Deployment and Maintenance of Europeana DSI core services - SMART 2016/1019

CONTRACT NUMBER - 30 - CE - 0885387/00-80

DELIVERABLE

C.3 Data access pattern report 2

Revision	1.0
Date of submission	30 April 2018 (M8)
Author(s)	Dasha Moskalenko, Europeana Foundation, Antoine Isaac, Europeana Foundation, Tim Hill, Europeana Foundation, Henning Scholz, Europeana Foundation, Hugo Manguinhas, Europeana Foundation
Dissemination Level	Public

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
0.1	19-4-2018	Dasha Moskalenko, Antoine Isaac, Ash Marriott, Tim Hill, Henning Scholz, Hugo Manguinhas	Europeana Foundation	Final version
1.0	30-4-2018	Victor-Jan Vos, Julia Schellenberg	Europeana Foundation	Final review

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

TABLE OF CONTENTS

Executive summary

This report describes the data we gather and the methods and tools we use to evaluate data access patterns of our users. The results gained shall help us to assist our users in finding the objects relevant to their needs.

The report provides an overview of the most popular blog posts, galleries, items and social media shares and analyzes those to find subjects our users were most interested in (between February and April 2018). Detailed statistics on API data access patterns are not yet fully extractable and will complement the evaluation in future reports.

Based on the evaluation so far we can conclude that most popular are items or content that was highlighted in editorial context (e.g. in a blog, gallery, exhibition) or on social media. Further, users were most interested in content that uncovered an untold story or existing content that was enriched with new facts. This result suggests that users value the relations that we can make between the items, as well as our editorial and social media efforts.

1. Introduction

The purpose of this report is to show how users interact with content and metadata on Europeana Collections, thus their most common data access patterns. Further, to analyse content and metadata accessed by users with the aim to anticipate and serve our users the objects they need.

Main question explored in this report is: "Which subjects were our users most interested in over the past three months February - April 2018?" This question is valuable to explore because finding subject groups that are interesting to our users could reveal clusters within existing themes or across themes. We can use these research findings to tailor the data access points of the site to the niche interests of our users and additionally uncover the unique selling points (USP) of Europeana Collections from a data perspective.

Methods used to collect data currently are highlighted and consideration is given to methods we have at our disposal but which are not yet fully operational.

The report concludes with an analysis of the data we have at hand including a number of assumptions about the most common data patterns.

2. Data access questions

The research question that we address in this report is "Which subjects were our users most interested in over the past three months February - April 2018?" This may be subjects (e.g. "women in the war"), specific time periods, artists or even broader concepts such a "historic artifacts."

Grouping our data access patterns by subject is a new method of grouping that we are exploring. On a higher level, Thematic Collections can be seen as grouping by theme to support users with specific needs and data access patterns. On a lower level, we look at which search queries users use most to find the data they are interested in. Finding subject groups that are interesting to our users could reveal clusters within existing themes or across themes.

We can use these research findings to tailor the data access points of the site to the niche interests of our users. We can use this information for the entity collection, to perfect our autocomplete, to focus our content sourcing and to steer our manual curation activities such as Thematic Collections, galleries, virtual exhibitions and browse entry points. It would also be interesting to explore if we are the only provider of data on these subject groups or have the largest collection. This is then one of the unique selling points (USP) of Europeana Collections from a data perspective.

3. Data access methods

We collect data on data access patterns from the following sources:

Google Analytics: this tool gives an overview of the items, blog posts and galleries that received most visits on Europeana Collections over a certain period of time. Additionally, we use it to identify most popular search queries executed on Europeana Collections. This gives us a good overview of the most popular content we hold and what our users are searching for on Europeana Collections (e.g. what search terms they use to find the information they are looking for).

API key logs: another source of information for funneling out most popular subjects on Europeana Collections is the use of our API keys[1]. We looked at how many active API users[2] we have (users that were either active for more than five days or made more than five

---

[1] An API user may reflect one or more actual users given that an API key can be associated to a single platform/portal such as Europeana Collections (which has its own API key).
[2] The Europeana REST API receives an average of 45 million requests from about 200 different users every month. From these we consider as active users the ones that demonstrate either a relatively frequent or high use of the API as they reflect services or applications that contribute to the reuse of the Europeana content in several of its markets.

accesses in a month) and what subjects they used to filter the data to extract the segment that they wish to access. For this activity we rely on our logs.

