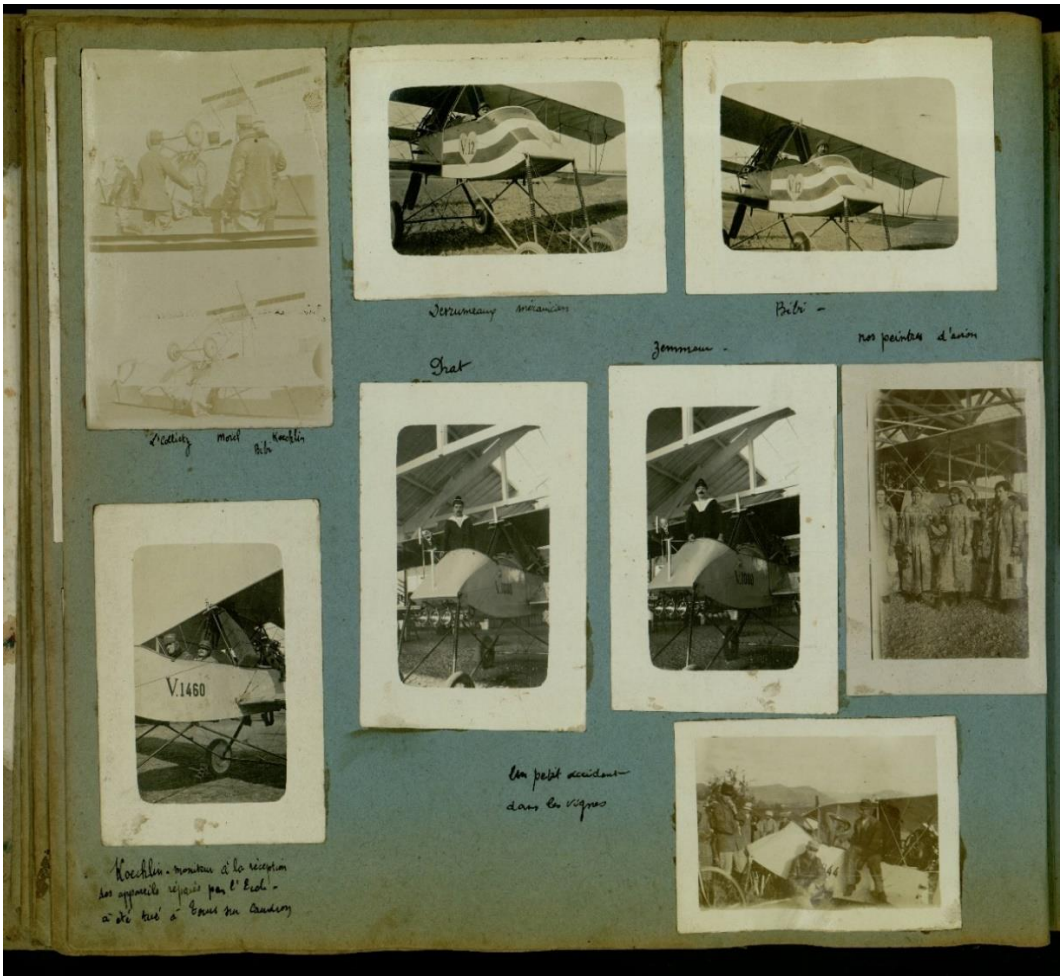
Data science contextualization for storytelling and creative reuse with Europeana 1914-1918.

Hagedoorn, Berber; Iakovleva, Ksenia; Tatsi, I

Europeana Research Grants
Final Report

Data science contextualization for storytelling and creative reuse with Europeana 1914-1918



Source: *Femmes peintres* (photographs of women responsible for painting canvas planes in World War I) by Louis Boissier. Photo by Archives départementales des Yvelines, Saint-Quentin-en-Yvelines (France). CC BY-SA. https://www.europeana.eu/portal/nl/collections/world-war-I

Author: dr. Berber Hagedoorn (principal investigator, b.hagedoorn@rug.nl), in collaboration with Ksenia Iakovleva (research assistant) and Iliana Tatsi (research assistant)
Affiliation: University of Groningen, the Netherlands
Date: 21 July 2019 (abridged version)

This report can be cited as: Hagedoorn, B., K. Iakovleva and I. Tatsi. (2019). *Data science contextualization for storytelling and creative reuse with Europeana 1914-1918.* Europeana Research Grants Final Report. University of Groningen, 21 July 2019, abridged version.

Table of contents

1 Data science contextualization for storytelling and creative reuse with Europeana 1914-1918

1.1 Set up of the project

In the research project 'Creative Reuse and Storytelling with Europeana 1914-1918', led by dr. Berber Hagedoorn (principal investigator, University of Groningen, the Netherlands), a combination of data science and qualitative analysis has been used to understand its platform engagement and map out requirements for creative reuse and storytelling with the Europeana 1914-1918 thematic collection, offering new contextualization of its textual and (audio)visual content. As a result, this study aims to provide insights into how Europeana 1914-1918 'affords' creative reuse and storytelling by researchers – both scholars and professionals/creators – as platform users, and how its linked (open) data can reveal 'hidden' archival stories, i.e. brought forth by cross-collection.

Our main starting point is that the selection of historical sources in a database adds another – more or less visible – layer of representation or interpretation (as Hagedoorn also discussed at the ENRS 2019 conference 'The Making and Re-Making of Europe: 1919-2019' in Paris in May 2019). Often, documentalists or users describing an item are more removed in terms of space and time from the personal story or perspective present in the historical source, which then leads to descriptions using more 'neutral' language – especially for (audio)visual content. Can data science offer opportunities to bring emotion 'back' into these sources? And can user analysis help here to better understand the value of such personal narratives in digital(ized) cultural heritage for creative reuse, storytelling and research?

In our contemporary media landscape, (audio)visual stories are no longer only told via mainstream broadcasting media, but are more and more told across different digital media platforms. The goal of this research project is to **use data science and qualitative analysis** to map how such storytelling is afforded by Europeana, and to develop models suitable for exploring creative reuse of its digital collections', taking the 1914-1918 Thematic Collection as a case study. Previous Media Studies research has studied how makers, together with users and algorithms, shape users' interaction with content on different platforms, in terms of political economy and platform 'politics' (Van Dijck, Poell and De Waal, 2018). We build on and move beyond such research by finding methods which offer new interpretations of the Europeana platform as a creative storytelling tool – and, hence, new interpretations of Europeana platform engagement – and how this engagement is shaped *in practice* by the interaction of the platform with different users.

'**Data science** is extracting knowledge/insight from data in all forms, through data inference and exploration' (RUG CIT)

'**Linked Open Data** is a way of publishing structured data that allows metadata to be connected and enriched and links made between related resources' (Europeana PRO/Api's)

Creative reuse can be understood as 'the process whereby one or multiple works, or parts thereof, are combined into a new work that is original, i.e. a non-obvious extension, interpretation or transformation of the source material' (Cheliotis 2007, p. 1). In the context of this project, the concept 'creative reuse' is found useful due to the focus it lends on, on the one hand, the creative and personal aspects of search and doing research (individual skills, search cultures, information bubbles…) as such self-reflexive elements should be emphasized more in doing contemporary research with digital tools (Hagedoorn and Sauer, 2019). And on the other hand, reuse as pointing to the fact that the selection of historical sources in a database adds another layer of representation or interpretation, as pointed out above. Cheliotis has in this context underscored how the practice of reuse is *widespread* in our society:

> *[Creative reuse] permeates many otherwise unrelated activities, from industrial manufacturing (building complex systems out of simple multi-purpose parts) to software design (code reuse), and from scientific publishing (reuse and citation of prior work) to fashion design (reuse of patterns, fabrics and designs).* (Cheliotis 2007, p. 1)

Creative reuse goes hand in hand with storytelling, which we understand in the broadest sense as narrativizing reality (a.o. White 1980) in online and digital contexts, and therefore a reliant on the contextualization of representations in a cultural heritage database to make data (re)usable. For instance, reuse by scholars and professionals as storytellers when carrying out different phases in their research and search processes (see also Hagedoorn and Sauer, 2019).

In order to understand how the thematic collections of Europeana 1914-1918 can be creatively reused for (digital) storytelling purposes, we study different ways that users can become engaged with the platform. To do so, we focus on the diverse stories that are present on the digital platform and the ways that they can be brought to the surface, at the same time offering new contextualization for these (audio)visual sources. This project helps in building expertise about the socio-technical practices of media users (principally, researchers) in relation to storytelling, search and research – especially for reuse in creative contexts – and in turn, generates knowledge, skills and tools for data science and qualitative analysis around (audio)visual data on media platforms, and the translation of interaction on a platform into data (the 'datafied experience').

Digital humanities methods have been incorporated for the analysis of historical resources and artefacts in (large-scale) projects, but scholars have gravitated more heavily into crafting archives and databases, instead of applying data science methods to existing ones (Manovich, 2016, pp. 2-3). Specifically, this project incorporates data science methods and qualitative analysis around linked (open) data on the media platform.

To do so, this project has a **mixed-method approach**. The principal investigator has developed, tested and improved **(1)** a model for platform analysis using data science, specifically topic modelling and sentiment analysis (using machine learning as well as manual annotation) and including (audio)visual sources **(= the focus of this progress report)**, and **(2)** a model for user studies using co-creative labs with different search tasks, talk-aloud protocols and post-task questionnaires, for user analysis, visual attention analysis (including an experiment with eye-tracking) and search task analysis, as well as questionnaires for survey analysis. By doing

so, a number of digital tools have been used and extended. The selected collection of stories in the Europeana 1914-1918 collection has been annotated, providing more contextual labels than the mere visual can provide. Statistics have been generated, and topic modelling and sentiment analysis carried out, along with the visualisation of examples of model clusters on the labels that the annotation created. This also included finding creative solutions for challenges regarding the study, especially the complex nature of (audio)visual sources and sources on the Europeana platform for applying data science methods.

This research has been developed in consultation and feedback sessions with both data science and digital humanities experts at the University of Groningen Centre for Information Technology (CIT), as well as with Europeana experts in user analysis and communication and the Europeana Research Coordinator. As a result, using protocols (methodological step-by-step plans) developed and designed specifically during this project – and which, importantly, can be reused in future research studies and for other Europeana collections, see also our recommendations for replication under each step in §1.2 – this project offers deeper understandings of Europeana as a creative storytelling platform, and models suitable for exploring and contextualizing Europeana's digital collections further.

This report before you focuses explicitly on the data science carried out during the project. Hagedoorn also employed a user-centred design methodology (Zabed Ahmed et al., 2006; Hagedoorn and Sauer, 2019) to analyse platform engagement of 100+ participants with the Europeana 1914-1918 collection, especially how users and technologies co-construct meaning. As previously argued:

> *Digital Humanities centres on humanities questions that are raised by and answered with digital tools. At the same time, the DH-field interrogates the value and limitations of digital methods in Humanities' disciplines. While it is important to understand how digital technologies can offer new venues for Humanities research, it is equally essential to understand and interpret the 'user side' and sociology of Digital Humanities* (Hagedoorn and Sauer, 2019, p. 3)

These user studies allowed for specific insights into how researchers – humanities scholars, creatives/media professionals as well as students – evaluate the role of creative reuse and storytelling when doing research into historical events and personal perspectives of World War I with the 1914-1918 collection. It is Hagedoorn's aim to publish the results of these co-creative design sessions in an academic journal publication.

1.2 Data Science models for exploring Europeana stories and creative reuse

This project delivers a **proof of concept** based on the following set-up. The data science analysis of the Europeana platform is split into several steps: selecting and collecting the data (scraping the site of the collection); translation of the descriptions from different languages into English (both automatic and manual); conducting sentiment analysis of the items' descriptions; topic modelling (both automatic and manual), and finally, annotation using both manual labelling as well as unsupervised machine learning for clustering data

(automated labelling) to offer new labels as contextualization for storytelling and creative reuse with/of the collection. Such steps also include some statistical text analyses and visualization of the results.

Using topic modelling and sentiment analysis, keywords and descriptions have been analysed, to answer questions about popular subjects and recurring themes. Specific attention is paid to what extent new contextualization and descriptions in terms of labels and sentiment detection can be offered by means of this approach (as a proof of concept), as well as in this manner offering new keywords and labels (which can function as sub collections, filters, or topics for searching the sources in the collection).

1.2.1 Selecting and scraping data

1.2.1.1 Methodology: data scraping

We created a dataset with information (metadata) about the items in the Europeana World War I 1914-1918 collection. The 1914-1918 thematic collection invites users to explore the untold stories and official histories of World War I in (currently) 374.715 items from across Europe (=198.641 texts; 172.635 images; 3.054 videos; 320 3D objects; and 65 sound recordings). These sources are aggregated from Europeana partner libraries, archives and museums[1], and at present 37.829 items in this total collection consist of so-called 'user generated content' contributed by either users online – as the website invites users to contribute their personal stories and content relating to World War I – or collected by Europeana during the 'roadshow' community collection days across Europe. The objects in the collection are digitized by professional documentalists.[2] All user generated content may be reused as open data (CC-BY-SA license). Content can also be accessed via Europeana's APIs.[3]

For collecting the data from Europeana 1914-1918, we used Selenium, the Python open source library for web scraping or data scraping (metadata in form of text). According to Rishab Jain and Kaluri (2015), this library has many advantages and supports multiple functionalities compared with licensed automation. It allows the designed scripts to communicate with the browser directly with the help of native methods.

For this study, several of the 1914-1918 sub collections or collection categories (called a 'topic' on the Europeana platform) are too small and specific, and/or the chances of not useful descriptions – which will not give relevant results in data analysis – are higher. Furthermore, in general the code for data scraping needed to be modified for every sub collection at least in part. The Europeana platform is quite unstructured, items occupy different positions, missing in some sub collections, and/or other issues. This is doable, but it does take the researcher more time to write code that is able to handle multiple possible versions of the pages.

[1] See the full overview of partner libraries, archives and museums on https://pro.europeana.eu/project/europeana1914-1918

[2] For further background on the Europeana 1914-1918 project in the Dutch context, see: https://www.slideshare.net/Europeana/het-europeana-19141918-project-in-nederland

[3] See https://pro.europeana.eu/resources/apis and https://pro.europeana.eu/what-we-do/creative-industries.

Therefore, a focus was placed on selected items in specific sub collections, for particular experiments within the overall project.

When developing the models for data science, a main focus is placed on the sub collections Women in World War I (sub collection containing 1.870 items in total) and Films (sub collection containing 2.726 items in total). In the first phase of developing models for topic modelling and sentiment analysis, the collections Official documents (123 items) and Aerial warfare (45 items) are also included and scraped. As outlined in the data science protocol (§1.2), the combined dataset is scraped from the Europeana page and translated, after which text-mining techniques will be implemented, such as topic modelling and sentiment analysis. The new stories (in terms of new labels and other forms of new contextualization) that might be discovered could be used to improve the filtering process and overall make for an improved user experience within the platform.

A portion of our analyses focuses more specifically on (audio)visual sources (such as Films, as well as Photographs, to uncover the added value of data science methods in offering new contextualization for storytelling with (audio)visual culture in a digital heritage database; since (audio)visual sources often offer more complex representations. As a case study, from §1.2.4 onwards, the differences in patterns and topics between user generated content and the linked (open) data from various institutions and collections currently present on the Women in World War I collection, will be analysed in terms of content, metadata, and intention. For the portion of the Photographs dataset (sub collection of 70,391 items) centred around the thematic axis of women (= 320 (audio)visual items), statistical analysis is carried out and topic modelling is performed on the labels and entities created using Vision API by Google Cloud. Data science methods will be incorporated to examine the differences and similarities with textual resources, drawing upon the transcribed documents and (audio)visual content of the WWI Diaries and Letters dataset (sub collections of respectively 846 and 482 items). Furthermore, since Europeana as a media platform supports the inclusion of user generated content, this research will also focus on identifying patterns between user generated content and linked (open) data from various institutions and collections.

Therefore, the following collections have been selected and scraped:

	Films	**Women in WWI**	**WWI Diaries and Letters**	**WWI Photographs**	**WWI Official Documents**	**Aerial warfare**
Type of dataset	(audio)visual sources	(audio)visual and text sources	Text sources	(audio)visual sources	Text sources	(audio)visual and text sources
Objects per dataset*	989	920	1400	320	123	45

** = after data scraping, cleaning and testing, final annotated number of items, with per item multiple new contextualization, such as sentiment calculation, labelling, etc.*

We scraped all selected metadata of the selected Europeana 1914-1918 sub collections using Python library Selenium. The scraped metadata was stored in a table in CSV format with a separate column for each type of content or information.

1.2.1.2 Results

☑ **Folder containing data science protocol, all datasets and scripts**
☑ **Our Python scripts for scraping**

The datasets of this research were extracted using the main Europeana 1914-1918 platform and the corresponding transcriptions website, for the WWI diaries and letters. In order to achieve this, **multiple scrapers have been written**, that adhered to the different kinds of data ((audio)visual/textual), as well as pre-processing techniques to clean and standardize the text data.

As a result, we retrieved a dataset with the following columns: **item number**; **title of item**; **description of item**; **type**; **provider** (=content provider); **institution**; **creator**; **first published in Europeana** (=date); **subject** (=list of different keywords); **language**; **providing country**; **item link**; **linked (open) data** YES or NO; and **collection** (=sub collection, e.g. films or Women in WWI).

1.2.1.3 Our recommendations for replication

It is possible to run our Python scripts for scraping, and subsequently retrieve the Europeana data as csv-files. It is also possible to directly download our files here. In order to run the scrapers, you have to install the following Python libraries on your PC: Selenium, Urllib, Pandas. For using the scrapers for retrieving data from other Europeana collections, you may need to modify them, to change in the scripts the names of HTML-tags where metadata is stored.

1.2.2 Translation: normalizing data into English

1.2.2.1 Methodology: automatic and manual translation

There are 24 languages in Europeana (Italian, Polish, Czech etcetera) of which the languages in our dataset were translated in two ways: using Google API+ Python library 'Google cloud'; and manual translation through Google Translate, for normalizing text into the English language. Since a large part of the data scraped was presented in different languages, we needed to translate it for our subsequent analyses. We used paid Google Cloud API for automatic translation, which is designed for translating large amounts of data. According to Li et al. (2014), Google machine translation lacks in the accuracy in grammar, complex syntactic, semantic and pragmatic structures. This results in nonsensical errors in grammar and meaning processing. Some languages

are translated more accurately than others, such as French into English (Shen, 2010) and Italian into English (Pecorao, 2012).

In order to translate the content of the platform in English, to be able to perform text-mining techniques, normalizing text into English is a necessary step. Even though Google machine translation might not be completely accurate in grammar, syntax, and structures, the overall meaning was deemed appropriate enough to carry on with machine learning techniques (Li et al., 2014). Importantly, whilst the grammar may not be perfect, the feeling remains, which is what is analysed in our next steps. The first part of normalizing texts into the English language is done automatically. Since many items contained descriptions in languages other than English, we decided to use Google Cloud API for automatic translation (importing it as a Python library), which lets websites and programs integrate with Google Cloud Translation programmatically. We implemented the following process into the Python script: if the language for the following item in the column 'language' was different from English, the description and header were translated into English by using the Google API ('google.cloud'). The translation was stored in the new column 'translated'.

