6 Internal Site Search Recommendations to facilitate the use of Europeana's cultural heritage collections in History Education

Research Framework on Improving Discoverability, Opening Up Historiana Project, part of Europeana Digital Service Infrastructure 4

EuroClio, European Association of History Educators, The Hague, the Netherlands

Fani Partsafyllidou, Research and Development trainee

June 2020

Contents

Fani Partsafyllidou, EuroClio - European Association of History Educators

Introduction

On the research

Needs assessment

In this paper, technical suggestions that will facilitate searching for historical sources in Europeana platform will be presented. These recommendations are made by EuroClio based on the preferences and search behaviour of the educational community. To find these out, EuroClio conducted a field research on needs assessment.

The findings showed that the more contextualized an item is, the more useful it is for history education, provided that the item is of historical interest in the first place. In fact, Items with adequate, comprehensible descriptions are 58% more likely to be included in a lesson[1].

Challenges

However, most objects do not have a description, which makes searching difficult. Meanwhile, the prospect of having all European items (48 million, in May 2020) curated by professionals is not probable in the foreseeable future.

Currently, stakeholders acknowledge the importance of curation, but find an obstacle in the perception that curation is not scalable, that millions of items cannot possibly be curated. This is exactly the issue we want to contribute solving in EuroClio.

Solutions

In this research we will share 6 recommendations for new functionalities in Europeana's internal search engine that will result in automated curation. These new functionalities use the data we already

---

[1] Fani Partsafyllidou, Digital innovation in History Education – a Field Research on Needs Assessment p. 23

Fani Partsafyllidou, EuroClio - European Association of History Educators

have on each item, then process them and combine them, creating more information. Out of those 3 recommendations:

· 3 filters determine → Is the item of historical interest? Is it curated?
· 2 functions enhance metadata
· 1 function improves sorting the results

Aim

Each sector approaches our cultural heritage in a different way; an artist needs to find items with high resolution, or with a specific colour -- a history teacher needs to find items that are dated and curated. This means that search engine optimizations has to act differently according to the user's profession or to provide different filtering options specialized for each sector. Ultimately, the aim of the research is to provide insight on how Internal Site Search can be customised for History.

On the project

A few words about Europeana

Europeana is the largest cultural heritage platform in Europe, hosting the digital collections of national museums, galleries, and archives. Any cultural heritage institution can join the Europeana community and make use of its technological services such as transcription, translation, etc. Although it was created by the European Union, it welcomes collections from other countries too. Europeana.eu is free, accessible by anyone. For the average user, searching in Europeana equals to searching 3.000 databases of cultural heritage.

A few words about Historiana

Historiana.eu is a platform for history teachers. It enables them to search Europeana for images, save them to their account, and then use them to create lessons, especially for interactive eLearning activities.

How is this research connected to Europeana?

This research is part of the activity 'Improving Discoverability' of the project 'Opening Up Historiana', a Digital Single Infrastructure activity, part of Europeana DSI-4. It is implemented with the financial support of the CEF Telecom Programme of the Innovation and Networks Executive Agency (INEA) of the European Commission. The aim of the project is to promote the digital collections of Europeana from the scope of historical education. This research explores the preferences and the search behaviour of history teachers, which is a valuable element of the Needs Assessment. Its purpose is to make sure that the technological developments in Europeana and Historiana meet the current needs of the European educational community

Recommendations

Filters

Three filters will be presented. They can be used independently, or they can be merged into one filter, called 'Recommended for History Education'. It should be noted that an item which does not show up under the filter 'Recommended for History Education' is not necessarily irrelevant for History Education. On the contrary, most items can be used to teach history if someone devotes time to find more information and contemplate on its historical value.

The question we should be asking is: Can the historical value of this item be understood by someone who has no ties with its origin country? Since Europeana has reached dozens of millions of items, history teachers need to be able to explore the collections effectively. 'Recommended for History Education' means that the item is more likely to be clearly useful to History Education regarding its content and the pieces of information provided.

A. Historical dating precision.

Challenge: Searching by date favours items with worse dating precision

When we are dealing with cultural heritage, and thus the past, the perception of time changes. Items can belong to a decade, a century, or even more, depending on the time period. However, here's where things get ineffective. An item whose dating span is two centuries will be visible more times in a search by year, whereas an item with a precise dating will show up less times. Meanwhile, the item with the exact dating is more likely to have educational value because it is related to a more specific historical context.

