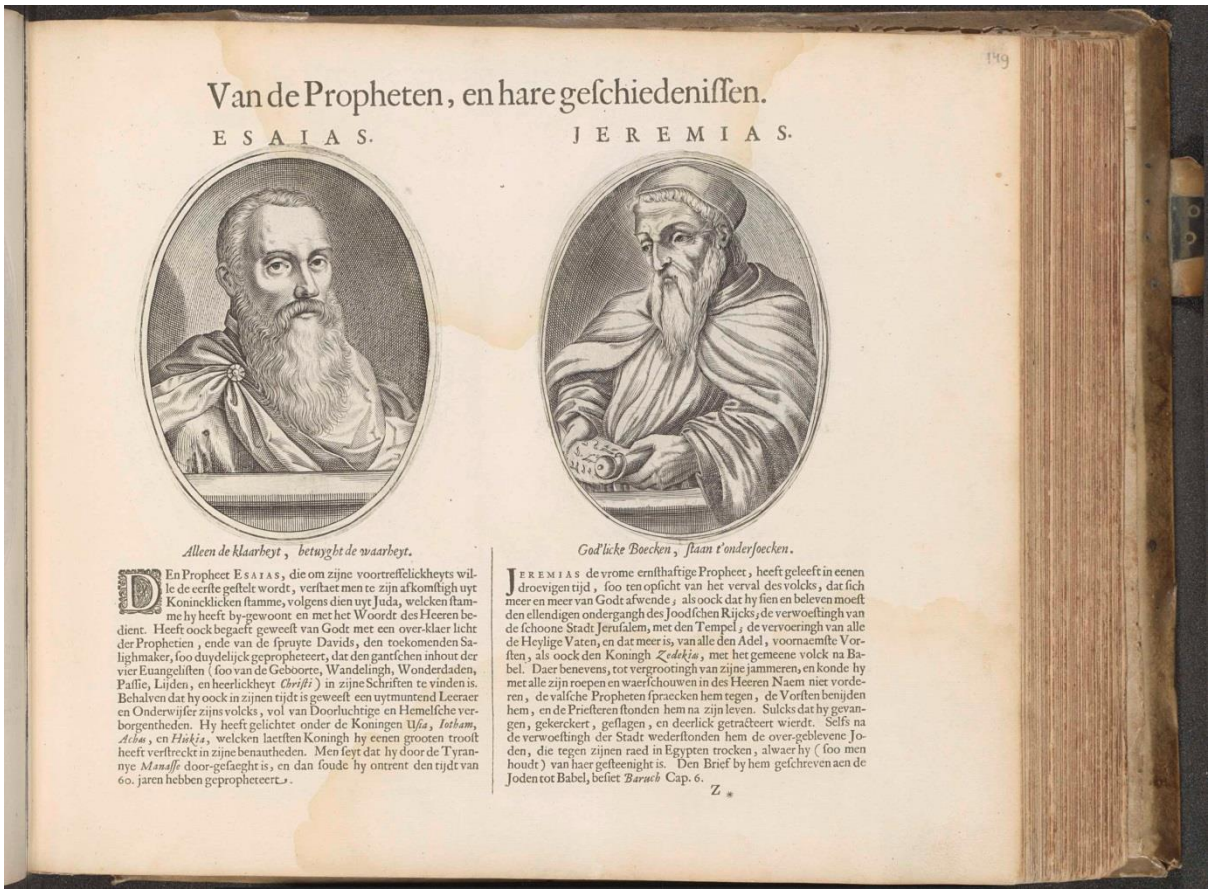
Europeana Research Grants
Final Report

Scholar Index: Towards a Serendipity Engine for the Humanities and Social Sciences.



Source: *Portretten van de profeten Jesaja en Jeremia; Van de propheten, en hare geschiedenissen; Den Grooten Emblemata Sacra, bestaande in meer dan vier hondert bybelsche figuren, zoo des Ouden als des Nieuwen,* Rijksmuseum, Public Domain.

Author: Matteo Romanello
Affiliation: École Polytechnique Fédérale de Lausanne, Digital Humanities Laboratory
Date: 20 August 2018

Results

Goal and Plan

The overall goal of the project was to interconnect the Venice Scholar (https://venicescholar.eu/) – a citation index of the literature on the history of Venice – with the digital objects contained in Europeana and related to the history of Venice.

The Venice Scholar was developed in the framework of the Linked Books project, aimed at indexing the literature on the history of Venice through citations to both primary and secondary sources. The platform currently contains more than 3000 digitized volumes (both journal issues and books), authored by circa 2800 authors over 200 years of historiography. The span of indexed materials is already substantial, with circa 4 million extracted references.

The first phase of the project (May 1 – June 19) was dedicated to 1) design of the interface components and 2) exploration of the Europeana APIs. The second phase of the project (June 20 – August 20) was occupied by 1) testing and debugging of the Venice Scholar API extension and 2) implementation of the new functionalities.

Development of new interface components

A new version of the Venice Scholar, with new interface components developed in this project, is accessible at http://www.venicescholar.eu/.

The two new components of the Venice Scholar are:

1) a sidebar displaying items from Europeana related to the page the user is currently viewing (e.g. an author page or a book page);
2) a visual gallery of Europeana contents related to the history of Venice, what we called *Venetica.*

Europeana Sidebar

The Europeana sidebar was added to all resources in the Venice Scholar, namely authors, books and journal articles. In any of these pages the user can now find a Europeana button, which displays the sidebar on the right when clicked upon. The result list can be scrolled and is paginated, meaning that more results (if available) are loaded on request (lazy loading), for the sake of performances. As an example, the contextual recommendations for author Gioseppe Paganelli[1].