However, in 600 cases (out of 2000 rows) it did not translate some descriptions due to more payment being requested by Python API Google Translate, which charges users for its use. As an output, instead of translated text, it gave the same untranslated description. Therefore, for normalizing the rest of the descriptions (which were not translated automatically) into the English language, manual translation in combination with Google Translate was used. We inserted the description into Google Translate in the original language, copied the translation in English (after a manual check) and stored it in the table.


ΤΟ ΚΑΘΗΚΟΝ
ΤΩΝ ΚΥΠΡΙΩΝ ΓΥΝΑΙΚΩΝ

Ἀφ' ἧς στιγμῆς ἐπὶ τὴν κυβέρνησιν τῆς ἐνιαίας Ἑλλάδος ἐτέθη ὁ ἀρχηγὸς τῶν Φιλελευθέρων, ὁ κεκηρυγμένος ὑπὲρ τῆς πολεμικῆς συμμετοχῆς καὶ τῆς Ἑλλάδος παρὰ τὸ πλευρὸν τῶν συμμάχων τῆς Συνεννοήσεως, δὲν ὑπῆρχεν ἀμφιβολία ὅτι ἐν πανστρατιᾷ θὰ ἐκαλεῖτο τὸ Πανελλήνιον νὰ διεκδι-

Η συμβολή των Κυπρίων γυναικών στον Α' Παγκόσμιο Πόλεμο.

The contribution of Cypriot women in the First World War.

Δημοσίευμα της εφημερίδας Ελευθερία ημερομηνίας 25 Αυγούστου 1917 με το οποίο ζητείται η σύσταση του Κυπριακού Ερυθρού Σταυρού που καλείται να στελεχωθεί με Κύπριες γυναίκες οι οποίες έχουν τη θέληση και ανθηρό το αίσθημα του καθήκοντός να βοηθήσουν στην ανακούφιση των πόνων των στρατιωτών παρέχοντας κάθε δυνατή περίθαλψη και στήριξη προκειμένου η Κύπρος να μην απουσιάζει από τις θυσίες της Ελλάδας για εθνική αποκατάσταση.

Image 1 *Example of Europeana 1914-1918 item and description 'The contribution of Cypriot women in the First World War'.*

1.2.2.2 Results

☑ **Our Python script for automatic translation**
☑ **Translated datasets**

The result was a new column with translations in our table. Part of the translations (600 out of 2000) were done manually, supported by Google Translate. Google translate technology is still not perfect, but our manual check has revealed that it almost always managed to save the real meaning of the text. This new contextualization can be found in the table.

1.2.2.3 Our recommendations for replication

Both automatic translation, sentiment analysis and noun extraction were done using one Python script, preprocessing.py, which can be found here. It demands installation of the following Python libraries: Pandas, NumPy, TextBlob, Goslate, OS, Google.cloud. For automatic translation to connect with Google API services, users have to set and use Google application credentials. You also have to create a billing account in order to pay for the translation. The library which we used for the connection with Google services, was deprecated and replaced with another one (the instructions for using it can be found here as well).

1.2.3 New labels as contextualization for storytelling and creative reuse with the collection

> *...usability is very much bound up with contextualisation. Users might be able to retrieve items, yet without context and a framework for interpretation, the cultural and material understanding of selected content remains limited.* (De Leeuw, 2011)

We used a number of different data science approaches to retrieve, gather and expand information for new labels as contextualization for storytelling and creative reuse with/of the collections. For instance, the labels we provide, can function as new filters and point to subtopics within a larger topic or collection. In the following pages, we offer the approaches we designed that other Europeana users/researchers can reuse, using our models (see the links to our datasets and scripts provided in this document). As part of the results, we also offer specific recommendations for when such replication should take place, and what researchers should take into account when they do so. Importantly, these approaches aid in:

- *defining new keywords, including topics which are impossible to find with an algorithm, by using a combination with manual approaches such as manual labelling (defining new keywords or topics manually and assigning them to items)*
- *improving the search algorithm in the collection (new keywords; new filters)*
- *creating meaningful links between items (new sub collections)*

- *this contextualization goes beyond present information in metadata such as descriptions*

- *Our files, scripts and datasets show the distribution of topics and sentiments among the items and collections and the variety (e. g. if there is a large difference between the lowest and the highest sentiment)*

1.2.3.1 Sentiment analysis

1.2.3.1.1 Methodology: sentiment calculation

After the translation, we conducted sentiment analysis of the translated data in order to get a sentiment value for every description. For this we used Python library TextBlob, which provides 'ready-to-use' tools for sentiment calculation or measuring sentiment (Gonçalves et al., 2013). It offers many useful functions for text analysis (part-of-speech tagging, noun phrase extraction, sentiment analysis, tokenization, words inflection and lemmatization, and spelling correction). The demand for the improvement of affective computing and sentiment analysis that extracts people's sentiments from online data, has been on the rise over the last decade (Cambria, 2016). Sentiment analysis, which is also known as opinion mining and emotions AI, uses natural language processing and text analysis to recognize, extract, assess, and examine affect and information that are deemed as subjective. Sentiment analysis has been more used for product reviews, market analysis, and marketing strategies, and analysis of trends on social media (Jussi et al., 2012). An essential function of sentiment analysis is the classification of the polarity of a body of text, as positive, negative or neutral, by looking at emotional and affective states.

Sentiment analysis was carried out using Python library TextBlob. It returns a polarity score, which is a float within the range [-1.0, 1.0], where -1 means that the text is 100% negative, and 1 means 100% positive. When calculating sentiment for a single word, TextBlob uses a sophisticated technique known as 'averaging'. It finds words and phrases it can assign polarity to (examples are 'great' or 'disaster'), and it averages them all together for longer text such as sentences. The algorithm for sentiment calculation was already implemented into the library, so we could not modify it in any way. It is based on a lexical-based method that makes use of a predefined list of words, where each word is associated with a specific sentiment. Lexical methods vary according to the context in which they were created.

Because sentiment in most cases was expected to be very low, for testing and evaluation, we ran it both without removing items with a 0 sentiment score, as well as with removing the items with a 0 sentiment score, as demonstrated in the visualizations of sentiments with and without 0's (for the goal of visualizing only items which have sentiment). In these graphs (see **Fig. 1** and **Fig. 2**), the horizontal axis of the plot represents all the items' descriptions of the scraped and translated dataset, the vertical sentiment score per item. Based on this evaluation, we decided to continue without removing items with a 0 score, as these items also demonstrated sentiment as shown in the graph on the next page.

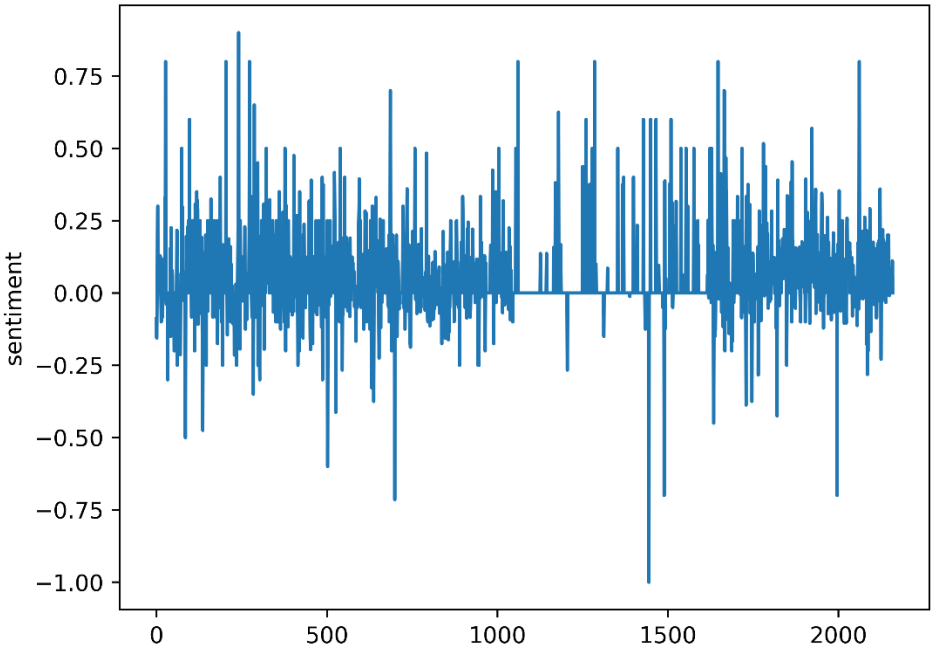


Fig. 1 *Sentiment with 0's*

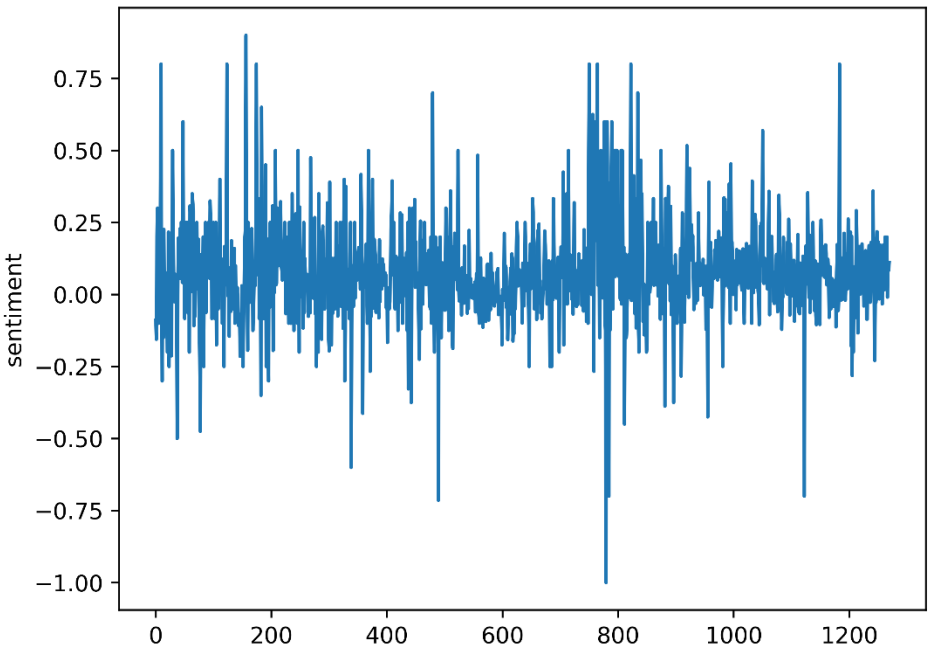


Fig. 2 *Sentiment without 0's*

1.2.3.1.2 Results

- ☑ **Our Python script for sentiment analysis**
- ☑ **Overview translated data with sentiment**
- ☑ **Sentiment calculation Women in World War I**
- ☑ **Sentiment calculation Films**
- ☑ **Sentiment calculation Official documents**
- ☑ **Sentiment calculation Aerial warfare**

Following the translation process, sentiment analysis was conducted for the dataset in order for each body of text to be assigned a sentiment value.. The Python library TextBlob allows for the processing of textual data. It provides an API for examining common natural language processing (NLP) functions, such as noun extraction, sentiment analysis, classification, etcetera (Loria, 2018, p. 1). TextBlob returns a polarity score within the range [-1.0, 1.0], respectively signifying negative and positive, by identifying words and sentences within a body of text and assigning subjective values to them. TextBlob is only one of a variety of such ready-to-use tools for sentiment calculation (Gonçalves et al., 2013). However, most of them include words from the texts with a high-score sentiment words (like tweets or reviews, where people describe their emotions vividly). Such software expects the same 'level' of sentiment in the input text. In our case, with descriptions of items connected with history, it was mostly detecting neutral or very low sentiment (since they do not contain informal words with high sentiment, which people use in tweets or in reviews). This new contextualization can be found in the table.

For example the item 'Hyänen der Welt' ('In the face of certain death') with the description: 'Drama in which two kidnapped persons, employees of a diamond cutting establishment, chase their kidnappers, a mine owner and his lover' offers a sentiment score of -0.6. This specific item itself may not be reused immediately as open data (as complex (audio)visual sources such as Films usually have copyright restrictions due to the many creatives involved), but contextualization in the form of a sentiment score can (1) support users in emotion detection for such items and in sub collections and (2) can provide researchers with an overview of sentiment present in certain collections or periods. Such an indication of sentiment present can support users when searching and selecting items for research. This is especially the case for creative reuse, when considering which items to contact content providers about to request a copy for reuse.

It must be noted that the scraped Europeana dataset offers challenges for sentiment analysis, usually because there are too many languages, and too little information in the text. The risk exists that we are just copying the data that already exist on the platform without much possibility to add value. Therefore, this approach as a *proof of concept* also works as a demonstration of the current extent of the possibilities of sentiment analysis (for researchers using domestic pc's) with the Europeana collection.

This analysis is followed up by annotation. Especially the manual annotation we carried out (this analysis follows in §1.2.3.3) gave us an opportunity to evaluate the results of sentiment analysis using calculation more precisely.

1.2.3.1.3 Our recommendations for replication

The sentiment analysis is running the script preprocessing.py. As an input, it takes csv-files with the data for four selected Europeana 1914-1918 sub collections (Women in WWI, Films, Aerial Warfare, Official Documents), then merges them in one table and gives a corresponding sentiment score to every item in it. The score appears in a new separate column in the table. Based on our tables, searching by sentiment score could possibly be implemented as a new search filter (we would recommend to do so in the form of a very easy to 'read' Likert scale), as participants during the user studies (uncovered in participant observation with talk aloud protocols) on their own initiative tried to search on positivity and negativity in the collection, generally to be able to research two different sides of a story (in this instance for the case of propaganda). They indicated the usefulness of being able to search on – as well as easy visualization of – positivity and negativity (source: focus groups March 14th, 2019 and May 22nd, 2019, at University of Groningen, the Netherlands), which a score could offer.

1.2.3.2 Topic modelling and noun extraction

1.2.3.2.1 Methodology: Automated topic modelling with LDA; noun extraction with TextBlob

Topic modelling is a machine learning and natural language processing method allowing for the discovery of stories in terms of more vague, abstract or 'hidden' topics within a collection. The keywords that are extracted from this process are clusters of comparable words. Analysed through a mathematical framework, the statistics of each word, can help deduce not only what each topic might be, as well as the overall topic balance in the whole collection (Papadimitriou et al., 1998; Blei, 2012). As a first step, we used the Python library TextBlob for noun extraction. This noun extraction in TextBlob uses the nouns which were extracted from the descriptions. Nouns extracted from every description were stored in a separate column in the table.

'display-case', 'photographs', 'right', 'son', 'brother', 'biplane', 'identity', 'tag', 'end', 'right', 'medal', 'family', 'disability', 'officer', 'whistle', 'handgun', 'pistol', 'protection', 'county', 'region', 'war', 'family', 'grandson', 'display-case', 'display', 'city'

Noun extraction using Python library TextBlob

Our next step was topic modelling – automated detection of a number of topics represented in our dataset. There are many ways of automated topic modelling in Python, most of them include machine learning and use Latent Dirichlet Allocation (LDA) (Řehůřek and Sojka, 2010; Jacobi et al., 2015), one of the most well-known algorithms for topic extraction. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics (Blei, 2003). For our research we mainly used the Python library for machine learning Scikit-learn. It has a module which conducts LDA and gives

a chosen number of topics represented by a chosen number of words as an output. In order to evaluate future results of topic modelling, we used some simple approaches for retrieving the most common words in the data.

We also used the Gensim library for Python which provides the LDA algorithm. Gensim is a Python library that can process raw texts in digital format and extract semantic topics in an automatic manner from them, without any human intervention. The algorithms in this library, one of which is Word2Vec, are unsupervised, meaning they need no human input in order to function; only text sources. The algorithms semantically detect the body of documents by analysing 'statistical co-occurrence patterns within a corpus of training documents' and once these patterns are located, any of the raw text documents can be 'queried for topical similarity against other documents' (Gensim). [4] In contrast to older text analytic methods where texts were treated as a whole, a much newer approach involves creating word representations. Those representations, called embeddings, are created using the algorithm Word2Vec, created in 2012 at Google by Mikolov et al. (2013). This involves creating high dimensional representations of words by utilizing their context (the window of words around the target word). This allows for the search of contextual similarities between words by training a Word2Vec model, using the datasets present in this research (for our case study see §1.2.4).

Before doing any analysis of the data it was necessary to remove stop words – words which are frequent in the texts but are not interesting for our research. Python library NLTK (Natural Language Toolkit) offers a list of such words (prepositions, modal verbs etcetera), but it was not sufficient for our study due to the complex nature of the dataset. For instance, many words in topics were representing nationalities and cities (British, Dutch, Spanish, Amsterdam, Moscow, etcetera). Therefore, we expanded the stop words list with more words. Moreover, for different collections we needed different stop words. For example, for 'Films' we had to exclude words such as 'film', 'video', and 'reportage'.

After cleaning data, the next step – creation or visualization of word clouds – was made by using the Python library WordCloud. We had to pre-process our descriptions and had to merge them into one large text, which was taken as an input by this library. As an output it provides a picture or visualization with many words of different sizes according to how often they are presented in our dataset. The second step was a simple extraction of the 10 most common words in the dataset and visualisation of them as a plot. Although a word cloud offers more words, this approach produces a more structured output.

1.2.3.2.2 Results

- ☑ **CSV-file initial topic modelling Women in World War I** – ***for topic modelling of this particular sub collection see §1.2.4.1 on discovering hidden stories and themes***
- ☑ **Our Python script for topic modelling**
- ☑ **Our Python script for making topics using noun extraction**
- ☑ **Noun extraction Women in World War I**

[4] For more information, please see: https://radimrehurek.com/gensim/intro.html.

☑ **Noun extraction Films**
☑ **Noun extraction Official documents**
☑ **Noun extraction Aerial warfare**

We conducted topic modelling with LDA. First, we defined the number of topics we want and the number of words which represent each topic. Then our program converts a collection of text documents to a matrix of token counts (to numbers), fit the data and gives the topics as a result. For such new contextualization, users/researchers can reuse our scripts in the topic modelling folder.

This analysis is followed up by annotation. Especially the manual annotation we carried out (this analysis follows in §1.2.3.3) gave us, as mentioned before for sentiment analysis, a good opportunity to evaluate the results of topic modelling more precisely. Some of the topics were not actually topics, but a number of frequent words not connected with each other. Sometimes part of the topic was correct, but there could also be some words present which did not fit the others. However, sometimes the words very accurately reflected the tendencies from the collections. For instance, in the 'Films' collection there are many films present about royals, which we see reflected in several topics.

