europeana connect

M3.1.4

Report on log file analysis of the Europeana Prototype

Europeana: an evaluation of users, usage and information seeking behaviour derived from web-server log-files October 2009—September 2010.

co-funded by the European Union

The project is co-funded by the European Union, through the ***eContentplus*** programme

http://ec.europa.eu/econtentplus

Österreichische Nationalbibliothek

EuropeanaConnect is coordinated by the Austrian National Library

Distribution

Version	Date of sending	Name	Role in project
0.9	2010-10-13	Europeana WG1.1	
1.0	2010-11-07	D J Clark	UCL Consultants Ltd.
1.1	2010-12-14	Public	

Approval

Version	Date of approval	Name	Role in project
1.0	2010-11-07	D J Clark	UCL Consultants Ltd
1.1	2010-12-14	D J Clark	UCL Consultants Ltd

Revisions

Version	Status	Author	Date	Changes
0.8	draft	D Nicholas, I Rowlands, DJ Clark.	2010-10-11	
0.9	pre-release	D J Clark, D Nicholas	2010-10-13	add presentation charts
0.91	revised	D J Clark	2010-10-22	check data
0.95	revised	D J Clark	2010-11-29	revised text
0.96	rev.	D J Clark	2010-10-29	check tables
0.97	rev.	D J Clark	2010-11-04	graphics redrawn with R and gplot
0.98	rev	D J Clark	2010-11-07	reformat for EC
0.99	rev	D J Clark	2010-11-11	reassemble broken doc
0.995	rev	Adeline van den Berg, Jill Cousins	1.11.10	Reviewing for Europeana Connect
1.10	Final	D J Clark	2010-12-14	Review items noted

Table of Contents

Introduction **6**

Context 7

Confidence ratings 10

Summary of key findings **11**

Patterns of use 12

Hourly patterns 12

Weekly patterns 13

Monthly patterns 14

Annual patterns 16

Use by country 19

User preferences **24**

Page type 24

Media format 24

Most popular providers and collections 25

Browser and platform 26

Search and navigation **27**

Analysis of home page traffic 28

Analysis of thumbnail page traffic 29

Analysis of full-doc/record page traffic 30

Search terms 31

Most frequent referring sites 33

Robots 35

Future analysis of Europeana users and Information seeking **36**

Annex: Global reach of Europeana **38**

List of tables

Table 1: Weekly pattern of use; whole world

Table 2: Monthly pattern of use; whole world

Table 3: Use by Europeana Connect, and its contributors and partners

Table 4: Comparative annual growth rates by EU-27 country

Table 5: Europeana use by EU-27 country

Table 6: National preferences for national collections: selected EU-27 countries

Table 7: Europeana insularity rating: selected EU-27 countries

Table 8: Most popular page types viewed

Table 9: User media format preferences

Table 10: Most popular content providers

Table 11: Most popular collections

Table 12: Search terms: from home page

Table 13: Search terms: advanced search

Table 13b: People are currently thinking about...?

Table 13c: typed-in searches

Table 14: Most frequent referrer hostnames

Table 14b: Google search strings

Table 15: Europeana: per million capita use by country; whole world

List of figures

Figure 0: Statistical outliers

Figure 1: Hourly pattern of use; whole world

Figure 2: Heat map of hourly pattern of use; EU-27 countries

Figure 3: Page views by day of week; whole world

Figure 4: Page views: 12 months, whole world

Figure 5: Heat map of monthly pattern of use; top ten countries

Figure 6: Seasonal trends in use; whole world

Figure 7: Cumulated annual use; whole world

Figure 8: Comparative annual growth rates; EU-27 countries

Figure 9: Absolute use mapped across the whole world

Figure 10: Per capita use mapped across the whole world

Figure 11: Heat map of national preferences for national collections; selected EU-27 countries