Users can now search by year, which is admittedly more than enough for the vast majority of users. However, this can be improved in a fascinating way. If users can filter by historical dating precision, Europeana will become the most important platform for finding new sources in History, both in schools and in academia.

Solution: Evaluating if an item's dating is precise or approximate. Filter accordingly

It is not only valuable, but also easy to add this filter. First, we define what historical dating precision is. The later something was created the easier it is to classify it in a time period. Thus, the concept of historical dating precision is relative. For example, for a photograph captured in 2002 we would expect to know the day it was created, while for a medieval manuscript at 50 years' period could be enough. Specifically, there are also exceptions to this rule, with the most well-known being the Dark Ages, in which we have less historical sources than Classical Antiquity from which we have more historical records. In this case it is not necessary that items from the Dark Ages can be easier to date than the ancient ones. Although this contemplation is interesting, it goes beyond the limits of the current research.

Below you can find an exemplary taxonomy of historical dating precision.

Relative Taxonomy (Recommended)		
If the items dates to	The dating is exact if its span is less than	The dating is approximate if its span is less than
40.000 BC – 20.001 BC	-	-
20.000 BC – 1201 BC	-	1000 years
1200 BC – 401 BC	50 years	100 years
400 BC – 399 AD	20 years	50 years
400 – 1199	50 years	100 years
1200 – 1599	50 years	100 years
1600 -- 1799	20 years	50 years
1800 – 1939	5 years	20 years
1940 – now	Day	5 years

Absolute Taxonomy (Not Recommended)		
If the items dates to	The dating is exact if its span is less than	The dating is approximate if its span is less than
anytime	25 years	150 years

Date entries in Europeana's metadata should be streamlined into the same date formats. Subtracting the value of the earliest possible dating (terminus post quem) from the value of the latest possible dating (terminus ante quern) gives us the estimated range. This number should be compared to the second column of the table. If it is smaller, then the item is classified as having an exact date. If it is larger, then it is classified as approximate.
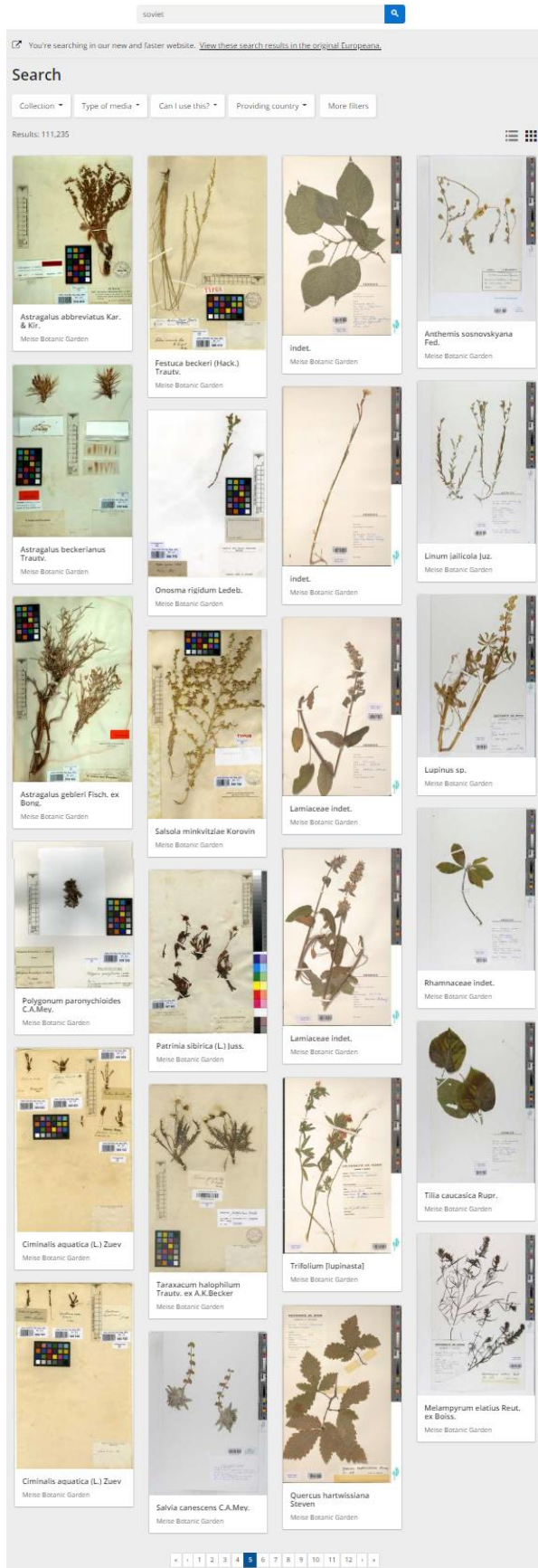
B. Face Recognition