Topic Number	**Words**
[0]	soldiers, general, line, seen, world, people, city, mark, corps, new
[1]	soldiers, army, emperor, troops, military, artillery, shot, shots, queen, world
[2]	troops, br. general, march, army, field, str, aircraft, soldiers, king
[3]	soldiers, column, troops, army, Limburg, horses, general, division, prince, king
[4]	story, love, army, Duyken, Pim, world, short, director, called, husband
[5]	army, troops, images, soldiers, emperor, shots, world, Wilhelm, shows, young
[6]	soldiers, troops, blood, gun, shown, military, gas, machine, field, small
[7]	soldiers, shows, army, committee, field, work, hospital, prisoners, camp, officers
[8]	world, gun, work, bridge, general, king, Mr, group, army, London
[9]	soldiers, hospital, band, London, military, general, lord, army, young, march

Fig. 3 *Topics Films sub collection*

Fig. 4 *Cloud: Films sub collection* (click here for the full visualization)

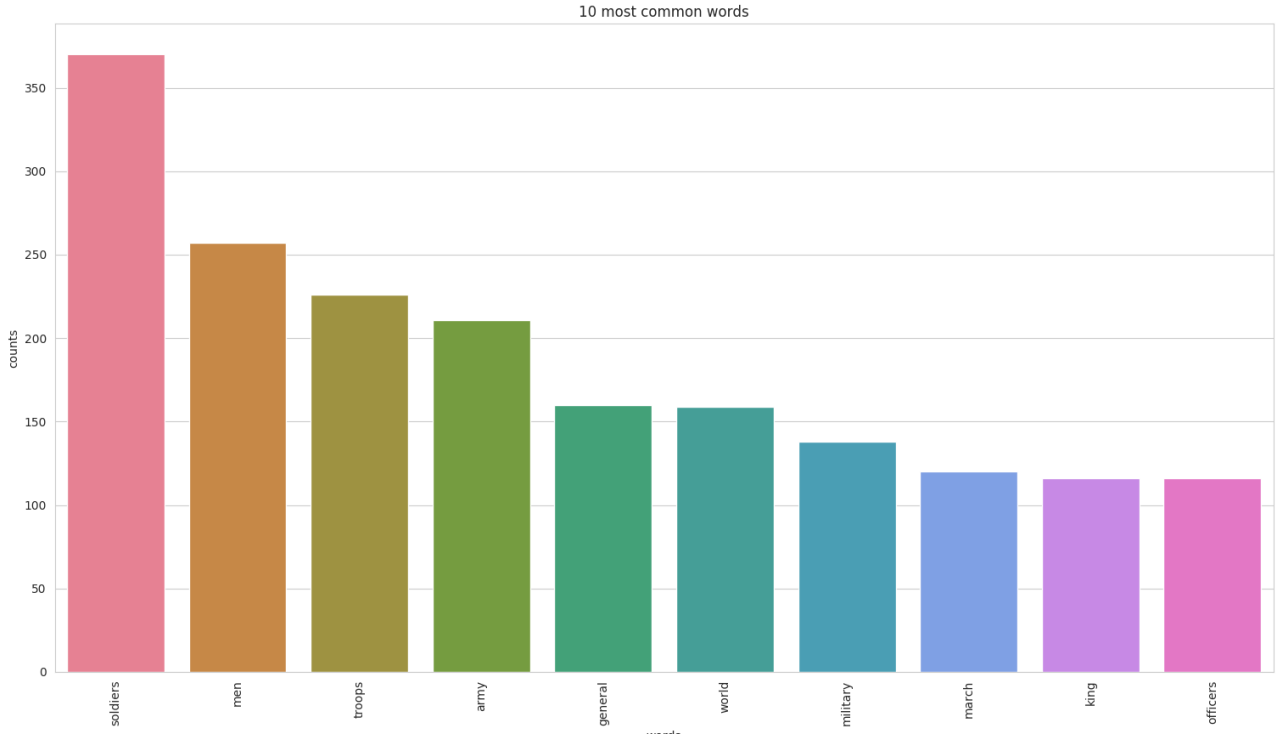


Fig. 5 *10 words: Films sub collection* (click here for the full visualization)

Fig. 6 *Cloud: Women in WWI* (click here for the full visualization)

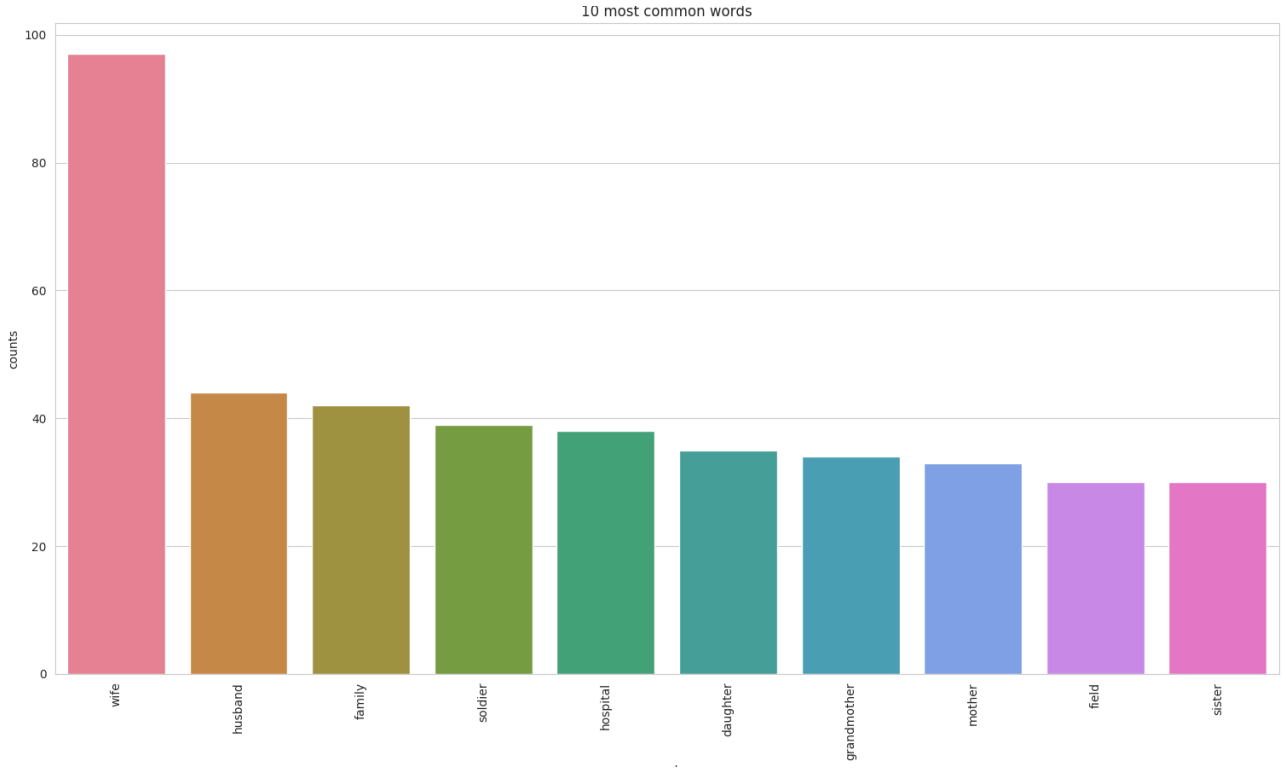


Fig. 7 *10 words: Women in WWI sub collection* (click here for the full visualization)

This topic modelling makes evident which are the **main 'hidden' topics or stories in these sub collections**, which is not evident from the filters on the Europeana portal. For instance, the topics or keywords we provide, can function as new filters and point to subtopics within a larger topic or collection.

1.2.3.2.3 Our recommendations for replication

Nouns are extracted automatically if you use the script preprocessing.py. They are stored in a new column in the table. For running topic modelling scripts it is necessary to install the following Python libraries: Pandas, OS, Re, WordCloud, Matplotlib, Scikit-learn, NumPy, Seaborn, Gensim. As an input it expects a csv-file with items' descriptions, as an output it gives a graph with 10 most common words, a word cloud and a list of topics which contain a number of keywords. Researchers can also change the number of topics and the number of keywords per topic (now both of them are set with 10).

If our scripts are applied to different Europeana collections, the list of stop words demands specific attention. For each collection it is necessary to create a separate list of stop words according to the topic. For instance, for the 'Films' collection we removed the words 'film' and 'movie', but for other collections they can be important and should not be removed.

1.2.3.3 Annotation using manual labelling

1.2.3.3.1 Methodology: labelling

To offer new keywords eliciting 'hidden meanings' in stories in the selected Europeana collections, and for improving user search on Europeana, we have manually annotated the two scraped sub collections, Women in World War I and Films. For this annotation we *tried* not to use the words which were already presented in the description, but either to use new synonyms, generalisation or possible associations to uncover hidden stories in linked (open) data. We also tried to assign to each item as many keywords as possible, so some of them have a long list of keywords, while others have only one or two.

1.2.3.3.2 Results

- ☑ **Annotation using manual labelling Women in World War I**
- ☑ **Annotation using manual labelling Films**

For the creation of such new meaningful keywords we carried out manual annotation (labelling). Our goal was to improve the search on the Europeana platform by defining topics which are impossible to find with algorithm (like 'domestic life') with manual approaches. Therefore, we tried to choose keywords which summarize the description or paraphrase the most important words in the description. For example, if the description mentions 'dragoons', we added the keyword 'soldier', which will help the users to get this result by searching for this word. The new contextualization can be found in these two tables: **Women in World War I** and **Films**. They can be used as new labels and keywords on Europeana in the future.

By using a combination with manual approaches such as manual labelling (defining new keywords/topics manually and assigning them to items) we importantly define and elicit topics **which are impossible to find with an algorithm.** An example is the topic 'domestic life', which is a key theme in the Women in World War I sub collection, but is currently *not* available for instance as a filter in search. Therefore, the labels we provide, can function as new filters and point to subtopics within a larger topic or collection.

Examples some of the labels' combinations using manual annotation:

- disabled_people, hope, life_after_war, domestic_life
- health_institutions, medical_research, blood_research, medical_equipment
- domestic_life, separated_family, betrayal, fate
- memories, friendship, united_nations, union
- family, memories, honour, nowadays, descendants
- memories, honour, nowadays, descendants, documents
- family, love, inspire, heroic, defense
- politics, ceremony, traditions
- soldiers, injured_people, victims_of_war, young_people
- war_consequences
- eyewitness, dignitaries, rich people, victory
- politics, ceremony, traditions, family, dignitaries, rich people
- before_the_war, domestic_life, traditions, travel
- family, couple, love, loyalty, fidelity, sacrifice
- marine, ships
- aerial, weapons, technology
- law_violations, cruelty
- before_the_war, politics, domestic_life, development, region
- assault, attack
- love_story, marine, ships, seamen, love
- nature, animals
- hatred, nationalism, nazi_ideology
- criminal, breaking_law
- business, workers, advertising
- excursion, sightseeing, documentary, tourism
- freedom, end_of_war, happiness, triumph, victory
- celebrities, biography
- death, suffer, injured_people, hostages
- affair, money, rich_people, poor_people
- injured_people, war_consequences, politics, food_supply
- destroyed_cities
- entertainment, culture, children
- hunger, food_supply, eyewitness, war_documents, freedom, victory
- industry, business, urban_life

1.2.3.3.3 Our recommendations for replication

We offer the following guidelines for manual annotation (labelling) for new contextualization:

- Do not repeat the words which are already in the description
- Use synonyms or generalization (e.g. for different items which mention kings, princesses, emperors etc. use a keyword 'royal people')
- Try to use as many synonyms as keywords as possible
- Try to use the same keywords for items with the same meaning so they can be filtered easily (e.g. not to use 'wounded people' for one item and 'injured people' for another item)
- Add hidden meanings (e.g. if the description states 'Anna had two children - Elisabeth and Jane', we can add keywords 'mother, daughter, family')
- Generalize actions (e.g. if the description states 'She cheated on him and married another man while he was in the army', we can add label 'betrayal')

1.2.3.4 Automated labelling: clustering with unsupervised machine learning

1.2.3.4.1 Reflection on supervised machine learning

Another goal of our project was to automate annotation or labelling of the items' descriptions with keywords (automatically generate keywords for the items). A general approach in case we have an annotated dataset would be to use supervised machine learning (Kotsiantis, 2007). For this we have to train the classification model on our annotated data and then to apply it to 'unknown' dataset. First, we counted unique combinations of keywords in the 'Films' sub collection and retrieved about 700 unique combinations out of 960 (we tried to be very specific in choosing keywords while annotating, so this was not a surprising outcome). Then we cut the number of keywords per item to 1 and got around 200 unique combinations. It was not valid for supervised machine learning, because too many keywords were 'outliers' - they were presented only in 1 or 2 items. Only 24 keywords were presented by more than 10 items. However, even when we cut the dataset only to the items which have these 24 most common keywords, we got a low accuracy of 35%. This can also be explained by the difference between 'human' and machine classification (Bhowmick, 2010): while automated models use exact similarities between the words from the texts with the same label, people just use their logic and associations which can give quite a different result.

1.2.3.4.2 Methodology: clustering (unsupervised machine learning)

After experimenting with the supervised machine learning, we decided that unsupervised machine learning would be the most appropriate method to use for this part of our study. Since supervised machine learning gave us a low accuracy, even with using small number of keywords, we decided to use another approach: clustering with unsupervised machine learning. For this we applied the same Python library as for topic modelling, Scikit-learn. The program, which we built in Python, uses the K-means clustering algorithm (Kanungo et al., 2002; Wagstaff, 2001). It splits all the data into the specified number of clusters, and in 10 (or another specified number) different circles it modifies the sizes of the clusters and fits the data to them in the

best possible way. After this process, we can extract the keywords which represent each cluster and assign them to the particular items. At the beginning, we needed to run some visualization in order to define the best number of clusters for our data (this is called 'Elbow plot'). The spots where the line 'breaks' and has a form similar to an elbow, are the best for visualization. Then we have to choose the number of clusters we want to have and the number of words in each cluster.

First, the program creates the word-document matrix (counts how many times each word is presented in each document). Second, it generates another one, with distances between different words (how close are two words to each other in each document). it defines *k* initial 'means', which are randomly generated within the data domain. Then it creates *k* clusters by associating every observation with the nearest mean. After that the centroid of each of the *k* clusters becomes the new mean. It creates new clusters around these means and repeats this process until convergence has been reached.

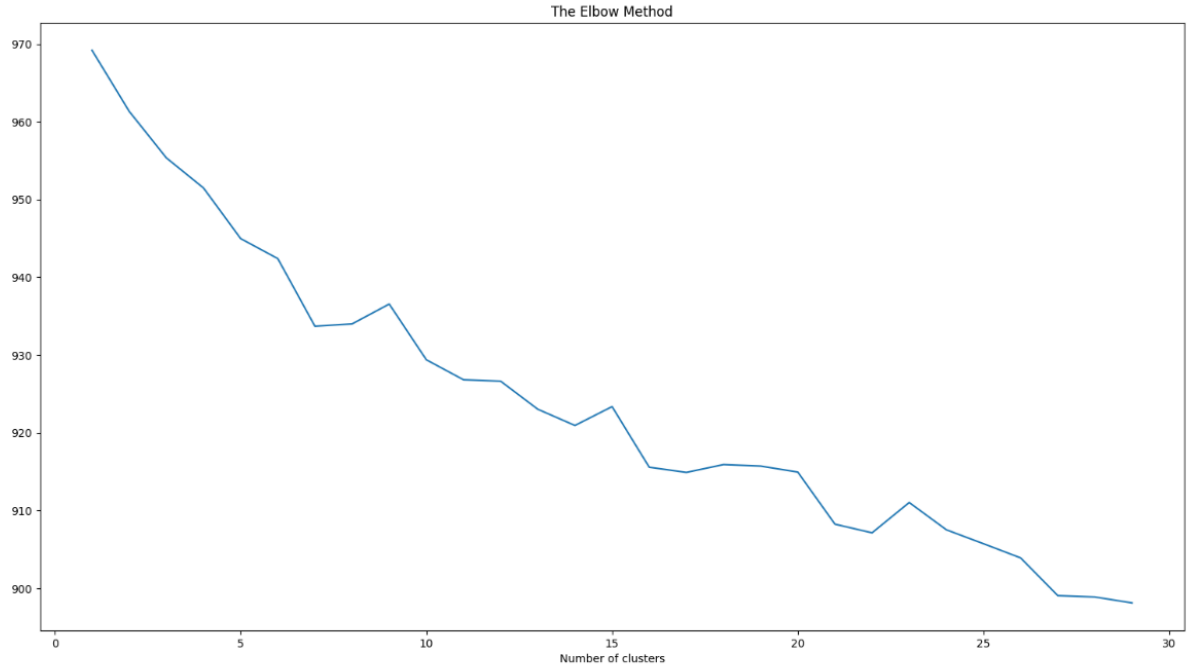


Fig. 8 *Elbow plot example*

1.2.3.4.3 Results

- ☑ **Our Python scripts for clustering using unsupervised machine learning**
- ☑ **CSV-file of Dataset labelled with 81 clusters**

As an output of this process we get the number of clusters we specified before. We can look through them and eliminate the ones which do not make sense. Then, the program assigns these clusters to corresponding items in the dataset and each item is given a number of keywords in the cluster. Our recommendation for use is to carry out unsupervised machine learning and clustering data with the different 'topics' or sub collections

within the larger 1914-1918 collection, because it will help to define subtopics within subtopics and make these more organized. We recommend to play with the number of clusters and to see which number gives the best result, and also to check the keywords in the resulting file, removing the keywords which do not make sense.

As a result, we get a new column in our dataset with keywords representing clusters. However, even after eliminating clusters which do not make sense, we often get incorrect results. If we choose 5-10 keywords per cluster, it is very likely that some of them will be correct while others will not (but choosing less may lead to inaccuracy too). For our project, **5 keywords per cluster** gave the **best performance**.

Two steps can be carried out for improving the result: (1) trying a different number of clusters/keywords and choosing the most accurate one, and a (2) manual check of the keywords assigned to the dataset and eliminating the ones which are not correct.

1.2.3.4.4 Our recommendations for replication

The Python scripts for clustering can be found here. First, for defining the clusters the script cluster_prepare.py should be run. It will show the plot with the line, which has some more or less recognizable breaks (or 'elbows'). You should remember the number on the x (horizontal) axis which corresponds to one of the 'elbows'. This should be a relevant number of clusters for your case. Then the script will ask you which number of clusters you prefer and you should enter this number. For running the scripts the following Python libraries have to be installed: NLTK, Re, Pandas, Sklearn (Scikit-learn), Numpy, Matplotlib.

At the beginning of the script you will find a list of stop words, which should be replaced by the corresponding one according to the collection analysed – we recommend to extend it after the first running of the script, after which irrelevant keywords will be clearer. After this you can run cluster_prepare.py again and see how removing stop words influences the result.

The script will save the clusters' numerical representations in the file Centroids.npy (researchers do not have to do anything with this file, it will be automatically used by another script). Then they should execute another script - cluster_run.py. It will read the clusters defined in the first script and ask which of them you would like to remove (some of them will not make 'sense'). After that, it will apply the rest of the clusters to your data (the file which you give as an input at the beginning) and save it as a csv-file (the example output file is here).