SPARQL: besides offering access for API users to Europeana Collections (through the Search API[3]), Europeana also offers a SPARQL[4] endpoint as a more advanced way for researchers to search through Europeana content. The SPARQL endpoint facilitates a more precise interaction with our data and enables one to search and correlate data based on specific metadata fields and data patterns[5]. Logging and analyzing SPARQL usage offers a granular view on how users, in this case researchers interact with our data and which subjects they find most relevant for their research.

OAI-PMH: not all users of the API interact with our data in a synchronous way. Some require that data is collected beforehand and indexed on their side so that they can provide a better service to their users. For this purpose, Europeana offers a OAI-PMH service[6] which allows API users to selectively gather a large amount of metadata records based on datasets and when the metadata records have been last updated (or created). This also makes it possible for API users to keep data up-to-date with the data that Europeana holds. Logging and analysing which datasets are being accessed and the frequency of their update gives us a general idea of what content is seen to be more relevant and the level of demand.

Broken links: we also analyse how many broken links there are on Europeana Collections that may be preventing users from accessing the content they need and damaging their experience. Similarly, we would also like to see if there is a correlation between the datasets with broken links and the popularity of a subject.

4. Data access measurements

4.1. Quantitative measurements based on user searches and visited content

Looking at the items, galleries and blog posts that received the highest number of unique visits from our users can illustrate which topics our users search for most or which topics our users are most interested in. The second step is to look for commonalities in these topics (e.g. time period, creator) that will allow us to group these by subject.

---

[3] Europeana Search API. View at https://pro.europeana.eu/resources/apis/search
[4] Europeana SPARQL API. View at https://pro.europeana.eu/resources/apis/sparql
[5] E.g. what are the French 18th-century painters with at least five artworks available through Europeana
[6] OAI-PMH service. View at https://pro.europeana.eu/resources/apis/oai-pmh-service

4.1.1 Top 10 blog posts

In the tables below you can find blogs post visited most by our users in the period between February to April 2018. Most viewed are recent editorials, however older editorials remain popular.

Blog post	Publication date	Number of visits	Topic
#ColorOurCollections and Europeana EYCH Colouring Book	February 2018	4030	Art, Campaigns, European Year Of Cultural Heritage, Downloads
Marco Polo – the man who brought China to Europe	January 2014	3001	Asia, China, Geography, Italy, Maps
Pictures in Focus: Street view in Amsterdam by George Hendrik Breitner	February 2018	1478	Photography, France, Amsterdam
Francis Bacon, Shakespeare and Secret Societies	January 2013	1443	History, Literature, Science
Nostalgic Viewfinder: The Instant Camera	March 2012	1179	Photography
Marie Curie: An Inspirational Woman for International Women's Day	March 2013	1154	History, Science, Women
Love across borders: the Belgian brothers who loved the same girl	March 2018	892	World War I, Love, User-generated content
Did Columbus really see mermaids?	January 2012	798	Art, History, Geography
5 Europeana Art highlights from 2017	December 2017	685	Art
Announcing our latest online exhibition Visions of War	February 2018	643	Art, World War I

Note: The pageview figures in this report represent when the user accesses that blog post in isolation. These figures do not account for viewing the contents of the blog through the overall blog feed, which accounts for 15% of the overall traffic to Europeana Blog.

4.1.2 Top 10 galleries

In the tables below you can find galleries visited most by our users in the period between February to April 2018. Most viewed are recent editorials, however older editorials remain popular.

Gallery title & URL	Publication date	Pageviews	Topic
Treasures of the Mauritshuis	September 2017	1795	Art
Treasures from the Finnish National Gallery	January 2018	988	Art
Clothes for the ballroom	April 2017	831	Fashion
Albert Edelfelt	January 2018	608	Art
Hats and Headgear	March 2018	581	Fashion
Famous Migrants	December 2017	496	Migration, History, People, Photography
European Landscape and Landmarks	May 2017	489	Art, Photography
Masculin Feminin	April 2017	447	Fashion
Women in World War I	March 2018	419	History, World War I, Women
The Kiss	March 2018	337	Photography, Love

4.1.3 Top 10 searches/query strings

The table below describes the top 10 query strings and number of times they were queried on Europeana Collections, during the period of February to April 2018.

Query string	Pageviews	Theme
Women in World War 1	1732	World War 1, Women, History
The Eastern Front	1420	World War 1, History

Letters from World War 1	1061	World War 1
Postcards from World War 1	961	World War 1
"Preparant els soldats"	919	History
World War 1 videos	871	World War 1
Leonardo da Vinci	830	Art, History
Paris	802	Multiple themes possible
Cockermouth	693	Multiple themes possible
Karl Heinrich Lammel	580	Photography

Europeana Collections had about 405,000 total queries, executed via the search box, in the period from February to April 2018. This is an average of about 5260 queries per day.