Challenge: Many results are irrelevant to History

Within minutes of browsing in Europeana's digital collections, one can understand the vastness of the term cultural heritage. Practically anything that has ever been created with a religious or aesthetic motive in the history of human kind is part of our cultural heritage. This becomes overwhelming for the average history teacher, who needs 2-3 sources for her lesson.

Solution: A Face Recognition filter

This is a filter that concerns Images in Europeana, which is currently the predominant type of media (26 million images, 20 million texts, May 2020). These are by far the two most popular media types, combined they amount for 95,83% of all Europeans items (46 out of 48 million). Searching texts is more straightforward, once they are transcribed -- the technology behind this procedure is currently under construction. The words inside the text can be searched as well. We need a different approach when searching images.

In our survey, in which we asked history educators from Europe which sources they are most likely to use in a lesson, we found that teachers would prefer the pictures that included a human figure. Two out of ten images did not include a person and were not a map, and these images were 95% least likely to be used in a lesson.

Assuming three categories that encompass everything: People, manmade objects, and nature. History is the study of human societies and as such, we are interested in people and manmade objects. Face recognition can speed up the process of finding a historical source by omitting irrelevant results. In the example on Screenshot (i) we search 'Soviet' using no filters. Botanical results start in the third page, and fully cover the fifth page. If we could use a filter expressing that we want results that contain people,

Screenshot i Fifth page of search query 'soviet' is full of plant specimens.
https://www.europeana.eu/en/search?page=5&query=soviet&view=grid

face recognition would remove an impressive amount of the irrelevant results.

Same goes for manmade objects. Needs assessment shows that images with people are preferred in education. This means that out of two results, one with a vase, and one with a person holding a vase, the latter would be preferred. This is attributed to the pedagogical value, the teacher wants to show the relationship between the person and the vase, not the vase itself – unless it is an Art History lesson. Another reason is that images with people are simply more appealing, which is reflected in school textbooks too.

Ideal case: Include option for 2+ people

On the other hand, a problem that may arise when using face recognition for historical purposes is that portraits might be overrepresented in the results. History is the study of human societies, not facial characteristics. It is not of a historical significance how do people look, yet vanity is a timeless human flaw so hundreds of thousands of close-up portraits have remained as cultural heritage. These portraits will bypass the face recognition filter. An advanced implementation of the filter could include the option to search for items with 2+ people in them, which depict the relationships between people.

C. Meta-curation

Challenge: Items often lack historical context

Items with adequate, comprehensible descriptions are 58% more likely to be included in a lesson, as mentioned earlier. However, the descriptions of the items are often left blank or mention the bare essentials such as the tile and other pieces of information that are found in the metadata. In these cases, the item cannot be used as a historical source because there is not enough information for the user to make sense of it from a historical aspect.

Meanwhile, there is no way for the user to know if the items include a description or not, except if she visits them, which is practically impossible for a large amount of sources.

Solution: Filter items as curated or not curated

The user has to be able to express that curation is essential as a criterion, if so. This can be done by selecting a filter to have the non-curated items removed. The result can be calculated based on the amount of words in the description. If it includes more than 70 words it should be considered curated.

Ideal case: Skip information from the metadata in the word count

To increase the chances that the description is meaningful, we should make sure that the description is original. Often, it is just a repetition of the title and the name of the museum where the item is archived. Comparing the words in the description and those in the rest of the metadata, we can find which words repeat themselves. Then, these words should be subtracted from the description's word count, and afterwards this number should be compared to the set limit of 70 words. This increases the accuracy of the filter.