At the end, we recommend to evaluate the results of the clustering. If it is observed that many clusters do not correspond with the items they are assigned to, the researcher should try to run all the processes again with a different number of clusters and keywords per cluster. After the best possible combination of these numbers is found, in order to use these keywords for labelling data we still recommend a manual check and an elimination of irrelevant keywords.

1.2.4 Discovering hidden stories and themes in Europeana 1914-1918 using data science methodologies: case studies[5]

Drawing upon and expanding the protocols outlined in §1.2, the following part of the project pays further attention to the possibilities for discovering hidden stories and themes in Europeana 1914-1918 using data science methodologies, by means of specific case studies.

1.2.4.1 Implementation of data science methods to discover hidden WW1 stories

Europeana 1914-1918 constitutes a large collection of people's stories and memories, either in (audio)visual or textual format, that are presented to users through the mediation of the platform. Therefore, Europeana stands as a mediator of stories and memories, for users that might find it inspiring to educate and inform themselves about historical happenings and events of the past through words, pictures, and sometimes narratives of people that lived at the centre of them. Since it is quite common for people to update their knowledge every time they experience something relevant on the matter, almost as if updating a sense of prosthetic memory, the browsing of the Europeana pages could potentially lead to the formation of new cognitive topics to substitute old, pre-existing ones (Rose, 1992). Therefore, Europeana users might engage in a seemingly update-like process, where they often renew their comprehension of historical and cultural events of the past. Europeana, as a facilitator of stories and simultaneously a media repository that people use, can shape their prosthetic memory in a subconscious manner, functioning hence as an 'active memory tool', through technology (van Dijk, 2004, p. 262). Furthermore, Schwarz (2010) posits that the present is in position to shape people's understanding of the past to the same extent that the past can influence present behaviour. Consequently, it is safe to assume that different people and different cultures can establish different ways of remembering and experiencing the past and the present.

Memory work up until the late 1960's was led by and assigned to privileged males, being identified as ''the preserve of elite males, the designated carriers of progress' (Gillis, 1994, p. 403). Therefore, this research on the contrary focuses on the stories that have been overlooked and erased by the dominance of the male centric canon. Hence, the main sub collections that are used to exemplify the formation of the users' cultural and public memory in this part, are the Women in World War I collection and a part of the Photos collection also centred around women. In order to explore and justify the different patterns between (audio)visual and textual resources, a combined dataset of the World War I letters and the World War I diaries of the Europeana 1914-1918 initiative, is also analysed:

[5] For more see also Tatsi, I. (forthcoming Summer 2019). Reimagining Storytelling: The discovery of hidden stories and themes in the Europeana 1914-1918 collection, by making use of data science methodologies. (Unpublished master's thesis Digital Humanities). Supervisor: B. Hagedoorn. University of Groningen, the Netherlands.

	Women in World War I	World War I Diaries and World War I Letters	World War I Photographs
Type of dataset	Text and (audio)visual sources	Text sources	(audio)visual sources
Number of objects per dataset	921	1400	320

The Humanities field traditionally regards textual corpora using qualitative methods, whereas digital humanities perceive them through various quantitative analyses. Therefore, this research will take advantage of various digital humanities methods and digital tools, carried out under a reflexive and a heuristic approach, especially since the digital sources of the Europeana 1914-1918 collection, will be used as tools to investigate and renegotiate research hypotheses throughout history (Teissier, Quantin and Hervy, 2018). This notion aligns closely to the question of **the implementation of data science methods to discover stories of the historical era of WWI that have been overlooked**. Furthermore, by unearthing stories that might not have made it into the spotlight before, new information might arise; information that could challenge historical events and the perception of the past as it is comprehended today.

As described in the data science protocol, all the sub collections are scraped from the Europeana platform on the basis of titles, descriptions, type of digital object, provider, institution, creator, when it was first published, individual subject, language, providing country, link to the page, and whether or not the data is available to use, and then merged with the respective translations (see §1.2). For analysing user generated content and linked open data, the source code for the scraping is further modified, in order to parse another attribute from the Europeana page: whether each particular object of the collection was **submitted by an individual (user generated)** or if it **belongs to an institution/collection (linked open data by content providers).** All of the data is stored in individual files, in .csv format.

Each dataset will be following the initial scraping process and translation, as presented in the protocol (§1.2). About 20%-30% of the descriptions has not been translated, hence this translation is carried out manually. Therefore, a single file in .csv format is produced having the same attributes mentioned above, including the collection each individual object belongs to and the translation of its description.

1.2.4.2 Uncovering hidden stories in the Women in World War I dataset using topic modelling

The Women in World War I collection was scraped from the Europeana 1914-1918 platform using the data science protocol (§1.2), which resulted to a .csv file of 997 items. The Google Cloud API was used to automatically translate about 70%, the other 30% was manually translated into English, supported by Google

Translate. After cleaning the data and removing duplicates or items with no useful information, the .csv file consists of 921 items.

Each item in the Women in World War I collection, is accompanied by a description, which depending on the item varied in sizes. Therefore, the first step in the process would be to analyse the descriptions of the items. However, as seen below, a data problem arises, concluding that the deviation of the description sizes was too big (3-386 words), something that could create problems with using standard text-mining techniques, such as topic modelling and clustering. Instead, custom labels were produced (§1.2), after a lengthy manual annotating process of the collection, where context and the most concise information from each item were extracted by the annotator.

	Descriptions size	**Labels size**
Mean	104.38	9.95
Min	3.00	1.00
Max	386.00	41.00

Statistics of descriptions and labels size

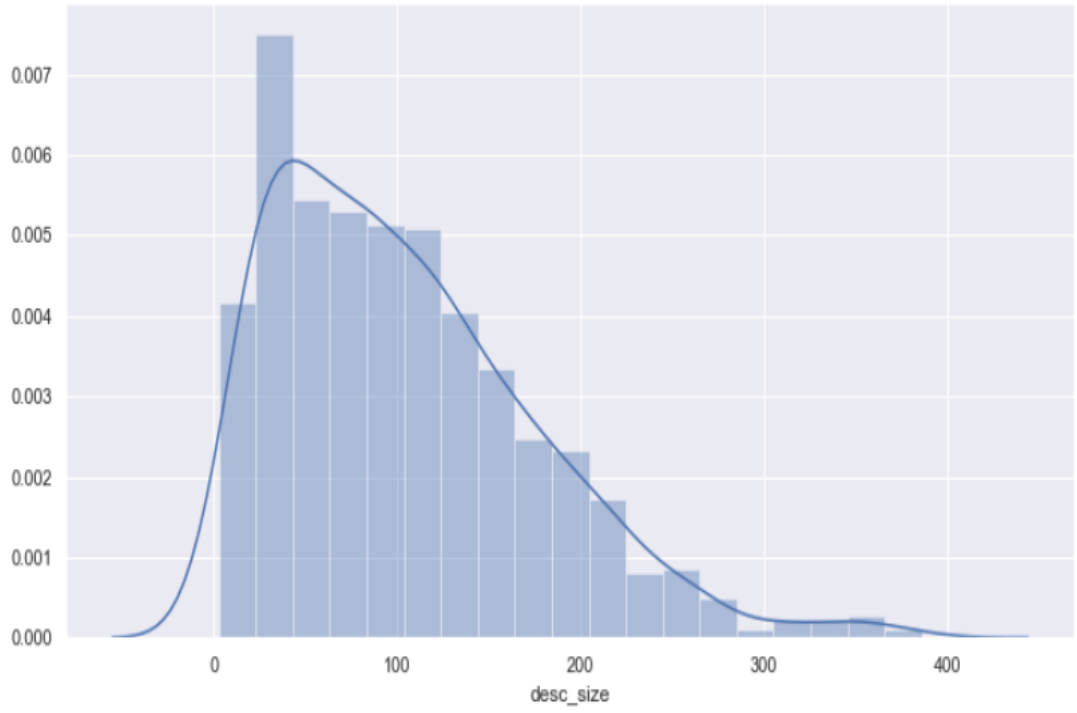


Fig. 9 *Word counts of description sizes (x axis: size of descriptions / y axis: frequency)*

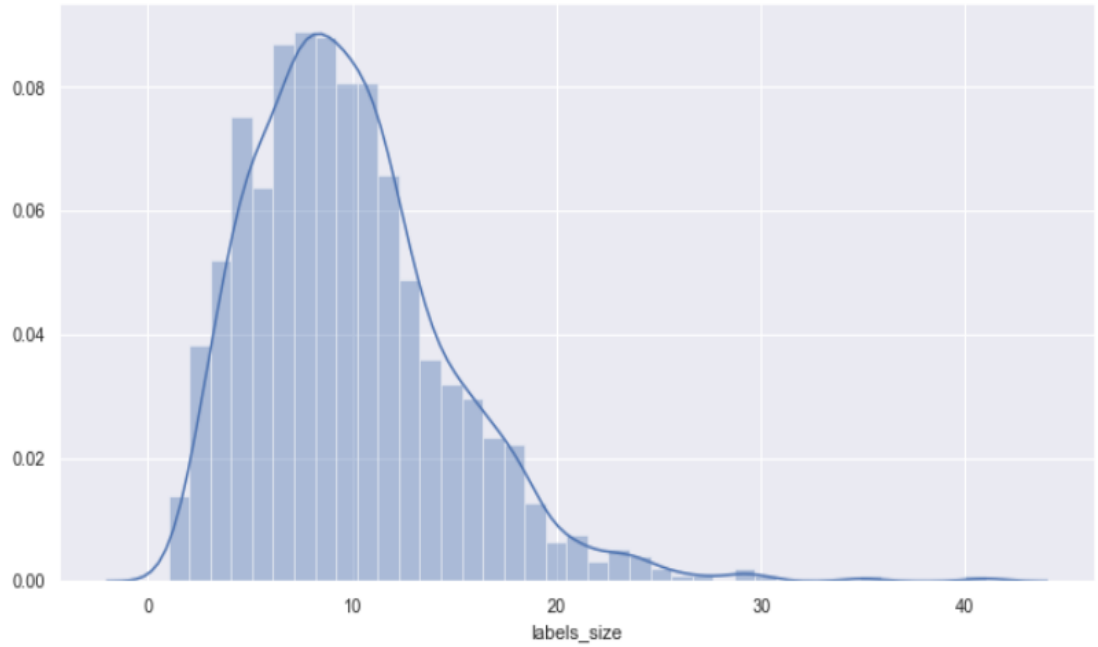


Fig. 10 *Word counts of label sizes (x axis: size of labels / y axis: frequency)*

After the annotating process, the produced labels allowed for the formation of a more representative and concrete dataset, which as seen in overview above 'Statistics of descriptions and labels size', has a range of 1-41 words, small enough to be manageable and concise, while simultaneously diverse enough to provide useful information. The representations of the word count of the datasets are also very telling (**Fig. 9** and **Fig. 10**). As seen in **Fig. 9** with words counts of description sizes, *descriptions* follow the Poisson distribution (Haight, 1967), whereas the *labels* (**Fig. 10** with words counts of label sizes), follow the normal distribution, which allows for the use of more standardized statistical methods using this particular dataset.

For the following step, a representation of the word frequency in the annotated labels is depicted, for the 30 most found labels (**Fig. 11**). As seen in **Fig. 11** *30 most frequent labels in the Women in WWI collection*, the most important words are as expected war, soldier, man. However, this is followed by the words woman and wife, therefore positioning the female presence well into a male-dominated historical period. The next words that follow revolve heavily around death, injury, and soldiers.

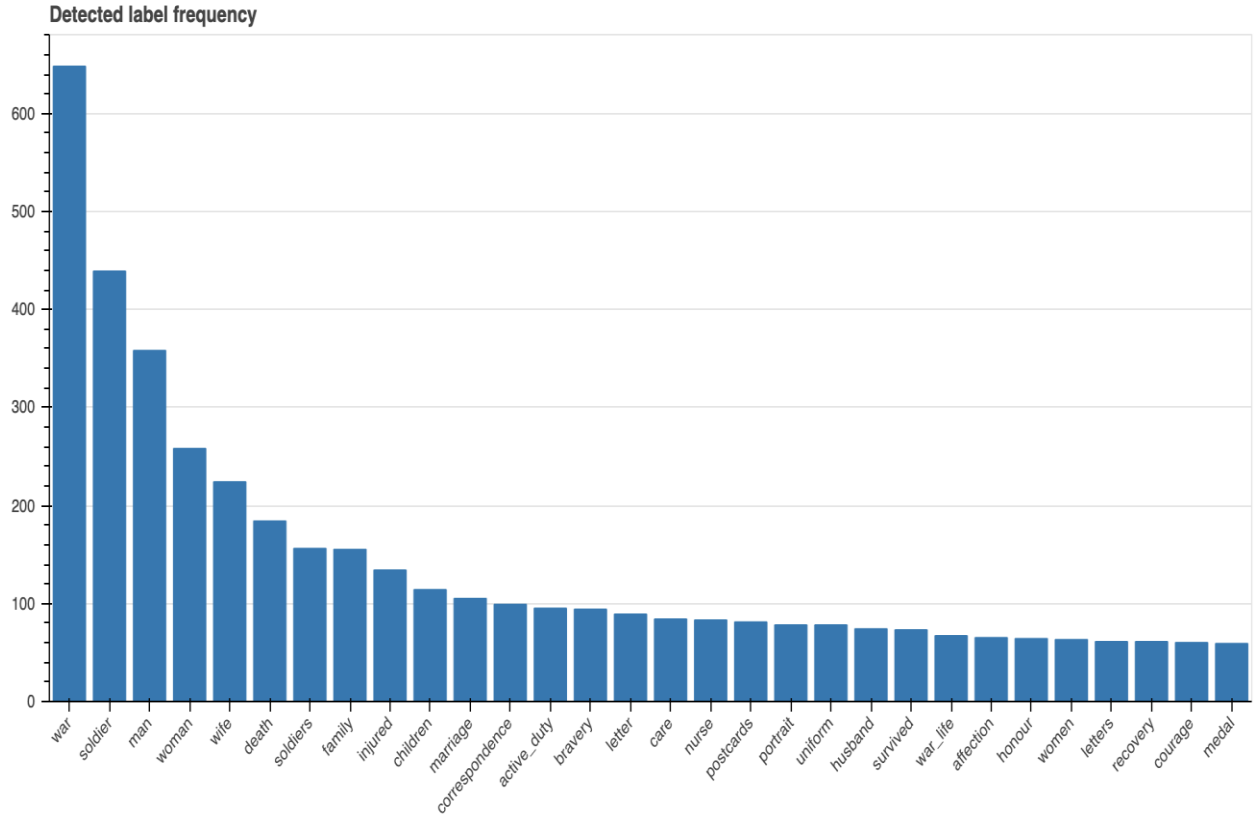


Fig. 11 *30 most frequent labels in the Women in WWI collection*

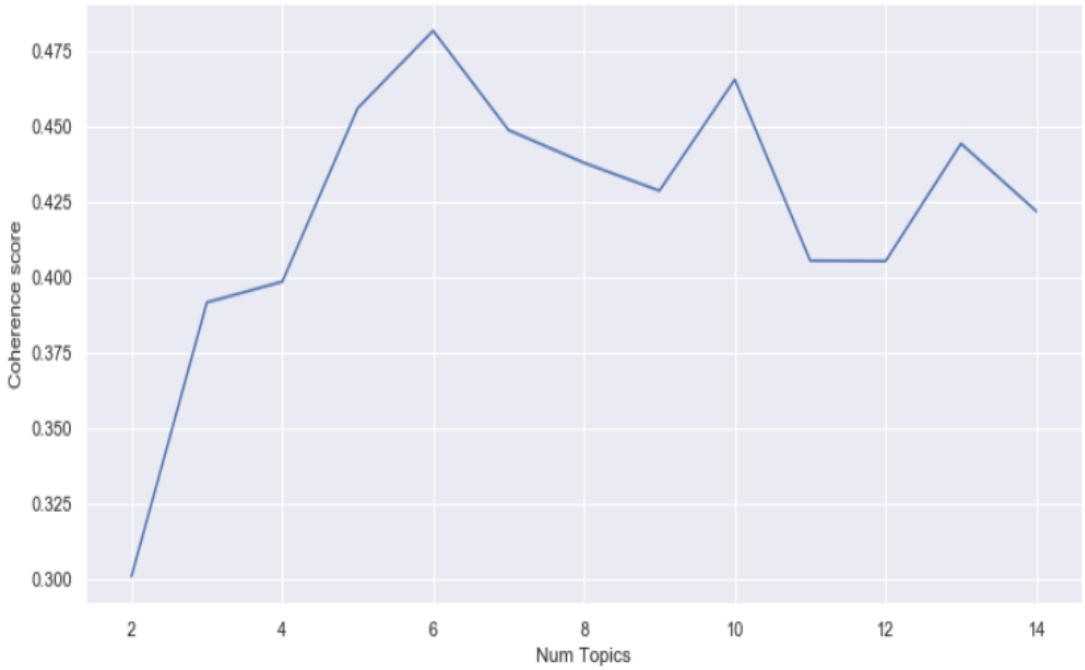


Fig. 12 *Coherence score from 2-14 topics*

Topic modelling is used, in order to extract possible contexts and topics of interest, by using the Gensim library for Python. This library provides the LDA algorithm, one of the most well-known for topic extraction. Topic modelling is a text-mining technique that enables the discovery of associated words in a text corpus, by identifying patterns. In our case, due to the absence of a large text corpus, we used the constructed datasets, i.e. the extracted labels, web entities, and the translated diaries and letters. By determining the words that most closely relate with each other, we can identify associated topics. In order for the number of topics to be produced, a coherence score was incorporated, in order to figure out the possibility of a good topic size. By experimenting from 2 to 14 topics, it seemed like the 6 topics might have had a higher coherence score, but the 8 topics made more sense to the annotator, so the number decided to remain at 8 topics (**Fig. 12**). The results of the topic modelling algorithm with 8 different topics, can be seen in **Fig. 13**, along with two examples of data visualization for two of the topics; model clusters extracted using LDA (**Fig. 14** and **Fig. 15**).

1.2.4.2.1 Results

- ☑ **CSV-files dataset case study Uncovering hidden stories in Women in World War I**
- ☑ **Scripts case study Uncovering hidden stories in Women in World War I**

The topics that were created by topic modelling are quite logical in terms of context. In particular, themes such as nurses taking care of injured soldiers, postcards and correspondence between families, as well as the bravery of soldiers are mentioned throughout the collection. That bravery often resulted in the award of medals and certificates, sometimes issued even posthumously to the widows. However, something that was not made clear by topic modelling, but was noted by the annotator of the dataset, is that for many of the widows that had their 'stories' present in the platform, it was very hard to be able to acquire pension from the government, sometimes even having to fight it legally (Topic [0]). Also, during the correspondence between soldiers and families, soldiers were not allowed to disclose either their locations or any military information whatsoever, since mailing services were heavily censored. Many times, soldiers had to cunningly hide information within their letters, either in coded writing or by writing under the stamp area (Topic [3]). Furthermore, one of the topics, mentioned is the involvement of women during the war, either in voluntary terms, such as nurses at military hospitals or the Red Cross or in the rarer instance of wealthy women, by giving money to charities and organising fundraisers. The word 'gender stereotypes' appears in this cluster, which the annotator used in items of the dataset, where the abilities of women to work hard or significantly contribute were either underestimated or ignored. For an abundance of items in the dataset, women that were left behind in the homeland, while male members of their family fought at the fronts, were usually in charge of keeping the household and the members of it afloat. However, what many of them received in return were letters and postcards ridden with anxiety, questioning their survival skills (Topic [7]). Moreover, in Topic [5], the correspondence between families and soldiers is also mentioned, with the exception that the correspondence in this topic includes words of affection, love, family, and were often accompanied by hand drawn pictures or handicrafts. This topic could allude to a more affectionate side of these soldiers, more prone to vulnerability and sensitivity. It is interesting to note that if these soldiers survived and returned home, they never again discussed the war with their families.