Topics requested are highly diverse, with about 241,000 unique query strings on Europeana Collections, in the period from February to April 2018. Comparing the total number of queries executed (c.405,000) with the total number of query strings (c. 241,000) shows that query strings are quite diverse and generally not used often. This is also because users use terms in many different languages or spell terms differently. Compared to the high diversity of query strings the amount of times our top ten query strings were queried seems to be noticeable.

4.1.4 Top 10 items accessed

The table below lists the top 10 items and number of times they were viewed on Europeana Collections, during the period of February to April 2018.

Object title & URL	Pageviews	Topic
Az eltűnt idő nyomában; Eltűnt idő nyomában | Marcel Proust	1921	Literature
Portrait of the son of the Catalan singer-songwriter Joan Manuel Serrat, Queco, born in Madrid on May 8, 1969 of his relationship with the Mercedes Doménech model	1683	Photography, Music
EXERCIT (sic) DE L'EST. PREPARANT ELS SOLDATS	1277	Film, War, History
Gabriele Ferzetti e la moglie, Maria Grazia Eminente, tra gli invitati del ricevimento	1234	Film, History

Elżbieta Starostecka i Włodzimierz Korcz z synem Kamilem w domu | PAP	1000	Photography
Matrimonio di Antonio Cabrini con Consuelo Benzi	965	Photography
ΙΣΤΟΡΙΑ ΤΟΥ 565 ΤΑΓΜΑΤΟΣ ΠΕΖΙΚΟΥ (ΠΡΩΗΝ 305 ΤΑΓΜΑ ΕΘΝΟΦΥΛΑΚΗΣ, ΜΕΤΕΠΕΙΤΑ 565 ΕΙΔΙΚΟ ΤΑΓΜΑ ΠΡΟΚΑΛΥΨΕΩΣ) ΤΗΣ 31/03/1979, ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ ΑΠΟ 01/01/1945 ΕΩΣ 16/08/1949	911	History, Military
Short Legends / author, Antonio Osete | Osete Pérez, Antonio (1876-1934)	653	Literature
The balandrito - Painting | Sorolla Bastida, Joaquín (Place of birth: Valencia, 02/27/1863 - Place of death: Cercedilla, 08/10/1923)	647	Art
Panorama de Paris pris du Pavillion de Flore	620	Photography, History

It appears that many of the top 10 items accessed lack a thumbnail and good quality data which prompts the question why these are the most accessed items. We hypothesize that this could be due to the spam traffic bots that regularly target Europeana Collections. This is a common issue that many websites struggle with. We will investigate if we can validate this initial hypothesis.

4.1.5 Items shared on social media that received most engagement

Engagement is a great indicator of popularity. Items that a lot of people find interesting usually receive a lot of engagement. Social media engagement measures the public shares, likes and comments, showing how many people interacted with content. It is a common measure of social media performance.

Post & URL	Publication date	Engagement	Topic
Find your crayons. Sharpen your coloured pencils. Arrange your felt pens. #ColorOurCollections is back! Learn more and download our colouring book http://bit.ly/2E5Owbe	05/02/2018	2431	Europeana EYCH Colouring Book
Happy Easter! How many hidden eggs can you spot	01/04/2018	1677	Easter Egg Hunt on an artwork

in this painting? 🥚🐣🐤🐥 Image: Market Square, with Flagellation, Ecce Homo and Road to Calvary in the background, Joachim Beuckelaer, Rijksmuseum, public domain Source: http://bit.ly/22DACD6			
Our #EYCH2018 Colouring Book features openly licensed content from 14 cultural institutions across Europe. Download it and enjoy discovering and colouring different forms and shapes of cultural heritage!🖼️🎨✏️ http://bit.ly/2E2jyRe #ColorOurCollections #EuropeForCulture	06/02/2018	1666	Europeana EYCH Colouring Book
Find your crayons. Sharpen your coloured pencils. Arrange your felt pens. #ColorOurCollections is back! Learn more and download our #EYCH2018 colouring book bit.ly/2E24dff #EuropeForCulture #coloringbook pic.twitter.com/dbx2fl3dfc	04/02/2018	1327	Europeana EYCH Colouring Book
PyeongChang 2018 Winter Olympic Games will start next week, so let's choose a winter sports image for our cover! Curling, ice skating, skiing or sledging? Vote by liking your favourite.	01/02/2017	1225	Monthly cover image voting