Metadata

A. Utilise information added by users

An upcoming feature of Europeana is that users will write comments under items. This will generate a vast amount of textual information. How is it going to be processed? What are the potential benefits for the internal site search?

Challenge: Users' comments will be diverse in topics and quality

Historical and artistic commentary can improve the internal site search, adding relative concepts and explaining what the source is about. This additional information can create more connections among the items and make them more searchable. However, users' input also has the potential to worsen internal site search, in the cases where comments are of low quality or irrelevant.

Solution: Enable users to tag their comments

The only way to make use of the valuable comments while discarding the rest is to apply a sort of evaluation. The most accurate way this can be done is if the users evaluate their comments themselves, by defining to which topic they refer.

A cultural heritage item can be analysed in three ways: as a work of art, as a historical source, or as a saleable object. Therefore, a tag system should reflect the typical discussions around cultural heritage items. It could include the following options, respective to the typical topics. The user can tag her comments accordingly. Of course, most people will write spontaneous remarks and personal afterthoughts. These comments should not by tagged under any category.

Tag	Topic
Depicted content	What is depicted and what does it mean?
Artist	Under which circumstances did the artist create it?
Artistic movement	How is it representative of an artistic movement?
Technique	What is noteworthy regarding the artist's technique?
Historical context	What was happening in the society at the place and time when it was created? How is it expressed?
Past and current ownership	How did it end up in the current institution?
-	Personal remarks

Overtime, this way the items will be curated by the users' community. Comments that are related to historical content will generate valuable keywords with related historical concepts and terms that will be searchable. The more searchable text associated with an image, the more visibility it gets.

Ideal Case: Tags personalise user's experience

If an item receives relatively many comments with the tag 'historical context' it indicates that it is of historical interest. If an item receives relatively many comments with the tag 'artistic movement' it indicates that it is related to Art Theory. In other words, the tags in comments suggest which groups of people will be more interested in which sources. A professional in the art business field is more likely to be interested in a source that has received many comments about 'past and current ownership'. Overall, tags on comments will enrich personalized user's experience.

B. Maps

Maps play a vital role in studying and teaching History. In this section we will explore how these items can be accessed in a more effective way.

Challenges: a. Taxonomy is unclear regarding maps

Currently, maps are characterised either as images or as texts interchangeably. I have failed to understand a distinction between the two given types. Please compare Screenshots I and ii below.
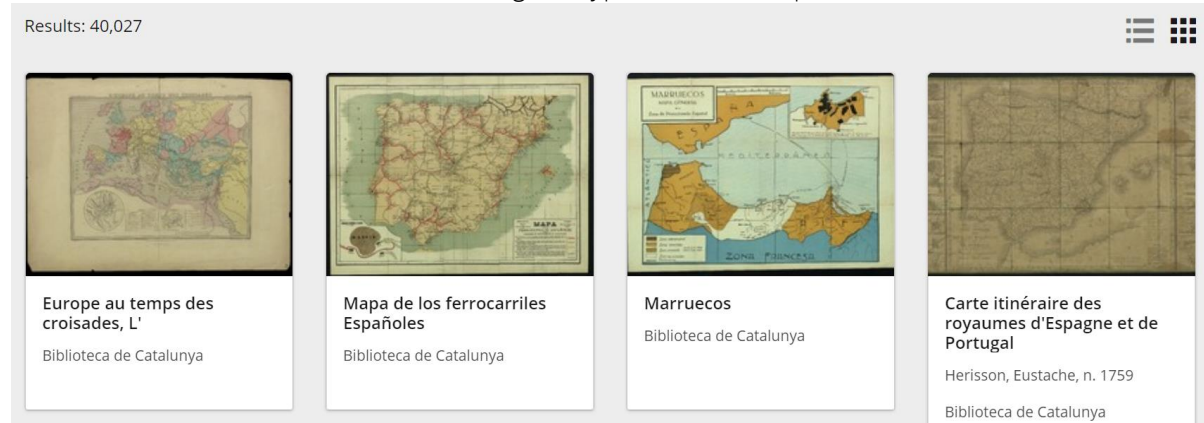