Consequently, it is rather obvious from the above remarks that machine learning techniques alone are not always enough to provide accurate results context-wise. It is very important in order to carry out a complete and concrete task in topic modelling, for the domain knowledge of the annotator to be involved. The results of the topic modelling algorithm with 8 different topics, can be seen in **Fig. 13,** followed with two examples of data visualization for two of the topics (**Fig. 14** and **Fig. 15**).

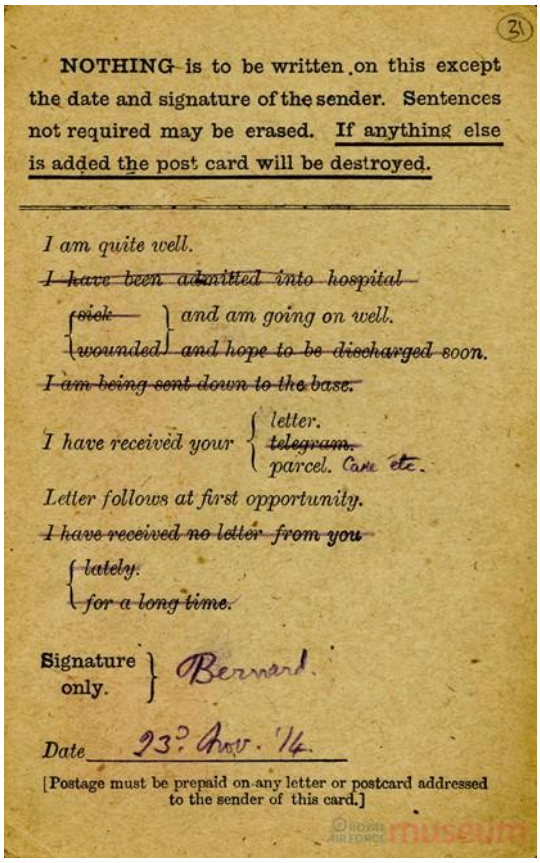


Image 2 *Left: 'I stand in gloomy midnight!' A field service postcard featured in the Women in WWI collection.*
Image 3 *Right: A censored field service postcard featured in the 1914-1918 collection.*

Topic Number	Words	Topics produced
[0]	courage, bravery, honour, medal, left_behind, certificate, woman, medals, widow, Irish	Soldiers fought with bravery and courage and either received medals upon their returns or their wives received their death certificates.

[1]	soldiers, active_duty, care, war, recovery, nurse, bravery, uniform, postcards, military_hospitals	Brave nurses worked at military hospitals and took care of injured soldiers until they recovered. Often, they received letters/postcards of gratitude.
[2]	nationalistic, patriotic, sadness, symbols, army, educated, training, women, young, possible_death	Many postcards contained patriotic and nationalistic symbols, which were often sent by young and educated people in the army or by women.
[3]	transfer, horse, family, hospitals, hard_work, censorship, hospital, help, brothers, doctor	Families worked hard to sustain themselves and send help to soldiers, who sometimes transferred or got injured.
[4]	war, soldier, man, injured, family, woman, children, correspondence, war_life, letters	Soldiers corresponded with their families, sending letters with their news about life at the front. Often, they got injured.
[5]	affection, woman, portrait, child, album, love, handicrafts, no_war_discussion, man, married	Many postcards featured family portraits of the soldiers or crafts on them, containing words of love and affection. Usually if more sensitive soldiers survived, they never mentioned the war again.
[6]	soldier, wife, death, marriage, man, war, letters, survived, worker, war	Many soldiers were workers before the war and they exchanged letters with their partners or got married upon their return, provided they survived.
[7]	sister, gender_stereotypes, postcards, elegant, photos, irish, red_cross, everyday_life, messages, fundraising	Rich women often helped the war cause by fundraising, whereas other volunteered at the Red Cross, contributing more than society thought possible.

Fig. 13 *The results of the LDA algorithm for 8 topics, in Women in World War I sub collection*

Fig. 14 and **Fig. 15** are two examples of the model clusters extracted using LDA. This graph creates a two-dimensional representation of discovered topics using the LDA algorithm (from Gensim) and the pyLDAvis Python library. The circles represent the extracted topics, and using dimension reduction algorithms it transforms them into a 2-dimensional plane. Their size represents the frequency of the times that a topic appears in the dataset corpus (the bigger the diameter, the more often a certain topic appears). In addition, the right side of the graph shows the most important associated words for the selected topic, as well as the number the times (this refers to the frequency of the times that a topic appears in the dataset corpus discussed above) appear in that topic compared to the whole corpus. For example: in **Fig. 15**, the topic [1] is selected, and the first, most important term, soldier, appears about 190 times, whereas it appears in the total corpus around 400 times.

Fig. 14 *Model Clusters Extracted Using LDA, Topic [0]*

Fig. 15 *Model Clusters Extracted Using LDA, Topic [1]*

For the purposes of examining if any differences exist between patterns in user-generated and open linked content, another attribute was created within the dataset, one that indicated whether each item was submitted by a person or was provided by a cultural institution. However, overall in the dataset there was a huge difference in numbers, with the user-generated content reaching 826 items and the linked open data reaching only 86. Therefore, before topic modelling is even performed at the two different data frames, it seems that it could be quite likely that results are not being as concrete as they would have been, if a more equal representation of content was noted. After an examination of the number of topics that made more sense to the annotator, the results of topic modelling for user-generated content (**Fig. 16**) and for linked open data (**Fig. 17**), are presented below. For instance, topic [3] that derives from the topic modelling algorithm for the user-generated content, consists of the words: war, soldier, man, death, wife, woman, correspondence, and refers to the correspondence between soldiers and their families back home or sometimes the notices of death they received.

Topic Number	**Words**
[0]	soldiers, war, medal, patriotic, soldier, postcards, wife
[1]	war, soldier, man, wife, letter, woman, family, death
[2]	soldier, death, woman, affection, wife, war, sadness, husband
[3]	war, soldier, man, death, wife, woman, correspondence
[4]	soldiers, care, nurse, woman, active_duty, recovery
[5]	war, soldier, man, wife, bravery, honour, death, marriage

Fig. 16 *The results of the LDA algorithm for 6 topics for user generated content in Women in World War I*

Topic Number	**Words**
[0]	man, woman, death, mass, commemoration, memorial, portrait, postcard
[1]	war, bravery, soldier, women, honour, active_duty

[2]	war, soldier, care, active_duty, nurses, letter, recovery, soldiers
[3]	woman, war, portrait, soldier, family, death, child, wife
[4]	war, family, letters, correspondence, man, wife, soldier, worker
[5]	war, woman, soldier, portrait, men, group, women

Fig. 17 *The results of the LDA algorithm for 6 topics for linked open data in Women in World War I*

It is quite visible from the presentation of the topics above, that they share more similarities than differences. Both user generated content and linked open data labels heavily revolve around the correspondence and the letters exchanged between families and soldiers on the front. Also, both labels feature the involvement of women as nurses -most of the time volunteers- that actively and with a lot of bravery, stepped up to the circumstances and risked their lives to take care of injured soldiers. A lot of notices of death are mentioned in the user generated content, which is quite logical since families received letters declaring their familiar ones injured and then succumbing to their injuries or being instantly killed on the battlefield. Most of these letters were accompanied by words of praise about the deceased individual's bravery or they were even awarded medals and certificates. Also, the word portrait prevails on the linked open data content, meaning a lot of the items, were either portraits of families, soldiers or at times even nurses and medical personnel. Concluding, a religious and ceremonious element is visible in the linked open data, with the words, commemoration, mass, and memorial featured in one of the topics.

1.2.4.3 Uncovering hidden stories in the Diaries and Letters in World War I dataset using sentiment analysis

This dataset comprises from diaries and letters acquired from the transcribed section of the Europeana 1914-1918 platform.[6] After the initial scraping of the platform, the dataset produced consisted of 1400 items, in many European languages, translated following the protocol (§1.2). The tools and libraries used to extract the data and analyse the results, are following.

For the next part of the dataset analysis, a set of language modelling and feature learning methods in natural language processing (NLP) was implemented, called word embedding. As elaborated previously, Mikolov's team at Google, invented Word2Vec, a word embedding tool, which can fast train vector space models (Mikolov, et al., 2013). During this process, words and sentences of a body of text are outlined to vectors,

[6] See: https://transcribathon.com/en/

meaning that Word2Vec takes the text and outputs work vectors. By creating a vocabulary from the text data that is used to train it, the tool consequently learns to represent words as vectors.[7]

1.2.4.3.1 Results

- ☑ **CSV-files dataset case study Uncovering hidden stories in WWI Diaries and Letters**
- ☑ **Scripts case study Uncovering hidden stories in WWI Diaries and Letters**

For the diaries and letters dataset from Europeana 1914-1918 platform, the reason for using word embeddings is to incorporate user input and find the closest word to one that is specified every time. The underlying assumption of Word2Vec is that two words sharing similar contexts also share a similar meaning and consequently a similar vector representation from the model. From this assumption, Word2Vec can be used to find out the relations between words in a dataset, compute the similarity between them or use the vector representation of those words as input for other applications, such as text classification. The results for the words nurse, war, German, and gas are presented in **Fig. 18** below, all words relevant to World War I. As is made visible, the word associations created by the implementation of word embeddings, are quite relevant and sensible context wise.

Words	**Associations**
nurse	charity, lily, nursing, voluntary, room, care, aid, red
war	conflict, period, campaign, warrior, he, corresponding, battle
German	captured, allies, protection, offensive, troops
gas	lines, fire, shells, subjected, enemy, attacks, night, violent, bombardment, attacked

Fig. 18 *Word embeddings associations in the World War I Diaries and Letters collections, where the Word2Vec algorithm produces a range of relevant words, based on one word from the user's input.*

As can be seen from **Fig. 19**, the five most frequent languages present in the Diaries and Letters in World War I collections are German, French, English, Dutch, and Greek, whereas the least frequent languages are Serbian, Spanish, Luxembourgish, Frisian, and Portuguese. This result is quite logical since the most dominant languages were quite involved in World War I and the least dominant originate from either country with small populations or countries that weren't actively involved in the Great War.

[7] See: https://code.google.com/archive/p/word2vec/.

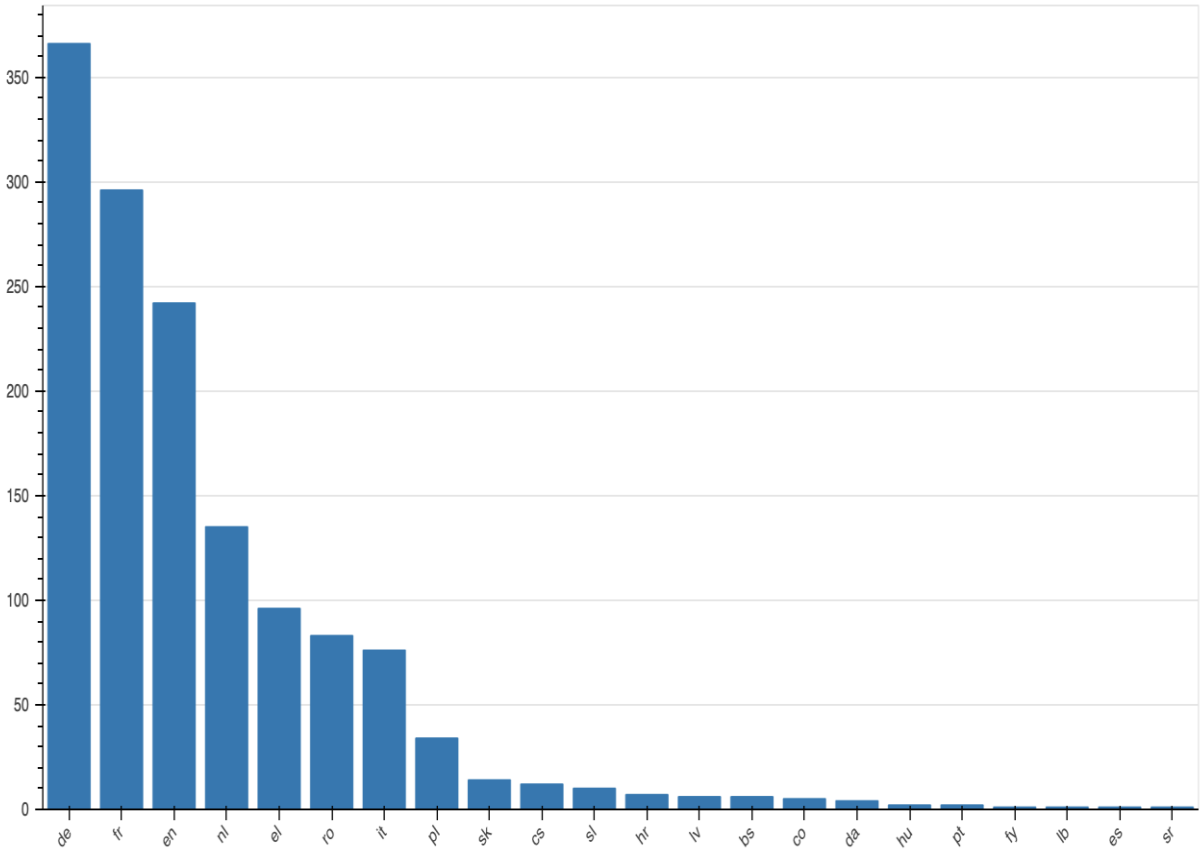


Fig. 19 *Distribution of origin in the World War I Diaries and Letters collections*

Language memo with Fig. 19:

de	German
fr	French
en	English
nl	Dutch
el	Greek
ro	Romanian
it	Italian
pl	Polish
sk	Slovak
cs	Czech
sl	Slovenian
hr	Croatian
lv	Latvian
bs	Belarusian
co	Colombian
da	Danish
hu	Hungarian

pt	Portuguese
fy	North Macedonian
lb	Luxembourgish
es	Spanish
sr	Serbian

An applicable methodology that was implemented on this dataset, was sentiment analysis. The Python library TextBlob was used, which provides pre-trained models that can quite accurately predict the sentiment of a sentence (an array of tokens), in a range of (-1, 1), -1 being the most negative limit, and 1 being the positive one. The float range used was [-0.5, +0.8], with -0.5 being 100% negative and 0.8 being 100% positive (**Fig. 20**). From the analysis, some interesting results emerge. First of all, there are no far-negative sentiments, as the most negative one is around -0.5, in stark contrast to the positive limit, which is 0.8. It is also quite clear that most of the items are found to have no sentiment, and if they do, it is more often than not positive. A visible cluster of positive sentiments near 0 (so around 0 - 0.5) could easily be expected in correspondence between soldiers and their families or diaries, where emotions such as hope, affection, love, longing, etcetera can occur.

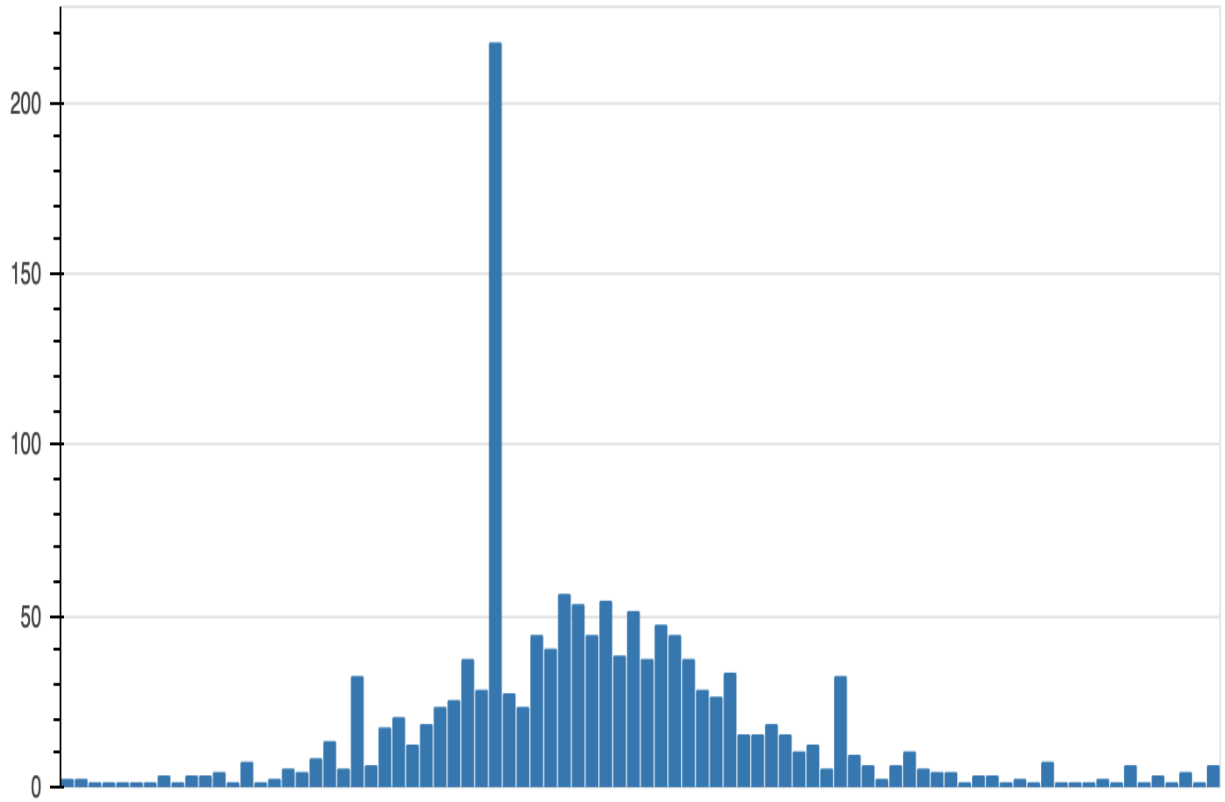


Fig. 20 *Distribution of sentiment in the World War I Diaries and Letters collections (x axis: individual collection items and their position on the sentiment scale / y axis: distribution, frequency of appearance).*

The majority of the items can be found in the middle, having been assigned a neutral value (**Fig. 20**). This is expected since the content of the dataset was tied to history and it probably did not contain high-score sentiment words, which can be found in social media content or product reviews. However, the texts that contain positive sentiments clearly exceed the negative ones, which could be expected in correspondence between soldiers and their families or diaries, where emotions such as hope, affection, love, longing, etcetera could be present.