Today is #WorldHeritageDay! No time to visit monuments and heritage sites? See them online on Europeana! 🏯⛩➡ http://bit.ly/2qGC8FZ	18/04/2018	1073	World Heritage Day - linking to search results
Today Lithuania marks 💯 anniversary of the restoration of the state. A great occasion to explore masterpieces of Lithuanian art in our gallery 🖼👉 http://bit.ly/2CmCa8J #RestoredLithuania100 #EuropeForCulture	16/02/2018	781	Art
#MuseumWeek has started today with the first theme #WomenMW. Explore our gallery celebrating artworks by a selection of significant female artists from across Europe and throughout history http://bit.ly/2FmbM4y	23/04/2018	587	MuseumWeek/Art
"I believe that a girl should not do what she thinks she should do but should find out through experience what she wants to do." Amelia Earhart was the first female aviator to fly solo across the Atlantic Ocean. #WomensHistoryMonth Image: Amelia Earhart, Tekniska museet, public domain Source: http://bit.ly/2H3z9NA	4/03/2014	534	WomensHistoryMonth, Photography
Missing #ColorOurCollections? We have something special for #Easter 🥚🖌 Enjoy!	31/03/2018	445	Easter

http://bit.ly/1LP9FVB			
Happy birthday to the Eiffel Tower! #OnThisDay in 1889, Gustave Eiffel celebrated the completion of his structural work by leading a group of government officials, accompanied by representatives of the press, to the top of the tower. The #Eiffel Tower was originally intended as a temporary installation for the 1889 World's Fair and was almost torn down in 1909. Fortunately, officials put a stop to this. It has become one of the most enduring symbols of France and the industrial age. Explore many historical items related to the #EiffelTower via Europeana: bit.ly/europeana-eiffel-tower	31/03/2018	391	Photography

4.1.6 Top 10 sources of traffic to Europeana Collections

Below are statistics from Google Analytics on the top sources of traffic to Europeana Collections, including the average number of pages visited per session. From these statistics, we can see that Google is our primary source of traffic (c. 63% of all traffic). However, we get higher quality engagement referrals from other services such as Facebook and the Blogs. Reasons for this could include social network circles having similar targeted interests in certain types of content, as well as users from blogs being regular or engaged consumers of cultural content.

Traffic Source	Number of users (proportion of overall) February - April 2018	Pages per session
Google - Organic Search	511,749 (62.73%)	2.88

Direct Traffic/No Source	191,001(23.41%)	3.5
Facebook Mobile	7,900 (0.97%)	1.86
Creative Commons Search	7,158 (0.88%)	2.97
noticias.universia.es	4,836 (0.59%)	3.33
Bing - Organic Search	4,456 (0.55%)	5.44
bibliotecaspublicas.es	4,376 (0.54%)	2.26
Pinterest	4,074 (0.50%)	4.26
Facebook Desktop	3,540 (0.43%)	6.07
Europeana Blogs	2,698 (0.33%)	9.07

4.2. Quantitative usage of the API

Users also access data via the Europeana REST API. Since January 2017, EF has collected monthly usage statistics for the Europeana REST API. From these statistics we derived two metrics to determine the number of active users: 1) number of API keys that exceeded the average of 5 calls a day; 2) number of API keys that were active for more than 5 days in each month. In the table below you find information of the months 1 February to 30 April 2018.

Period	Number of API keys that…	
	exceeded the average of 5 calls a day	**were active for more than 5 days in each month**
February	87	81
March	65	66

4.2.1. Content accessed via SPARQL endpoint

The SPARQL endpoint was developed as a pilot service and therefore does not yet track usage. As part of our efforts to bring this service into a full production state, we will start logging all user requests and collect the relevant content statistics for future reports.

4.2.2. Content accessed via OAI-PMH harvesting

The OAI-PMH service is still in an alpha state and therefore does not yet track usage. This will be addressed once this service is moved into beta state by adding logging of all user requests. We will collect the relevant content statistics for future reports.

4.3. Link maintenance activity

As reported in the previous C.3 report (M4) we worked on broken links as part of a bigger analysis of Europeana datasets to improve data according to the Europeana Publishing Framework[7] (EPF). We identified that about 1200 datasets (about 13 million records) are not compliant to any of the EPF tiers (e.g. have an insufficient thumbnail size, lack of direct links to digital media) and about 60 datasets (1.2 million records) of those have broken links. The majority of these datasets have been de-published since September 2017 and all datasets larger than 5,000 records are free of broken links since April 2018.