Screenshot ii First page results in the query 'map', free to use, Item type: Image
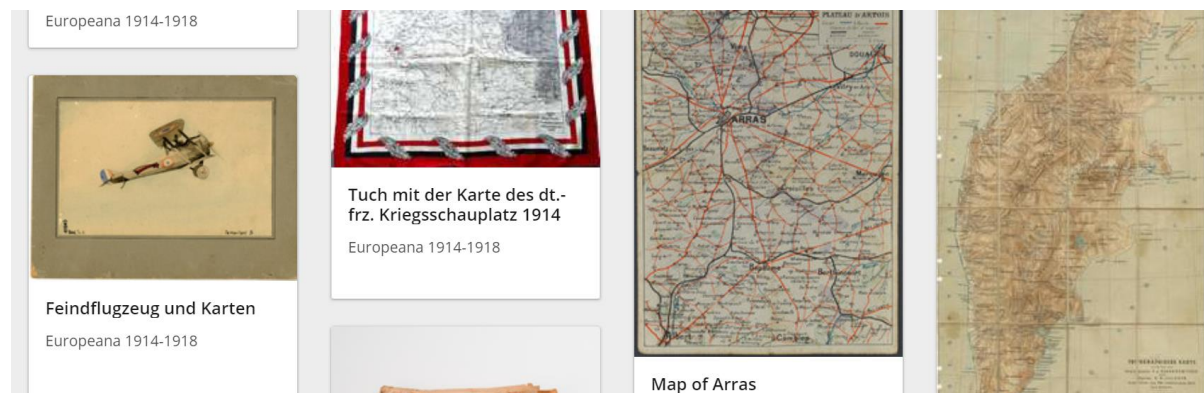
Screenshot iii First page results in search query 'map', free to use, Item type: Text

Providing a consistent way to present maps can speed up the user's experience.

Challenges: b. Users cannot find maps based on their content

A researcher cannot search the name of a city, a lake, or any other geographical term and find it within a map. Instead, she has to guess the exact title of the map. This is not ideal, since small landmarks

Fani Partsafyllidou, EuroClio - European Association of History Educators

are not guaranteed to be mentioned in maps, so using the wider geographical area as a term may not be fruitful.

Item i Map of Crimea https://www.europeana.eu/en/item/9200579/agmdy33a
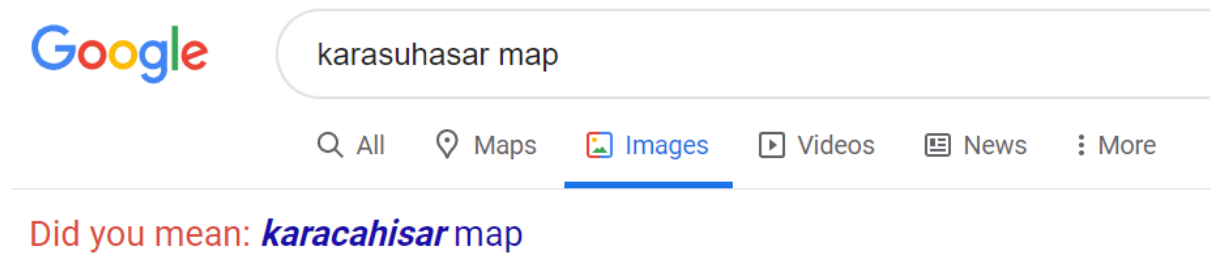
Solution: Classify maps as texts, run text recognition

Maps need to be tagged as text instead of image. Currently, maps are most commonly characterised as images. It would be helpful if photographs of maps are characterised as text instead of image, not to mention more accurate. The core pieces of information in a map are the geographical names, which are written inside the map in the form of a text. The colours and the design are secondary elements that support the text.

More importantly, choosing text as source type for maps is the most useful and effective way to go. It will enable the text-processing services to be used on maps. Handwritten maps will be transcribed in the same process the manuscripts are, providing a list of geographical names for every map. All those names then become searchable. Subsequently, processing maps as texts enables translation, which further improves the search.

This can boost Europeana's traffic; given that two-thirds of traffic to Europeana Collections comes from Google[2], anyone who searches an unknown geographical name will be lead to Europeana, if there is a map mentioning this place and the map is processed as text. For example, the location 'Karasuhasar' which is seen in Item I above, returns zero results in Google using the query 'Karasuhasar map. All in all, if a map has an x number of geographical names inside it, if transcribed it will appear as a result in x more search queries than now. Currently, only the title is a keyword. In the aforementioned example, this map will show up only to the users who type Crimea.