NB: Note that in the figure representing the distribution of sentiments in the extracted documents (**Fig. 20**), the x axis represents the sentiment range, from negative -0.5 to positive +0.8. The bars themselves are just distributions that show a general picture of the sentiments inside the diaries and letters. The large bar in the middle is the count of the diaries and letters that have no sentiment, or better, the sentiment was not identified. It is clear from the graph, that most of the tests had neutral to positive sentiments, as this is obvious from the size of the bars on the right of the middle bar.

In reflection, sentiment analysis as a methodology is often used in political contexts, as a measure of public opinion, something which could be useful if the Europeana datasets could be explored and clearly annotated by domain experts. A political subcategory could provide a significant insight on the subject matter. However, the use being made here is subtler, as it is a useful sample of the mean emotional state of the authors, while at the same time the topics are explored (as explained in the topic modelling part of this report).

1.2.4.4 Comparison with Cloud Vision API when using data science methodologies with (audio)visual sources

This dataset (the Europeana 1914-18 sub collections: Women in WWI, Diaries, Letters, Photos) contains (audio)visual content, including images. Since the previous dataset contained only text documents, it is rather interesting to work with images and be able to analyse their content. In order to do that, the Europeana 1914-1918 platform was scraped using the Gensim Python Library, as to acquire all the photographs that were connected to the keyword 'women'. The resulting .csv file contained 339 items, out of which only the 320 were usable, these were later uploaded to a pre-trained model, namely Google Vision API. The latter allowed for the retrieval of various, diverse information on the images.

1.2.4.4.1 Results

- ☑ **CSV-files dataset case study comparison with Cloud Vision API when using data science methodologies for (audio)visual sources**
- ☑ **Scripts case study comparison with Cloud Vision API when using data science methodologies for (audio)visual sources**

The results that were retrieved from the upload of the images on the Google Vision API, were labels and web entities. The labels refer to the context of each image, with examples being vintage clothing, family, art, etc.

Web entities describe the context around the photograph as an item, i.e. what collections it was found in, such as Europeana or other possible websites that the picture might have been found on. By performing statistical analyses, two graphics were created, one for the label frequency (**Fig. 21**) and one for the entity frequency (**Fig. 22**). Furthermore, topic modelling was used to analyse the labels and the entities scraped from the dataset, using the Python library Gensim and the algorithm LDA. The results are presented below (**Fig. 23** and **Fig. 24**), along with two examples of visualization graphs: two respective model clusters extracted using LDA for topic modelling (**Fig. 25** and **Fig. 26**).

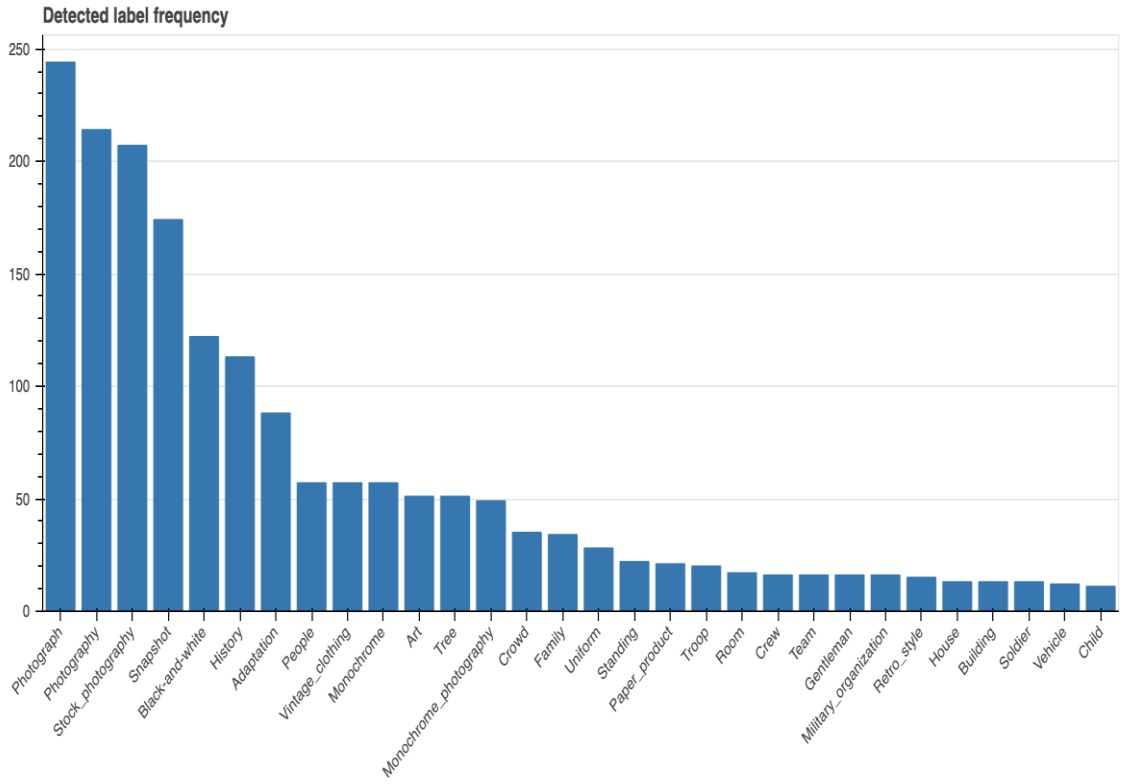


Fig. 21 *Label Frequency of Photographs in World War I dataset*

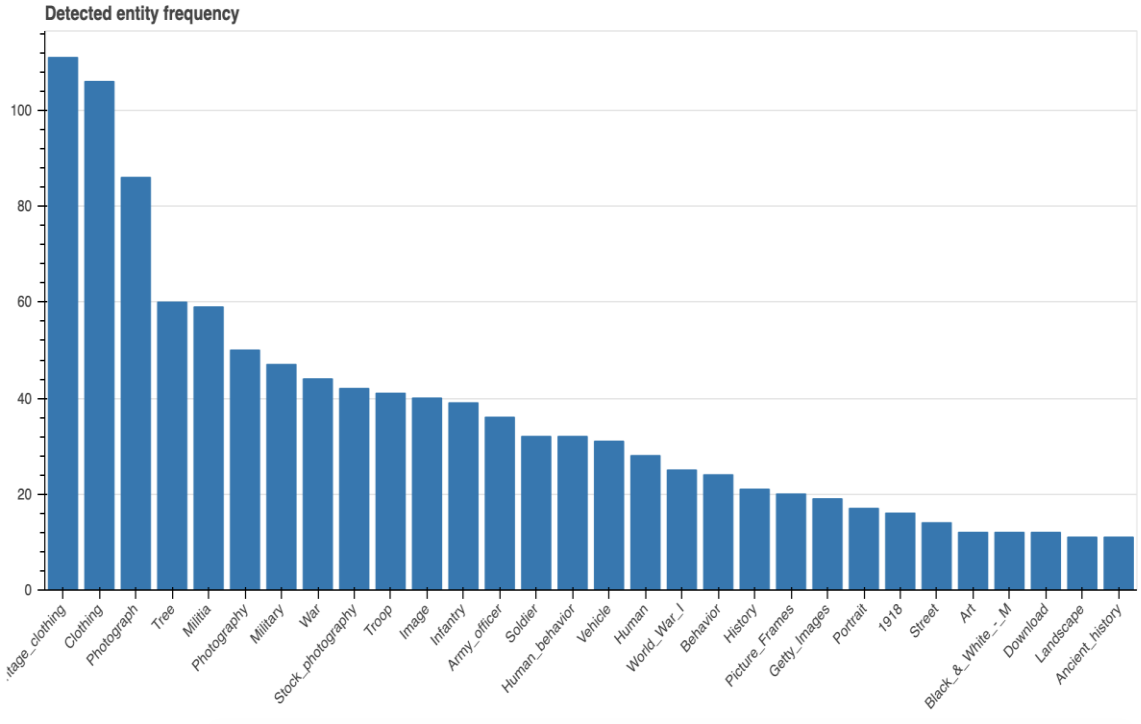


Fig. 22 *Entity Frequency of Photographs in World War I dataset*

Topic modelling was used to analyse the labels and the entities scraped from the dataset, using the Python library Gensim and the algorithm LDA. The results are presented below (**Fig. 23** and **Fig. 24**), along with two examples of visualization graphs, one for the labels (**Fig. 25**) and one for the entities (**Fig. 26**). An example of LDA for topic modelling follows.

```
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel, LdaModel

import pyLDAvis
import pyLDAvis.gensim


# CHANGE THOSE TO MAKE IT WORK
DATA = labels # entities / labels
TOPIC_NR = 6
WORD_NR = 8
```

```
data_token_sentences = [e.split() for e in DATA.to_list()]
dictionary = corpora.Dictionary(data_token_sentences)
corpus = [dictionary.doc2bow(e) for e in data_token_sentences]

ldamodel = LdaModel(corpus, num_topics=TOPIC_NR, id2word=dictionary, passes=15)
topics = ldamodel.print_topics(num_words=WORD_NR)
for topic in topics:
    print(topic)
```

Topic Number	**Labels**
[0]	Photography, Photograph, Snapshot, Black-and-white, Stock_photography, Monochrome, Monochrome_photography, Adaptation
[1]	Vintage_clothing, Photograph, Retro_style, Snapshot, Lady, History, House, Hairstyle
[2]	People, History, Crowd, Photograph, Troop, Photography, Stock_photography, Snapshot
[3]	Paper, Text, Paper_product, Letter, Font, Document, Handwriting, Bishop
[4]	Photograph, Uniform, Family, Vintage_clothing, Crew, Team, History, People
[5]	Stock_photography, Photograph, Photography, History, Adaptation, Snapshot, Tree, Paper_product

Fig. 23 *The results of the LDA algorithm for 6 topics for labels*

Topic Number	**Entities**
[0]	War, World_War_I, World_war, Soldier, History, World
[1]	Landscape, Photograph, Tree, Library, 1918, Winter
[2]	Water, Tellurium, Iodine, Vlorë, Silicon, Black_and_White, M

[3]	Vintage_clothing, Clothing, Troop, Tree, Vehicle, Militia
[4]	Stock_photography, Photography, Image, Photograph, Getty_Images, Download
[5]	Vintage_clothing, Clothing, Photograph, Human_behavior, Human, Behavior
[6]	Tree, Phenomenon, History, Photograph, War, Europeana_1914-1918
[7]	Military, Militia, Army_officer, Infantry, Soldier, War

Fig. 24 *The results of the LDA algorithm for 8 topics for entities*

Although the results in topics make sense, it is clear that the Cloud Vision API (being pre-trained in all kinds of images) can only do so much for the purposes of this research. The labels are quite generic and not as useful, providing little actual context for the themes presented in each image. It is therefore quite clear that a human annotator, especially one with domain knowledge on the subject, can provide labels that are more distinct and consequently more useful in creating different categories, topics, and themes for the Photographs in World War I dataset centred around women.

In addition, in order to perform sentiment analysis on the Photos Collection, the facial recognition service of the Cloud Vision API was implemented. However, while exploring the different photographs, it was noted that due to their bad quality, their blurry undertones, as well as the limited size of the dataset, the software could not pick out any distinctive emotions or recognize facial expressions. Therefore, the results of the sentiment analysis for this comparison part were deemed inconclusive and were not included in the results section.

Fig. 25 *Model Clusters Extracted Using LDA [labels]*

Fig. 26 *Model Clusters Extracted Using LDA [entities]*

1.3 Affordances of storytelling and creative reuse with Europeana 1914-1918: reflections on Europeana as 'an active memory tool'?[8]

A recurring practice in digital platforms with an educational and research character, over the past few years is to ask the public to participate and contribute to the work carried out; something that not only allows the curators of the platform to augment the already available heritage collection data, but to also engage with the general public in a deeper and a more meaningful level; platforms such as EUscreen or the digital version of the Library of Congress.[9] Projects that deal with mediated memories, identical or similar to the historical era on which the Europeana 1914-1918 revolves around, have also been launched.[10] Europeana incorporates a Reuse Team that aspires to augment the reuse of Europeana collections, targeting professionals functioning within the fields of research, education, as well as the creative industries. In particular, Europeana invests in collaborations with selected professionals deriving from the above-mentioned fields who act as mediators in order to improve outreach and engagement within the relevant audiences. Europeana's way to achieve this is through providing online spaces via Europeana Pro, reliable channels of communications, and most importantly access to relevant content, such as via APIs.[11]

The Europeana 1914-1918 project was developed in partnership with the Oxford University Computing Service (OUCS), and while collaborating with the Nationalbibliothek, the project toured eight different cities in Germany in 2011, people were asked to come forward with documents and memorabilia, to have them digitized on the spot.[12] Experts were also present on location, in order to identify authentic military artefacts and also talk to the public about their objects. People were also invited to upload digitized scans of their possessions online. Since most of the material uploaded online was previously unpublished, the newly-found digital collection, was a rich source of information that were available for research, not only on military fronts, but also on various European countries during World War I. The main aim of the Europeana 1914-1918

[8] For more see also Tatsi, I. (forthcoming Summer 2019). Reimagining Storytelling: The discovery of hidden stories and themes in the Europeana 1914-1918 collection, by making use of data science methodologies. (Unpublished master's thesis Digital Humanities). Supervisor: B. Hagedoorn. University of Groningen, the Netherlands.

[9] See: http://euscreen.eu/ and https://www.loc.gov/film-and-videos/collections/.

[10] For the project Living Legacies, see: https://nigradfair.org/sites/LivingLegacies1914-18/, this project provides communities with access to information, expertise and support for projects that explore the impacts that World War One had in Britain and Ireland, and the war's continuing legacies today. Furthermore, while striving to explore and analyse the casualties of the Great War and its significance for Canada, Antonie et al. (2016), created a methodology that allowed them to integrate more than one historical source and extract various results. Therefore, a new dataset with geographical data about Canadian soldiers who fought in World War I, was produced. This dataset presented social and history researchers with the opportunity to explore, analyse, and create new hypotheses about soldiers and their demographic information (p. 192). Finally, another project that implemented digital methods to historical datasets, revolved around the presence of Belgian refugees in Wales during World War I. It aimed to not only demonstrate the totality of archival records, but most importantly to allow new insights and conclusions to be formed about the local reactions to the largest refugee movement of the twentieth century, see also: http://www.walesforpeace.org/wfp/news-article.html?id=48

[11] See: https://pro.europeana.eu/page/re-use-team

[12] See: https://pro.europeana.eu/project/enrich-europeana.

initiative was for the individual, unseen, family stories to take their place next to the official historical sources, and significantly contribute to the formation and enhancement of cultural memory (Purday, 2012, p. 8).
A large part of cultural institutions, tasked with -amongst others- preserving and exhibiting the past, often reinforce dominant narratives and power structures with their practices, all the while maintaining and memorializing content that associates with power and control. In order to locate and unearth the stories and testimonies of people functioning outside these power structures, researchers have turned to alternative archives, such as 'visual images, music, ritual and performance, material and popular culture, oral history, and silence' (Hirsch and Smith, 2002, p. 12). By allowing users to upload content that they own, that might be significant and emotional to them, transforms users from mere consumers of content to **agents** who engage with history and actively contribute to collective, public memory (Owens, 2013, p. 128). Furthermore, in the case of Europeana, the engagement with the public can be achieved from a **multicultural perspective**, since the platform is a melting pot of languages and historical sources deriving from all over Europe. This exact engagement of 'end-users' and the multicultural element render the narrativization of heritage a rather dynamic and volatile process, that manages to establish 'multiplicity' in the quest for discerning the past (Rahaman and Tan, 2009, p. 110).

Any Europeana user is allowed to submit digital material for evaluation on the site, which if deemed reliable by an expert's team, will find its way to the platform amongst thousands of personal stories and objects, contributing thus to a digital tapestry of memories. Europeana hence, functions as a mediator between the general public and the stories that find their way in its digital domain, by freely allowing access to its site. Therefore, within these 'mediated' memories that appear on site, both media and memory hold key roles between the person and the society; with the users being able to intercept the past in their own way and adapt this interpretation to their perception of the present (Van Dijk, 2004, p. 262). However, according to Van Dijk (2004) individual memory as a supplementary factor to the historical narrative can be easily dismissed, with personal memory being perceived as a cultural phenomenon that harms the notion of remembering. It could be suggested that personal memories, narratives, and testimonies coming from individuals that do not function within academic scholarship, could harm historical integrity and authenticity. In order to comprehend what should and should not find its way into the official historical narrative -becoming thus a part of public memory- any claim to historical truth should be considered. User generated content such as personal stories, even though they allegedly form perceptions of the actual past, could easily adopt myth-like qualities. For example, people who belong to a larger collective, sometimes they share stories amongst them of the people they believe to be (Poole, 2006, pp. 157-158). However, this does not mean that their common stories might not accurately represent the historical past, therefore subjecting them to the same amounts of criticism as history.

By elaborating on people's stories on the Europeana 1914-1918 collection, a sense of biographical memory is created. According to Kuhn (1995), the forging of individuality occurs through autobiographical story forms that go on to shape aspects of personal cultural memory. The sub collections of Europeana 1914-1918, offer diverse acts of memory to the users, by allowing them to navigate through pictures, albums, documents, and at occasion videos, an action of remembering what Kuhn (2000) has coined as 'memory work, ... [an] active practice of remembering which takes an inquiring attitude towards the past and the activity of its

(re)construction through memory' (p. 286). In particular Kuhn (2000) goes on to suggest photographs to be 'far from being transparent renderings of a pre-existing reality, embody coded references to, and even help construct, realities' (p. 183). On the other hand, Rose (1992) posits that memories are created anew every time people remember, and media tools and technologies are so overwhelmingly full of these exact memories, that they essentially become inextricable. Hence, the memory process contains also the making of the above memory products, a creative process that allows for the establishment of a perpetuity between the past and the present; time and memory constantly influencing each other (p. 264). Being able to often reconstruct memories, allows each user's individual memory on the Europeana platform to be formatted, while simultaneously influencing the collective memory and identity or at least steering it to specific directions.

History in the 21st century is not only more accessible, but also formulated by digital resources. Therefore, the challenge expands beyond the primary steps of digitizing archives and resources. Specifically, the challenge for researchers translates to being able to implement already existing content for research purposes. In academia, this exact furthering of any research is in position to not only allow new approaches to historical hypotheses, but also function as the basis for interdisciplinary research, juxtaposing content and information from many possible fields, thus enabling researchers to ask new, innovative, and -previously- unattainable questions. As mentioned above, Europeana 1914-1918 offers the possibility to its users to upload digitized copies of objects they own that are relevant to the Great War. This open-source feature of the platform allows for a wide range of possibilities to be engendered within the digital realm of remembering. Most importantly, the concept of juxtaposing already established archives with recently emerged digital collections in a hybrid form, could function as the point of reference for a new 'scholarly research infrastructure', not just by merging together textual resources or textual with (audio)visual ones, but rather introducing new elements and 'secondary digital outputs', such as data visualizations (Hughes, 2016, p. 226). Socially, the above juxtaposition could form the basis not only for discourses on the power and sociocultural credibility of archives, but also open up a much wider discussion on digital collection and their influence in the formation of cultural memory.

