). Later on, the Docker based installation was used to retrieve named entities using English, German and Italian versions of DBpedia Spotlight.

TINT (Version: 0.2)
TINT NER is the first tool we used for performing named entity recognition for the document in Italian language. The NER functionality included in Tint is based on the CRF Classifier included in Stanford CoreNLP. The model provided with TINT is trained on the Italian Content Annotation Bank (I-CAB), containing around 180,000 words taken from the Italian newspaper L'Adige, and used for the Entity Recognition task at Evalita 2009, the Temporal Expression Recognition and Normalization Task at Evalita 2007 and the Named Entity Recognition Task at Evalita 2007.

Dandelion
Dandelion is a named entity extraction & linking API that performs very well even on short texts, on which many other similar services do not. It currently works on texts in English, French, German, Italian, Portuguese, Russian, Spanish and many other languages. With this API it is possible to automatically tag texts, extracting Wikipedia entities and enriching data. Dandelion APIs were also tested for comparison.

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