---

[2] Beth Daley, Making it easier to search and browse collections online
https://pro.europeana.eu/post/making-it-easier-to-search-and-browse-collections-online

Fani Partsafyllidou, EuroClio - European Association of History Educators

Google   karasuhasar map

Q All   Maps   Images   Videos   News   More

Did you mean: *karacahisar* map

Your search - **karasuhasar map** - did not match any image results.

Screenshot iv Geographical name depicted in Europeana item i returns zero results in Google

Sorting

Diversification of results

Challenge: Repetitive items

The order in which the items appear in the search results is crucial. Only the 1000 first results are accessible, so the order determines which results will be shown. What we would like to see is an additional criterion to the sorting of these results, apart from relevance, and that is: variety. To demonstrate that current behaviour of ordering that we would like to alter, two screenshots are presented below. These are the two first pages of results after searching 'red flag'.
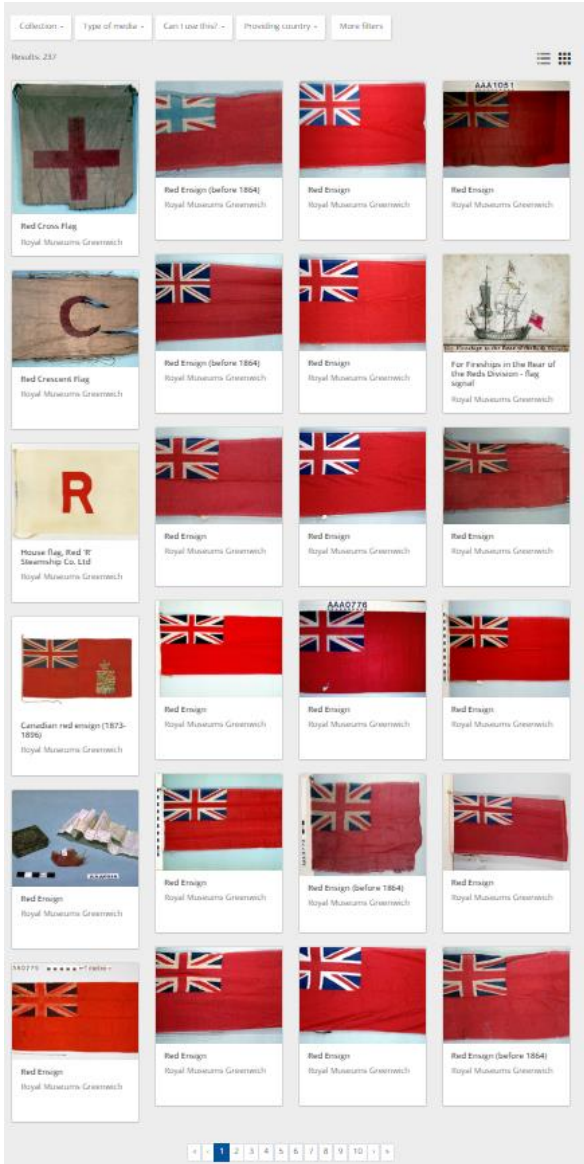
Screenshot v First page after searching 'red flag'