Figure 12: Growth of mobile access; whole world

Figure 13: Analysis of home page traffic

Figure 14: Analysis of main search page traffic

Figure 15: Analysis of main content page traffic

Figure 16: 250 most popular search terms: typed into search box on home page; whole world

Figure 17: 250 most popular search terms: advanced search; whole world

Figure 18: Page views by humans, outliers and robots; whole world

Figure 19: Monthly robot and human activity compared; whole world

Figure 20: More signal less noise

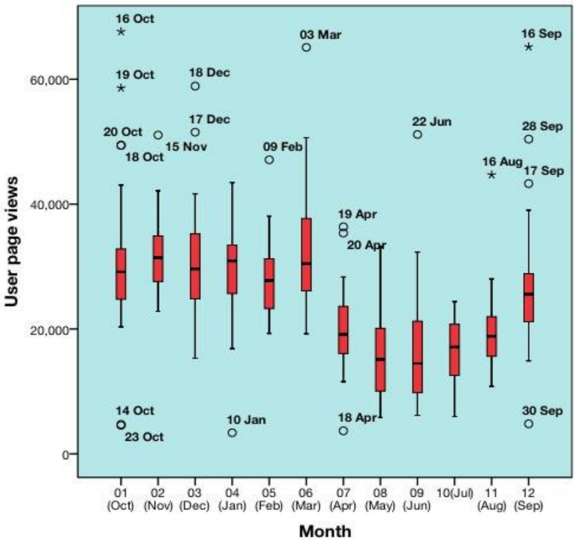
Introduction

This report presents an evaluation of usage and information seeking analysis derived from 12 months Europeana logs (October 2009—September 2010) of the Europeana Prototype. Our deep log analysis provides more detailed, accurate and robust data that can be obtained by Google Analytics. Nevertheless, what is provided here is still very much a work in progress. The log evaluation, like the Europeana service itself, is a prototype. Now, with a full year's historic data as baseline, we are in a position to measure the service over the next year as the project pursues a policy of active marketing to end-users..

An evaluation needs to consider, who the users are, and for whom value is added? We can identify several Europeana user groups and various concepts of value that can be measured. There are public users and, in more formal contexts, education and research users; there are content providers and those who provide curatorial value; nor should we neglect that the project itself forms a digital economy of software creation and experiment.

These are broad questions but not irrelevant to how we seek to extract information and insight of value from the analysis of log files. We look for evidence of the public visibility of Europeana: how many referrals, from where and when and how? We can establish what kind of users Europeana is attracting: casual browsers or deep researchers? We can ascertain how many leads Europeana provides to the providers' site: what evidence is there of wider access being provided, or added value for the digital scholar? Europeana provides a vector to promote and develop a digital economy: what new technologies are succeeding, does new content or new features build a community? Finally, for Europeana as a self-sustaining enterprise, where can log analysis add value, identify key interests, track an emerging market?

The most notable feature of an exploratory data analysis of europeana.eu log-files is the proliferation of unaccountable bursts of activity. On first encounter we mark them as outliers, aberrations to a presumed regular statistical pattern. But on setting aside and labelling the most outstanding anomalies we discover the general pattern is still irregular, a feature of the landscape. The diversity of interests, the multiplicity of cultural resources in one virtual place, is revealed in the texture of the data. We apply our normal methods of differentiation and classification, we reduce 150 million log records to ten million instances of what came next to the screen when a link was followed, but an overall pattern eludes us.



There is not a single European digital library, there is a diversity of creatures in this virtual laboratory, and here lies the major challenge for evaluator and policy maker.

Context

This section offers some context for our statistical analysis and frames how the data should be understood. We present an explanation of terms, methods of data processing and our level of confidence in the analyses we present; given that there are inevitably issues beyond our control and we want to be as transparent as possible.