Venetica

The purpose of *Venetica* is to display resources that are related to the contents accessible through the Venice Scholar interface, mostly scholarly publications related to the history of Venice[2]. The

[1] https://www.venicescholar.eu/results#europeana=59857acac12c608927dd9769&rT=authors&type=publications&refcat=&refid=

[2] The implementation of Venetica can be now found at https://www.venicescholar.eu/venetica.

resources are retrieved from Europeana based on frequently used keywords as well as names of popular authors, information that are available in the Venice Scholar database.

The user can select either authors or keywords from a pre-populated list of filters (a maximum of 10 filters is allowed). Each selection refreshes the gallery of results. When multiple filters are selected, clicking on a filter will:

a) highlight all corresponding search results
b) display them at the very top of the gallery/list
c) display a grey cross (top right corner of each filter) to remove the filter from selection.

A list-based view of the search results is also available, to allow for viewing the results in a tabular format, and for sorting according to various criteria (title, provider, license, date, language).

Methodology

Our usage of Europeana's Search API can be broken down into three cases:

1. search based on author's name;
2. search based on publication (book or article); keywords from title are used;
3. search based on keywords.

The biggest challenge we faced was to find contextual recommendations for those authors whose name does not have any direct match in Europeana (this is the case with most of modern or contemporary authors in the Venice Scholar database).

In these cases, we decided to use the titles of related publications as a form of "context" *about* the author. We take authored publications, cited publications as well as publications citing the work of a given author. From all these publication titles we extract keywords by filtering out stopwords and by using TF-IDF – term frequency–inverse document frequency – as a way of removing less informative words. We then use these keywords to make a series of queries to the Europeana API.

We did an evaluation of the contextual recommendations for authors, based on a random sample of 5,000 authors (approx. 10% of the total). We found that only less than 5% of the authors do not have any Europeana recommendations. About 42% of the authors yields at least one match when using the author name as the search key. The keyword-based strategy described above allows us to display some recommendations in the remaining cases (53%). As far as the number of results returned is concerned, 50% of the authors has 9 results or less, 20% have between 9 and 45 matches, 20% has between 45 and 730 results, while only 5% has more than 4,300 results.

Serendipitous findings

The original goal of integrating Venice Scholar with Europeana, was to enable the serendipitous discovery of contents in the latter by using some contents from the former as "seeds". While only a more thorough user evaluation will tell us if this goal was achieved – and to what extent – I would like to give at least one example of the serendipitous discoveries we made during this project.

The contextual recommendation for the journal article titled "Testamento del doge Agostino Barbarigo" and published in 1909 looks at first surprising: our search in Europeana's API returns two portraits of prophets Isaiah and Jeremiah, engraved in a 17c book which is part of Rijksmuseum's collection. But from the object's metadata we learn that two portraits, respectively of Agostino Barbarigo (the *doge*) and Amerigo Vespucci, were used in this engraving as images of the prophets.[3]

Use of the data

Both new functionalities – Europeana sidebar and Venetica – reuse the same set of metadata from Europeana: object title; thumbnail's URL; direct link to digital content; link to Europeana page; content provider; license; object type.

A new endpoint was added to the Venice Scholar API (http://api.venicescholar.eu/v1/europeana), which takes care of querying the Europeana API according to the search logic described above and of returning the search results to the front-end for display in the interface.

Challenges and issues

The main challenge we faced in this project is related to information retrieval and concerns the quality of search results we obtain from Europeana's API. In order to enable serendipitous discovery, a good balance needs to be struck between having too many – and too noisy – results and having few but very precise matches. In information retrieval terms, this means finding an appropriate balance between recall and precision.

Another issue that we will address in the coming months is how to make more data-driven the selection of search seeds displayed as filters in *Venetica*. For the time being, these seeds are the results of manual selection and curation (and thus static), but in the longer term we want to remove this pre-selection and replace it with a ranking or filtering based on other criteria (relevance in corpus, author's date of birth, or number of received citations).

Publications

- The Scholar Index (of which the Venice Scholar is the first working instance) was presented at the Digital Humanities conference, held in Mexico City.[4]
- The source code of the Venice Scholar was released under an MIT. [5]

Lessons learned

The extraction of keywords from the publication titles is a topic that could have benefitted from further research and experiments. The TF-IDF-based we employed is one the many possible strategies to distill keywords from text. Other more sophisticated approaches permit, for example, to have

[3] https://www.venicescholar.eu/results#europeana=595fa1e4fe7683316b2e01aa&rT=articles&type=references&refcat=&refid=

[4] The slides can be found at https://doi.org/10.5281/zenodo.1299239 (see especially slides 12-13). The conference paper is published at https://dh2018.adho.org/linked-books-towards-a-collaborative-citation-index-for-the-arts-and-humanities/.

[5] The user interface components are part of the Venice Scholar codebase (https://github.com/ScholarIndex/VeniceScholar), while the extension to the Venice Scholar API is part of the Linked Books code (https://github.com/ScholarIndex/LinkedBooks/tree/europeana/api).

keyword phrases (i.e. keywords consisting of multiple words, e.g. “editoria veneziana rinascimentale”). This was already anticipated, and it was not the main focus of the project, but it would be interesting in the future to explore and compare other ways of extracting words.

Co-financed by the European Union
Connecting Europe Facility