The process for detecting broken links in the Europeana Collections galleries is now semi-automated. When a displayed item in a gallery encounters an error in retrievability, the data partner services team gets notified and can take appropriate action.

5. Analysis of the data access measurements

Looking at the top 10 most popular blog posts and galleries on Europeana Collections between February and April 2018 we can see an interesting pattern emerging (e.g. niche targeted content is the most popular). Biographical blog posts depicting famous artists receive a lot of visits, as do galleries designed using items from one museum or one painter.

It appears that our users appreciate when we dive deep to explore a particular historical figure (person) or an object (topic) providing information about their history, accomplishments or uses and their evolution through the years. From this we could hypothesize that our users are more interested in the story rather than the item.

Looking at the top 10 most popular query strings, the majority are related to the World War 1 theme. Based on this we can conclude that our users are highly interested in the First

---

[7] The EPF clarifies the relationship between the quality and openness of data and the impact it can have. Implementing this framework means that EF and all its data partners work towards high content quality so it is compliant with the higher tiers of the framework. In the EPF four tiers of participation were developed, to allow data partners to decide what they want to achieve with their digital collections, based on their own agendas and capabilities. The higher the quality of data provided by data partners, the more benefits they will create for their audiences. View at https://pro.europeana.eu/post/publishing-framework

World War theme, specifically in the subjects "Women in World War 1" and correspondence during World War 1. The reason why people are interested in the subjects "Letters from World War 1" and "Postcards from World War 1" could be because they want to uncover personal stories from that time period. As hypothesized above it appears that our users are more captivated by the stories rather that the content or items themselves.

This is interesting to explore because it would indicate that the 51,5 million items that Europeana Collections offers are not the only reason why users visit Europeana Collections. Such a diverse and rich collection presents the opportunity to find links between the items that can tell a story about a person, object, place or event. This seems to be the main perceived value of our trusted collection by our users.

Looking at the top 10 most viewed items we can spot two patterns: a) most of the items come from one of these two aggregators: EuroPhoto and The European Film Gateway; b) a fair percentage of them are items from Spain. To understand what could have contributed to the popularity of these items we looked at the top sources of traffic for this period. Unfortunately, we couldn't spot any clear correlation between the two. We plan to dive deeper into internal and external factors that could have potentially contributed to the popularity of this content and items.

A prominent internal factor that could have contributed to the popularity of certain items, are galleries and blog posts or our promotional activities on social media. The follow up question to ask here is "How much of this popularity has to do with our own promotional activities?" To find this out we need to compare the most popular content items and search queries to items shared on social media that received the most engagement.

Looking at the posts that received the most shares on social media we can conclude that participatory posts asking for people's opinion or offering people to download (e.g. colouring book for free[8]) receive most engagement. Additionally, posts that dive deep to explore a particular historical figure, object or place are also very popular which is similar to what we concluded for the most popular content. The one clear overlap between the most popular blog post and the most engaging social media post is the promotion of the Europeana colouring book.

Analysing the external factors we can conclude that the reason the blog post and the social media post about the colouring book were so popular is because this was our contribution to Colorourcollections.org[9]. This initiative invites libraries, archives, and other cultural institutions around the world to share free coloring sheets and books based on materials in their collections. Also, public holidays contribute to the popularity of posts. For instance, during Easter the Happy Easter! social media post performed very well and in March, the

---

[8] Colouring book for free. View at https://www.facebook.com/Europeana/posts/10155461846643668
[9] Colorourcollections.org. View at http://library.nyam.org/colorourcollections/

month of the international women's day, the post about Amelia Earhart received a lot of visits.

A major external factor that could have contributed to the popularity of a certain subject is market trends. For example this year the city of The Hague in the Netherlands is celebrating the year of the famous Dutch painter Mondrian. This could have sparked an interest from our (Dutch) users in Mondrian and his artwork. The follow up question to ask here is "Which external factors could have contributed to the popularity of these subjects over the past three months?"

6. Conclusion

The main conclusion that we can draw from this research is that one of the main values of the diverse and rich collection we hold are the relations that can be made between the items. This can be done by potentially uncover an untold story or enriching an existing one with new facts.

Our users value when we make connections for them in blog posts and galleries. Content that tells a story portraying the life on an artist or the history of a museum or its collection therefore receives the most attention.