For the case of the diaries, letters, and photographs in Europeana's 1914-1918 sub collections presage people's decision when it comes to how and which memories are deemed preservable. Especially photo albums and scrapbooks would fall under what Foucault (1972) would characterise 'normative discursive agencies', mechanisms that on some levels shape people's memories or at least steer them a particular way. Nevertheless, the above agencies often become censored and filtered according to social contexts and cultural norms, something that often, in terms of cultural memory, creates problems in the distinction between the personal and the collective and consequently between the personal and the society (van Dijk, 2004, p. 266). Therefore, according to Van Alphen (1999), cultural memory lies at the intersection of personal and the collective; with memory and culture as concepts, corresponding better to something that people 'create', something that allows them to form their individual and collective identities, rather than something that is owned (p. 268).

Modern historians, such as Huyssen (1995), distance themselves from the notion of 'representation' when it comes to memory. He claims that representation is what follows memory, since the latter is never authentic, but rather restrained to be expressed through images, videos, and texts; therefore, stranded in

'representation'. Therefore, it is safe to assume that every segment of memory presented in the Europeana platform, lies at the intersection of public and private. Van Dijk (2004) when it comes to the representation of memory suggested that:

> *The past is not simply there in memory, but it must be articulated to become memory. The fissure that opens up between experiencing an event and remembering it in representation is unavoidable. Rather than remembering or ignoring it, this split should be understood as a powerful stimulant for cultural and artistic creativity.* (Van Dijck 2004, p. 2)

Over the last few years a shift has been noted in the academic world, a shift from the representation of memory to the 'mediation' of memory. Online media and sites are historically considered 'collective mediations of the past, where authentic or collective experiences are moulded into prefigured technological and narrative frames' (van Dijk, 2004, p. 271). A memory that is frozen and expressed through words, is considered to be a 'technologized' memory of sorts, usually present in online media, which simultaneously assists the human memory, all the while being perceived as a menace to the printed world; a reality which purists fear that hurts the concept of remembrance (Ong, 1982). Therefore, media is put in a rather precarious position by many, boosting creativity and at the same time contaminating memory. According to Urry (1996), the almost exclusively online presence of media, alters the way that images of the past are formed now. Instead of depicting the past, contemporary media generate specific memories, that sometimes might serve specific social and cultural power structures. With the abundance of digital tools freely available online, it is easier than ever to produce images and consequently memories. Therefore, the mediation of memory is not a static concept, but it can rather be negotiated constantly and consequently re-formulated. This is a rather significant process, especially regarding memories that have been established within problematic frameworks and sociocultural contexts that do not correspond to the present era, and therefore need to be interpreted in a different light, so as to address problematic interpretations of history.

The juxtaposition of memory and feminist studies assumes that the contemporary is construed by a past that often is problematic and impugned. Both fields suggest that the reason to study the past moves away from a purist academic worldview and allows for a better interpretation of the present, as well as highlight the importance and value of personal memories and experience within a universe of alike-minded ones. As mentioned above, more often than not, memories are created within the framework of dominant, power structures and social systems rooted on inequality. Therefore, the mediation of those memories in the present, leads to a distorted or at times inaccurate interpretation of the past. Hence, active remembering, and – especially at the present time – active, digital remembering, is the tool to challenge forgetting, erasure and oppression, rendering the proper inquiry of cultural memory to an act of political activism (Hirsch and Smith, 2002, p. 13).

Feminist scholarship has strived to bring stories that have been silenced or erased from the historical narrative to the surface and include them in the hegemonic cultural memory. Therefore, conjecturing cultural memory through feminism, allows for the challenging of the 'cultural recall and forgetting' (Hirsch and Smith, 2002, p.

11), as well as identifying the political importance and the reason behind the stories that have remained in the dark. Hirsch and Smith (2002) when discussing recent academic work on cultural memory, posit that:

> *… developments in feminism and work on cultural memory demonstrate that the content, sources, and experiences that are recalled, forgotten, or suppressed are of profound political significance. What we know about the past, and thus our understanding of the present, is shaped by the voices that speak to us out of history; relative degrees of power and powerlessness, privilege and disenfranchisement, determine the spaces where witnesses and testimony may be heard or ignored* (Hirsch and Smith 2002, p. 12).

In the case of the Great War, similar to many historical events throughout time, the role of women has been considerably minimized, since history was mostly written by men; a strange result if one considers that the world has always been equally populated by both. This erasure of women in history and especially during war times, results in a 'weird, unreal, uneven' representation (Ferrus, 2010, p. 65). As seen by an initial, tentative exploration of the Women in World War I collection from the Europeana 1914-1918 project, women had a significant social and cultural role in public spaces, paving the way for the independent modern woman.

Drawing attention once again to the present study, the analysis of resources in a more traditional, textual form, but also (audio)visual ones; will have the internet as a space of presenting their full research potential. As mentioned above, this project revolves around the notion of discovering hidden stories within the canon representation of significant historical periods, such as the Great War. By unearthing the alternative stories and social memories that might be facilitated within the sub collections of Europeana 1914-1918, it will also be made possible to also unearth the social, political, cultural, and economic contexts, where disparities were dominating people's lives.

1.4 Overall recommendations

Digital humanities projects that incorporate data science methods in literary or historical corpora have been on the centre of attention of many scholars, for the past decades. For example, Stanford University has launched a literary lab that juxtaposes an abundance of textual resources with data analysis methods, in order to determine the most useful and interesting subsets of different literary waves.[13] Coming from the same university, the *French Revolution Digital Archive* (*FRDA*) is a collaboration of the Stanford University Libraries and the *Bibliothèque Nationale de France* (BnF), in order to engender a digital version of the key research sources of the French Revolution; by making them available internationally, scholars are able to explore the historical resources using data science methods.[14] The present research will be relevant and probably interesting as a juxtaposition of methods for digital humanities professionals, as well as individuals functioning within the historical and cultural sector.

[13] See: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.
[14] See: https://frda.stanford.edu/.

For large datasets, such as the ones featured at the Europeana 1914-1918 collection, applying digital humanities methods and using digital tools offer possibilities that qualitative methods in the humanities do not, with algorithms playing a vital role in people's everyday life. Not only are they able to process an immensely wide array of inputs and variables for decision-making, but they also do it with an agility and an accuracy that transcends any possible set of human skills, and at a fraction of the cost. In particular, audio visual resources and similar kind of interactive media, as well as textual resources offer a solid ground for digital analysis within the Europeana platform. A lot of merit could be added in research processes and results by not only focusing on the exploration of individual resources, but rather on the analysis of larger patterns within the dataset, since algorithms are able to analyse massive amounts of data in a quick, efficient, and inexpensive manner (Manovich, 2016, p. 2).

During the digital turn that has been gaining ground over the past decades, technological and cultural research often coexist and are implemented in such a way that the outcomes complement each other. Working with big cultural data helps with the continuous challenging of standard methods and approaches in the fields of social science and the humanities. Cultural analytics, as an imminent trend of the 21st century is defined as the application of computational methods used for the research of big data sets and flows (a.o. Manovich, 2016) Therefore, using digital tools allows for the juxtaposition of computational techniques and cultural data sets with more traditional methodologies within the humanities. However, the developers of said tools should make sure that they are mastered 'from the inside out', making any potential biases unequivocal to the public (Tenen, 2016). The digitized past along with the 'semantic knowledge representation methods' could allow not only for the exploration of the resources, but also for the possible unearthing of hidden histories within Europeana's sub collections (Hughes, 2016, p. 225); both aspects being of great use for the purposes of this research.

It is important to point out here the value of making a connection to the 'user side' or 'sociology' of digital humanities. User-centred design methodology (see e.g. Zabed Ahmed et al 2006) allows for studying engagement when using the platform – so studying *in interaction* and *in practice* – and thus studying more specifically how users and technologies co-construct meaning. This includes but also goes beyond user evaluation. For instance, from a critical digital hermeneutic perspective, the provided 'maps' or overviews of the data, topics and sentiments in the Europeana 1914-1918 sub collections from the previous section (§1.2) will be compared with the participants' answers in Hagedoorn's user studies, which allows us to deeper understand what kind of stories and platform engagement Europeana 1914-1918 affords the most, and how this can be built upon.

The value of 'raw' authentic materials to be able to reuse it is emphasized by the users in the study, and that for a large portions of the participants, items on the portal need to be better contextualized to make the reuse value apparent (as one participant put it for the purpose of the 'validity of source[s], placed in right context of time'). This is especially the case for sources of a more audiovisual nature. Especially the audiovisual annotation needs to be improved, also in particular for topic modelling (for e.g. filtering). Users also would like to see more examples of creative reuse as inspiration. Participants during the user studies' talk aloud protocols on their own initiative tried to search on positivity and negativity in the collection, usually to

be able to research two different sides of a story. Our data science methodologies offer different avenues to offer new forms of contextualization, which can also be used as keywords or filters. The user analysis also points to the need to update the filter option: for instance, whilst a large portion of the participants were interested in videos or user generated content, it took them quite a while to discover that filters for these categories existed (or they did not find them at all).

For the present study, we reached the best results with .CSV-files that we self-annotated. For more efficient data analysis in the future we would recommend Europeana to introduce the minimal standard for descriptions. For instance, many photos in the World War I collection have only a name of a person and a year in the description (especially in the Women in World War I collection). It leads to retrieving names and years in topics when we do topic modelling and in clusters when we do clustering. At the same time, we cannot retrieve information about what is actually in the picture (e.g. 'young woman with her daughter and husband at the beginning of the war'). This can also be achieved by automated image recognition (software which recognizes the objects in pictures).

The research can be developed further in several directions. Most of the steps were made for two collections – Women in WWI and Films. All of them can be repeated for other sub collections of Europeana 1914-1918 and for other Europeana collections (but specific attention should be paid to stop words, which must be changed for every specific case). Other sub collections can be scraped in the same way (but corresponding HTML-tags on web-pages of items from other collections should be changed in the script if they are different from the ones used for 'Films', 'Women' and others). The link to the collection specified in the script should also be changed.

The authors of Python library TextBlob, which was used for sentiment analysis in our research, do not specify the particular algorithm which is executed for scoring sentiment. When calculating sentiment for a single word, TextBlob uses a sophisticated technique known as 'averaging'. It finds words and phrases it can assign polarity to (examples are 'great' or 'disaster'), and it averages them all together for longer text such as sentences. Sometimes the results of this analysis do not seem logical. For instance, the example of sentiment analysis in TextBlob tutorial is the sentence 'TextBlob is amazingly simple to use. What great fun!' which is very positive. However, it gets a sentiment score 0.39166666666666666 and it is not clear why it is not higher.

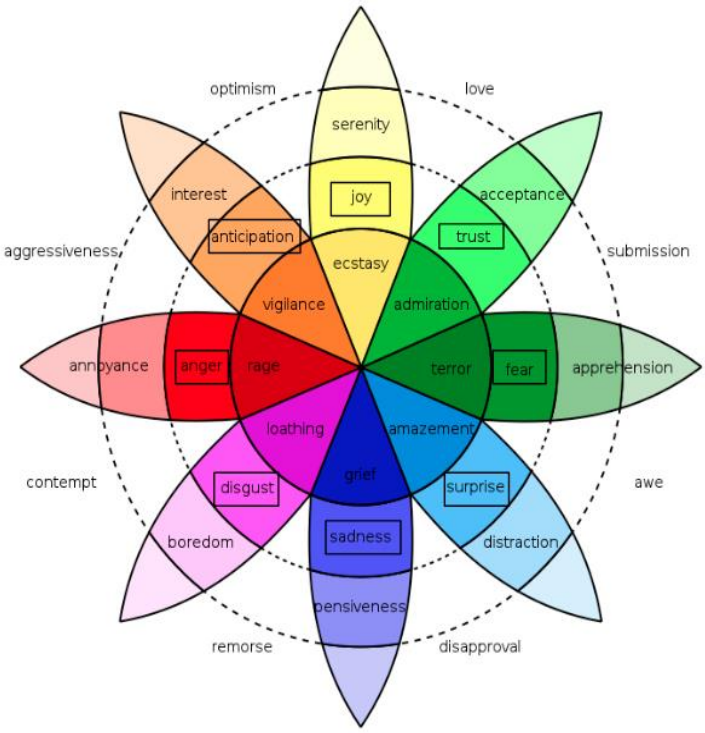
However, there are more efficient methods of sentiment analysis by using machine learning (Maas, 2011; Gautam, 2014), which can be applied for further development of this research in future. First, these methods demand manual annotation of the dataset by humans. Second, one of the existing machine learning models can be used to train on this data (the model 'learns' from the data some features of positive, negative or neutral sentiment in this particular dataset). Then this trained model can be applied on the new data (which was not used for training) and give more representative results than ready-to-use models like TextBlob.

This similar approach was used by Gautam (2014) for the sentiment analysis of Tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between these two. For this they first pre-processed the dataset, and extracted the adjectives that have some meaning (which is called feature

vector), then selected the feature vector list and thereafter applied machine learning based classification algorithms, namely: Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature.

Experiments were conducted using well-known methodologies for the analysis of (audio)visual and textual data. The use of state-of-the-art libraries and pre-trained models i.e. Word2Vec, topic modelling, and the Google Vision API confirmed the original hypothesis of the study, that it is indeed possible to discover hidden patterns, themes, and 'stories' in the data. Both the human-annotated labels and the labels from the pre-trained models, presented valid clusters of different topics, which could potentially be used in the Europeana archives to improve filtering options, recommendations, and search tags. In this manner Europeana connects more with the cultural heritage affective turn, for instance through filters and tags related to emotion. It is important here to keep the connection with the human annotator, who can recognize emotion whilst annotating labels.

It is also quite clear that although the pre-trained models provide a solid foundation for the above-mentioned improvements, the use of one or more human annotators (especially ones with domain knowledge on the subject) would provide deeper and more comprehensive connections between the items of the datasets. It would be well worth it to have two or more annotators to make sure that the produced labels are sensible and not conflicting. This can be achieved by implementing some inter-annotator agreement methodologies, such as the Kappa score.[15]



[15] For more information, please see: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

Fig. 27 *The Plutchik wheel of emotions.*[16]

There is such a strong need for human annotation because of the missing sentiment in the descriptions, as language is very benign and neutral (which makes sense if a documentalist transfers information to a digital heritage database). Sentiment analysis can help to get emotion back, but you do need to have a well-trained model that is based on a large number of words (as sentiment models are more trained in market research, political campaigns, product reviews, etc.).

Another worthwhile investment would be to make greater use of sentiment analysis tools and more specifically, examining the usage of the Plutchik Wheel of Emotions (see **Fig. 27**).[17] This methodology greatly improves on the basic sentiment analysis algorithms, in the sense that it does not just score texts according to the positivity or negativity of the feeling, but provides a wide range of sentiments. By implementing the Plutchik wheel of emotions, especially in textual resources, it is easy not only to parse individual opinions, but also to be able to assign values to them, which distance themselves from the elementary positive/negative distinction. The Sirrocco opinion extraction framework, based on the Plutchik's Wheel of Emotion, is able to parse large amounts of text into subjects and opinions.[18]

This, according to Sokolenko (2017) allows for a triad between author, subject, and opinion to be formed. Since most texts, vary in factuality, opinions could either be presented in one sentence or instead be elaborated on for several paragraphs. The Sirocco opinion extraction Framework hence, examines individual sentences and then proceeds to analyse larger parts of texts, so as to be able to form opinions. In order for the subjects to be determined, a 'parsing tree' is used, a tree that represents the syntax of a string, in regard to context-free grammar (Chiswell and Hodges, 2007, p. 34). Then, Named Entities and Noun Phrases – usually capitalized – are extracted, and grouped together using NLP algorithms, based on their role within the sentence (Sokolenko, 2017). Opinions are then sorted into emotions and qualities, using the Plutchik's framework on human emotions, which is the core of Sirrocco Opinion Extraction Framework. Plutchik elaborated on 8 basic emotions, which can be paired up, therefore, being ideally suited for 'algorithmic implementation' (Sokolenko, 2017). The emotions are: anger, trust, joy, anticipation, fear, disgust, sadness, and surprise, with more complex emotions being expressed as combinations of the above.

[16] Retrieved from https://www.6seconds.org/2017/04/27/plutchiks-model-of-emotions/ [16.06.2019]
[17] See: https://github.com/NVIDIA/sentiment-discovery
[18] See: https://github.com/datancoffee/sirocco

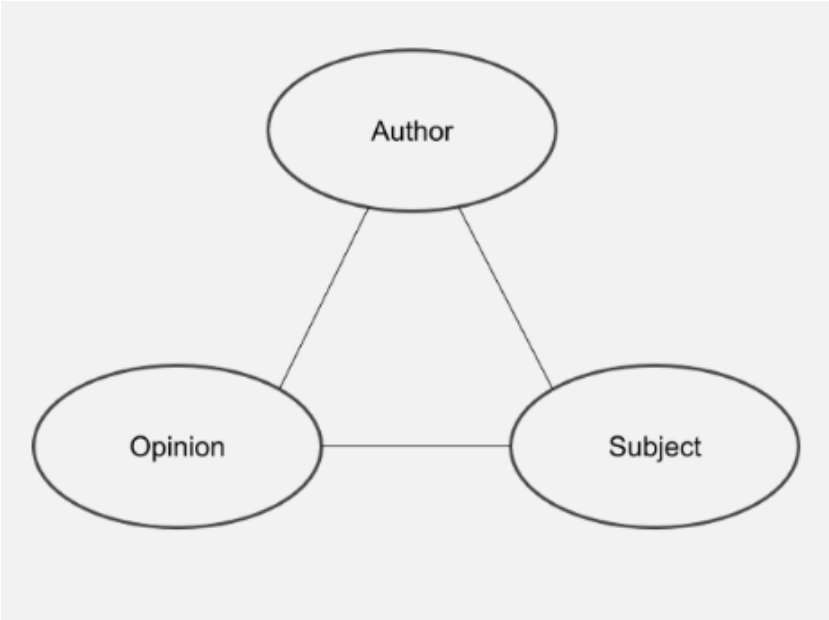


Fig. 28 *The Author-Opinion-Subject Triad.*[19]

We also reached conclusions on the perceived user experience. Europeana is a platform dedicated to the European digital cultural heritage, featuring both user generated content and linked open data, deriving from many different cultural institutions, such as museums, archives, and libraries throughout Europe. What renders the platform quite successful and widely usable amongst different target audiences, is its focus on the creative reuse. It is not just a portal that features audio visual and textual content, but rather a structure that allows the users to download, share, distribute, and use for academic, educational, and research purposes. Also, a major asset in the platform would be the numerous APIs, that allow the users to build applications focused on the collections that come from different European cultural institutions. from paintings to (audio)visual content. In particular, the SPARQL, the Record and the Entity APIs take advantage of structured metadata, and are able to return more specific and detailed results to the user who might be searching for something particular or detailed. Moreover, the user has the opportunity to contribute themselves in the formation of contextual labels by annotating, using the Annotations API.