We notice that there is repetition of one result, the 'Red Ensign'. Although this is indeed a relevant item, so prioritizing it makes sense, the repetition of the same item in minimum variations is not helpful. Likewise, the second page is filled with identical items titled 'prayer flag'.
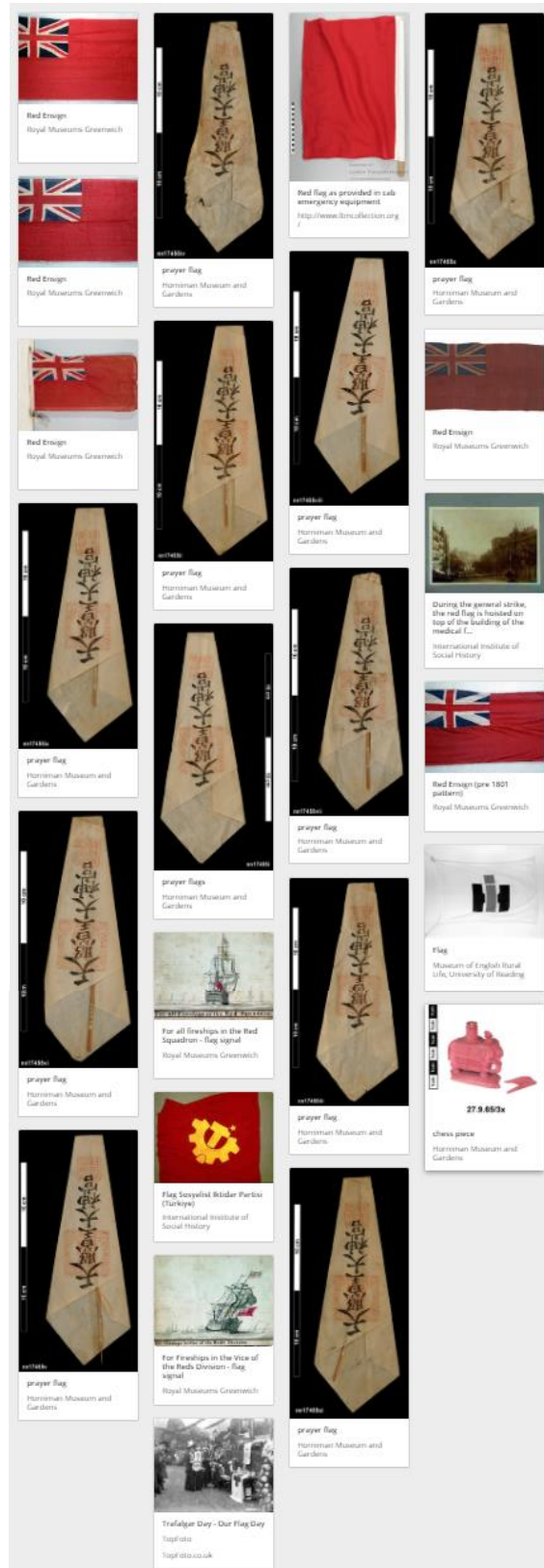
Screenshot vi Second page after searching 'red flag'

Fani Partsafyllidou, EuroClio - European Association of History Educators

Solution: Image processing to disperse identical items to separate pages

A technical way to overcome this challenge would be to have the content of the images analyzed. Then, if the content is close to identical these items should be dispersed to separate pages.

Ideal case: Apply image processing only to the first results

In case the analyzing of content is a heavy task which would cause the page loading time to increase, it can be partially applied. For example, out of the 1000 results that make it to the search result pages, image processing can be applied only to the first 100 results. This would create a balance between a fast and an accurate response, thus optimizing the user experience. Of course, if within the 100 first results there are 20 identical, they would be replaced with the next items and the analysis will have to run for another circle. The numbers 100 are indicative, and it should be tested how many times would the image processing have to run on average to generate 100 first results that are not identical. The heavier the task proves to be, the less images it should be applied to.

Summary

▇ Filters   ▇ Metadata   ▇ Sorting

	Challenge	Recommendation
1.	Searching by year favors items with worse dating precision	Add a filter with options: · Precise Dating · Approximate Dating
2.	Many results are irrelevant to History	Add a filter with options: · People Depicted · No people depicted
3.	No way to choose items with a description	Add a filter with options: · Curated · Not Curated
4.	Many results lack historical context	Utilise information added by users
5.	Taxonomy is unclear regarding maps, users cannot find maps based on their content	Classify maps as texts, run text recognition
6.	Repetitive items	Disperse identical items to separate pages using image processing

	Recommendation	Implementation	Ideal Case
1.	Add a filter with options: · Precise Dating · Approximate Dating	Simplify time formats, calculate time spans, use recommended taxonomy	←
2.	Add a filter with options: · People Depicted · No people depicted	Apply face Recognition on images	Include option for 2+ people
3.	Add a filter with options: · Curated · Not Curated	Count the amount of words in the items' descriptions	Do not count the words found elsewhere in metadata
4.	Utilise information added by users	Enable users to tag their comments to evaluate their relevance	Tags personalize user's experience
5.	Include maps' content in metadata	Classify maps as texts, run text recognition	←
6.	Disperse identical items to separate pages	Use image processing	Apply only to first results