First, from a cultural heritage perspective, within the datasets we annotated, a lack of diversity within the European context was noted for the linked (open) data. Whereas the lion's share of the open content came from users, something that speaks volume of the work that the platform does for user engagement, the volume of the linked open data that came from cultural institutions was comparably quite low. In particular, for our Women in World War I dataset, 826 items came from users, whereas only 86 came from museums, archives, and libraries. Moreover, Eastern Europe has a rather small representation, compared to Western Europe. Most of the linked open data comes from Germany, France, Netherlands, and the UK, whereas Southern and Eastern Europe could be represented more.

[19] Retrieved from https://medium.com/@datancoffee/opinion-analysis-of-text-using-plutchik-5119a80229ea, [20.06.2019]

From a technical perspective, we believe that the filtering processes for the collections that we worked with, could be improved, especially where the topics are concerned. By using the Cloud Vision API on collections that had previously been scraped, we realised that the sentiment labels that were produced could be implemented either as filters for searching purposes or in the form of tags. Whereas users can currently search based on topics, places, and people (usually contained in age or gender), emotions are not present in the search options. However, most cultural heritage professionals would agree that in the 21st century, museums and cultural institutions have taken an **affective turn**, reinventing their spaces as locations where people go to feel while simultaneously have their feelings challenged by featured narratives. This reinforced emotional role of the museum can be attributed to empathy; a rather important emotion for many visitors that can be triggered not just by following guided tours or reading interpretive material, but also simply by being present on location; deriving from that notion.

Hence, we do not see the reason why a digital space for cultural heritage, such as the Europeana 1914-1918, which includes a lot of stories that 'dabble' in emotions, could not adopt this prevailing affective turn. Therefore, Europeana could benefit from creating more detailed tags or tags that are based on sentiment and not on merely on descriptive practices. This addition could be carried out by either using an API, such as Cloud Vision or even better by incorporating human annotators in the process of labelling. Our project provides a set-up for this which can be replicated. Moreover, this approach could additionally improve user engagement, since by inviting users to create their own story categories based on the topics that the algorithm picked up, digital storytelling and creative reuse with the Europeana collection could also be further developed.

2 Overview of datasets and scripts

Data scraping

☑ Folder containing data science protocol, all datasets and scripts
☑ Our Python scripts for scraping

Translation

☑ Our Python script for automatic translation
☑ Translated datasets

Sentiment analysis

☑ CSV-files dataset case study Uncovering hidden stories in WWI Diaries and Letters
☑ Scripts case study Uncovering hidden stories in WWI Diaries and Letters
☑ Our Python script for sentiment analysis
☑ Overview translated data with sentiment
☑ Sentiment calculation Women in World War I
☑ Sentiment calculation Films
☑ Sentiment calculation Official documents
☑ Sentiment calculation Aerial warfare

Topic modelling

☑ CSV-files dataset case study Uncovering hidden stories in Women in World War I
☑ Scripts case study Uncovering hidden stories in Women in World War I
☑ CSV-file topic modelling Films
☑ CSV-file initial topic modelling Women in World War I
☑ Our Python script for topic modelling
☑ Our Python script for making topics using noun extraction
☑ Noun extraction Women in World War I
☑ Noun extraction Films
☑ Noun extraction Official documents
☑ Noun extraction Aerial warfare

Annotation – manual labelling

☑ Annotation using manual labelling Women in World War I
☑ Annotation using manual labelling Films

Annotation – automated labelling

☑ Our Python scripts for clustering using unsupervised machine learning
☑ CSV-file of Dataset labelled with 81 clusters

Cloud Vision API

- ☑ CSV-files dataset case study comparison with Cloud Vision API when using data science methodologies for (audio)visual sources
- ☑ Scripts case study comparison with Cloud Vision API when using data science methodologies for (audio)visual sources

3 Thank you

The principal investigator expresses their thanks and gratitude to Ksenia Iakovleva, Iliana Tatsi, Dimitrios Soudis, Susan Aasman, Jonas Bulthuis, Mark Span, all participants, and Alba Irollo, Milena Popova and Lorna Hughes and the Europeana Research Grant Program team.

4 Bibliography

Adabala, N., Datha, N., Joy, J., Kulkarni, C., Manchepalli, A., Sankar, A., and Walton, R. (2010). An interactive multimedia framework for digital heritage narratives. *Proceedings of the International Conference on Multimedia - MM '10*.

Antonie, L. et al. (2016). Historical Data Integration: A Study of WWI Canadian Soldiers. In Antonie, L., Gadgil, H., Grewal, W. and Inwood, K. Paper Presented at *2016 IEEE 16th International Conference on Data Mining Workshops*, Washington: IEEE Computer Society.

Assmann, J. (2008). Communicative and Cultural Memory. In A. Erll and A. Nünning (Eds.), *Cultural memory studies: An international and interdisciplinary handbook* (Media and cultural memory 8). Berlin: Walter de Gruyter.

Assmann, J. (2011). *Cultural Memory and Early Civilization: Writing, Remembrance and Political Imagination.* Cambridge: Cambridge University Press.

Associated Press. (2016, December 20). AP makes one million minutes of historical footage available on YouTube. Retrieved May 5, 2019, from https://www.ap.org/press-releases/2015/ap-makes-one-million-minutes-of-historical-footage-available-on-youtube

Bal, M., Crew, J. and Spitzer, L. (Eds). (1999). *Acts of Memory. Cultural Recall in the Present*. Hanover: University Press of New England.

Bernad Monferrer, E., Mut Camacho, M. y Fernández, C. (2013), Estereotipos y contraestereotipos del papel de la mujer en la Gran Guerra. Experiencias femeninas y su re ejo en el cine. *Historia y Comunicación Social, 18*, 169-189.

Bhowmick, P. (2010). Classifying Emotion in News Sentences: When Machine Classification Meets Human Classification. *International Journal on Computer Science and Engineering 2* (1), 98-108.

Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM, 55* (4), 77-84.

Borowiecki, K. J. and Navarrete, T. (2017). Digitization of heritage collections as indicator of innovation. *Economics of Innovation and New Technology, 26* (3), 227-246.

Burke, P. (2008). *What is Cultural History?* Malden, MA: Polity.

Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems, 31* (2), 102-107.

Chiswell, I. and Hodges, W. (2007). *Mathematical Logic*. Oxford: Oxford University Press.

Conrad, S., Krujiff, E., Suttrop, M., Hasenbrink, F., and Lechner, A. (2003). A storytelling concept for digital heritage exchange in virtual environments. *Virtual Storytelling, Proceedings, 2897*, 235-238.

Cheliotis, G. (2007). Remix culture: an empirical analysis of creative reuse and the licensing of digital media in online communities. School of Information Systems, Singapore Management University, 10 January 2007, p. 1-12. Available at: https://www.academia.edu/4312099/Remix_culture_an_empirical_analysis_of_creative_reuse_and_the_licensing_of_digital_media_in_online_communities

Cutting, J. E., Brunick, K. L., Delong, J. E., Iricinschi, C., and Candan, A. (2011, 01). Quicker, Faster, Darker: Changes in Hollywood Film over 75 Years. *I-Perception, 2*(6), 569-576. doi:10.1068/i0441aap.

Dalbello, M. (2004). Institutional Shaping of Cultural Memory: Digital Library as Environment for Textual Transmission. *The Library Quarterly: Information, Community, Policy, 74* (3), pp. 265-298.

Dashtipour K, Poria S, Hussain A, et al. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cognit Comput*, 8, 757-771.

De Kosnik, A. (2017). Rogue Archives: Digital Cultural Memory and Media Fandom. *European Journal of Communication, 32*(5), 504-505.

De Leeuw, S. 'European Television History Online: History and Challenges,' *VIEW: Journal of European History and* Culture 1,1, 2012, 3-11.

Ferrús I Batiste, J. (2010). Las mujeres en la historia. *Igualdad de Género en el ámbito público y privado,* F. Isonomia, Universitat Jaume I, Castellón.

Gillis, J. R. (1994). Memory and Identity: A History of a Relationship. In J. R. Gillis (Ed.), *Commemorations: The Politics of National Identity* (1-24). Princeton: Princeton UP.

Gonçalves, P. et al. (2013). Comparing and combining sentiment analysis methods. *COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks*, 27-38.

Goodman, S. and Parisi, L. (2010). Machines of memory. In: S. Radstone and B. Schwarz (Eds.). *Memory: Histories, Theories, Debates,* pp. 343–362. New York: Fordham University Press.

Haddon, L. (2011). 'Domestication Analysis, Objects of Study, and the Centrality of Technologies in Everyday Life,' *Canadian Journal of Communication 36* (2), 311-323.

Hagedoorn, B. (2015). Towards a participatory memory: Multi-platform storytelling in historical television documentary. *Continuum, 29*(4), 1-14.

Hagedoorn, B. (2016). Doing History, Creating Memory: Representing the Past in Documentary and Archive-Based Television Programmes within a Multi-Platform Landscape (Doctoral dissertation). Zutphen: CPI Koninklijke Wöhrmann.

Hagedoorn, B. and Sauer, S. (2019). The Researcher as Storyteller: Using Digital Tools for Search and Storytelling with (audio)visual (AV) Materials. *VIEW Journal of European Television History and Culture, Special Issue on (audio)visual Data in Digital Humanities*, eds. Pelle Snickars, Mark Williams and Andreas Fickers, Spring 2019. Open access, online multi-media article. Available at: https://www.viewjournal.eu/articles/abstract/10.0000/2213-0969.2018.jethc159/

Hagedoorn, B. (2019). Creative Re-Use and Storytelling with Europeana 1914-1918. Report Europeana Research, 25 June.

Haight, F.A. (1967). *Handbook of the Poisson Distribution*. New York: John Wiley and Sons.

Halbwachs, M. (1992). *On Collective Memory*, (Ed. and Trans.) L. A. Coser. Chicago: University of Chicago Press, (orig. written 1925 as *Les Cadres sociaux de la m´emoire* and published 1950 as *La M´emoire collective*, Presses Universitaires de France, Paris).

Haskins, E. (2007). Between Archive and Participation: Public Memory in a Digital Age. *Rhetoric Society Quarterly,* 37, 401-422.

Hirsch, M. and Smith, V. (2002). Feminism and Cultural Memory: An Introduction. In M. Hirsch and V. Smith (Eds.), *Signs*, Vol. 28 (1), Gender and Cultural Memory Special Issue (1-19). Chicago: The University of Chicago Press.

Hoskins, A. (2009). Digital network memory. In A. Rigney and A. Erll (Eds.), *Mediation, Remediation, and the Dynamics of Cultural Memory,* pp. 91–106. Berlin: Walter de Gruyter.

Hoskins, A. (2018). *Digital memory studies: Media pasts in transition*. NY: Routledge.

Hughes, L. M. (2016). Finding Belgian refugees in Cymru1914.org: Using digital resources for uncovering the hidden histories of the First World War in Wales. *Immigrants and Minorities, 34* (2), 210-231.

Jacobi, C., van Atteveldt, W., and Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, doi: 10.1080/21670811.2015.1093271

Jussi, K., Sahlgren, M., Olsson, F., Espinoza, F. and Hamfors, O. (2012). Usefulness of sentiment analysis. In *European Conference on Information Retrieval*, pp. 426-435. Berlin, Heidelberg: Springer.

Kanungo, T., et al. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24* (7), available at: https://ieeexplore.ieee.org/document/1017616

Keightley, E. and Schlesinger, P. (2014). Digital Media - Social Memory: Remembering in Digitally Networked Times. *Media, Culture and Society, 36* (6), 745-747.

Keller, P., Margoni, T., Rybicka, K., Tarkowski, A., and IViR. (2014). Re-use of public sector information in cultural heritage institutions. International Free and Open Source Software Law Review, 6(1), 1-9.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica 31*, 249-268

Kuhn, A. (2000). A journey through memory. In S. Radstone (Ed.), *Memory and Methodology*, (183-186), Oxford: Berg.

Li, H., Graesser, A.C., and Cai, Z. (2014). Comparison of Google Translation with Human Translation. *FLAIRS Conference*.

Loria, S. (21 November, 2018). Textblob Documentation Release 0.15.2. Available at https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf

Maas, A.L. et al. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* - Volume 1 (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-150.

Manovich, L. (2016). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. *Journal of Cultural Analytics*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, 3111-3119.

Moravec, M., Bolton, E., Hawkins, K., Heggan, S., Rao, J. and Smalley, H. (2017). The Great War Through Women's Eyes. *Pennsylvania History: A Journal of Mid-Atlantic Studies 84* (4), 452-461.

Nora, P. (1989). Between Memory and History: Les Lieux de Memoire. *Representations, 26*, 7–25.

Nosrati, F., Crippa, C., and Detlor, B. (2018). Connecting people with city cultural heritage through proximity-based digital storytelling. *Journal of Librarianship and Information Science, 50*(3), 264-274.

Owens, T. (2013). Digital Cultural Heritage and the Crowd. *Digital, 56* (1), 121-130.

Papadimitriou, C., Raghavan, P., Tamaki, H. and Vempala, S. (1998). Latent Semantic Indexing: A probabilistic analysis (Postscript). *Proceedings of ACM PODS*, 159-168.

Poole, R. (2006). Memory, history and the claims of the past. *Memory Studies, 1* (2), 149-166.

Poort, J., van der Noll, R., Ponds, R., Rougoor, W., and Weda, J. (2013). The value of Europeana: the welfare effects of better access to digital cultural heritage. (SEO-report; No. 2013-56). Amsterdam: SEO economic research/Atlas voor gemeenten.

Pruulmann-Vengerfeldt, P. and Aljas, A. (2009). Digital Cultural Heritage - Challenging Museums, Archives and Users. *Estonian Literary Museum, Estonian National Museum, University of Tartu, 3* (1), 109-127.

Purday, J. (2012). Europeana: Digital Access to Europe's Cultural Heritage. *Alexandria, 23* (2), 1-13.

Rahaman, H. and Tan, B-K. (2009). Interpreting Digital Heritage: A Conceptual Model with End-Users' Perspective. *International Journal of Architectural Computing, 9* (1), 99-113.

Reading, A. (2003). Editorial. *Media, Culture and Society, 25* (1): 5–6.

Reading, A. (2012). Globital time: time in the digital globalised age. In: E, Keightley (Ed.), *Time, Media and Modernity,* pp. 143–164. Basingstoke: Palgrave Macmillan.

Řehůřek, R., and Sojka, P. (2010). 'Software Framework for Topic Modeling with Large Corpora.' In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA.

Rishab Jain C. and Kaluri, E. (2015). 'Design of Automation Scripts Execution Application for Selenium Webdriver and TestNG Framework' in *ARPN Journal of Engineering and Applied Sciences*, *10* (6), 2440-2445.

Saleh, B., Abe, K., Arora, R. S., and Elgammal, A. (2014, 08). Toward automated discovery of artistic influence. *Multimedia Tools and Applications, 75*(7), 3565-3591. doi:10.1007/s11042-014-2193-x.

Sanders, E. B.-N. and Stappers, P.J. (2008). 'Co-Creation and the New Landscapes of Design,' *Co-Design*, *4* (1), 5-18.

Savenije, B. and Beunen, A. (2012). Cultural Heritage and the Public Domain. *Liber Quarterly: The Journal of European Research Libraries, 22*(2), 80-97.

Schich, M., Song, C., Ahn, Y., Mirsky, A., Martino, M., Barabasi, A., and Helbing, D. (2014, 07). A network framework of cultural history. *Science, 345*(6196), 558-562. doi:10.1126/science.1240064.

Schwartz, B. (2010). Culture and Collective Memory: Two Comparative Perspectives. In J. Hall, L. Grindstaff and M-C. Lo (Eds.), *Handbook of Cultural Sociology*. London: Routledge.

Serrà, J., Corral, Á, Boguñá, M., Haro, M., and Arcos, J. L. (2012, 07). Measuring the Evolution of Contemporary Western Popular Music. *Scientific Reports, 2*(1). doi:10.1038/srep00521.

Smith, D. A., Cordell, R., and Dillon, E. M. (2013, 10). Infectious texts: Modeling text reuse in nineteenth-century newspapers. *2013 IEEE International Conference on Big Data*. doi:10.1109/bigdata.2013.6691675.

Sokolenko, S. (2017, May 04). Opinion Analysis of Text using Plutchik. Retrieved June 20, 2019, from https://medium.com/@datancoffee/opinion-analysis-of-text-using-plutchik-5119a80229ea

Sturken, M. (1997). Tangled Memories: The Vietnam War, the AIDS Epidemic, and the Politics of Remembering. Berkeley: University of California Press.

Tasker, G., and Liew, C. (2018). 'Sharing my stories': Genealogists and participatory heritage. *Information, Communication and Society,* 1-18.

Teissier, P., Quantin, M. and Hervy, B. (2018). Humanités numériques et archives orales: cartographies d'une mémoire collective sur les matériaux. *Cahiers François Viète*, Centre François Viète, Université de Nantes, Actualité des recherches du Centre François Viète, III (4), pp.141-177.

Tenen, D. (2016). Digital Humanities and its Methods. Blunt Instrumentalist: On Tools and Methods. In L. F. Klein (Ed.). Gold, M. K. Debates in the Digital Humanities 2016. Minneapolis; London: University of Minnesota Press.

Terras, M. (2015). So you want to reuse digital heritage content in a creative context? Good luck with that. *Art Libraries Journal, 40* (4), 33-37.

Toms, E., and Duff, W. (2002). 'I spent 1,5 hours sifting through one large box.' Diaries as information behavior of the archives user: Lessons learned. *Journal of the American Society for Information Science and Technology, 53*(14), 1232-1238.

Underwood, T., Black, M. L., Auvil, L., and Capitanu, B. (2013, 10). Mapping mutable genres in structurally complex volumes. *2013 IEEE International Conference on Big Data*. doi:10.1109/bigdata.2013.6691676.

Van der Akker, C. and Legêne, S. (2016). Museums in a Digital Culture: How Art and Heritage Become Meaningful. Amsterdam: University of Amsterdam Press.

Van Dijck, J. (2004). Mediated memories: personal cultural memory as object of cultural analysis, *Continuum: Journal of Media and Cultural Studies, 18* (2), 261-277.

Van Dijck J, Poell T. and De Waal M. (2018). *The Platform Society: Public Values in a Connective World*. Oxford: Oxford University Press.

Volcani, Y. and Fogel, D. B. (2001). System and Method for determining and controlling the impact of text. *USA,* (7).

Wagstaff K., et al. (2001). Constrained K-means Clustering with Background Knowledge. Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577-584.

White, H. (1980). The Value of Narrativity in the Representation of Reality. Critical Inquiry, 7.1: 5-27.

Wildemuth, B. and Freund, L. (2012). 'Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors,' in Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, ACM.

Zabed Ahmed, S.M., McKnight C. and Oppenheim, C. (2006) 'A User-Centred Design and Evaluation of IR Interfaces,' *Journal of Librarianship and Information Science*, *38* (3), 157-172.

Zorich, D.M. (2003). A survey of digital cultural heritage initiatives and their sustainability concerns. *Council on Library and Information Resources*, Retrieved from http://www.clir.org/pubs/reports/pub118/contents.html.

Co-financed by the European Union
Connecting Europe Facility