Europeana

> "*a multilingual point of access, a network and a channel for digital content distribution.*"

Europeana —the european digital library—originated with a 2005 proposal supported by six european heads of state (France, Poland, Germany, Italy, Spain, and Hungary): the Digital Libraries Initiative. It is a project to "to make all Europe's cultural resources and scientific records —books, journals, films, maps, photographs, music, etc.—accessible to all, and preserve it for future generations". Europeana is conceived as a single access point for all these digital materials: the wandering scholar no longer has to travel the length and breadth of Europe seeking the original, digital copies are accessible online. It is also intended to provide stimulus to a 'digital economy', content creation and to 'democratise access to culture and knowledge'.

The europeana.eu website was launched in November 2008 as a "multimedia online library". Analysis of the server log-files is part of the Europeana Connect project which commenced in May 2009. After an initial assessment of sample files in the summer of 2009 arrangements were made to transfer the server logs on a daily basis to the research team at UCL Department of Information Studies. This automated transfer of the complete files has been in operation since October 2009. Thus now, in September 2010, we are able to present a report covering a full twelve months of stable operation of the europeana.eu web-site. A full-year's data enables some inferences of seasonal patterns to made. However during the spring and summer of 2010 a major upgrade of the site, the 'Rhine release', required some recalibration of our log-file analysis to account for new and remodelled features of the site. It is probable that the Rhine release process itself contributed to significant use of the site by developers and other project partners and thus the pattern of use over the year may not be a reliable indication of normal stable usage.

Data Processing

A full year of daily server log-files, 7GB of compressed data, records of over 150 million requests to the server. Much of this detail is of interest only to the system administrator, it records the performance of the server not matters of value to the provider or user. On the other hand such a record is neutral, untainted by predetermination of what should be worth recording.

The process of reduction and refinement that leads to the tables and charts in this report begins with an expansion: each request in the log is parsed and expanded to form a record with as many as several hundred attributes. Not all attributes are present in every record but overall we can identify several thousand attributes. The 12-month log-file thus appears as a very large spreadsheet table: 150 million rows and more than four-thousand columns. The aim of data-mining is not only to summarise these records in convenient tables, it must also find the hidden patterns and connections, cell to cell, within the whole table.

Page-view

We analyse logs from a user perspective; the fundamental unit is the '*Page View*': what new display results from clicking on a link or typing in a URL. By new display we mean a complete page refresh: thus changes to the display such as pop-ups on mouse-over or the suggestions displayed when typing in a search box are not considered a new page.

For Europeana.eu a canonical sequence of page views would be: the Home page, a search result displayed as a set of thumbnail images, a detailed record, and a 'click through' to a provider site. This last item opens a new window on another site; strictly speaking this is not therefore a 'Page View' of the europeana.eu site (and would not count as such for advertiser oriented 'page impressions' counters such as Google Analytics), however we are able to record these and they are included in our Page View counts as 'redirect'. Additional analysis of 'shownAt' (the link text 'View in original context') and 'shownBy' (the link on the main image on the record page) is used to discover the popularity of providers and content.

For every 'Page View' a large number of requests are sent to server and logged, these are filtered and combined to create a single record of each page view. The residue from this process we classify as 'page furniture'. We keep a record of this for book-keeping purposes, to have a complete account of all the log records we process. This is a technical measure, for the purposes of a user-centric log analysis the 'page furniture' is of no importance: all the important data in consolidated in the record of a Page View.

A similar consideration applies to 'Errors': if a request to the server is invalid or fails a log record is created: we keep account of these but they are not relevant to a user-centric analysis.

Thus the basic accounting is that over 12 months the europeana logs consist of some 150 million records, of which 15% count as errors and 40% page furniture. Our initial processing results in some 67 million 'Page View' records.

Robots

We analyse logs from a user perspective; we want to understand the behaviour and patterns of use of real, human users: the greater part of the Page Views we record originate from various automated agents, we need to discount these. The most obvious, the most numerous of these is *Googlebot*.

A common referral path to Europeana is via a Google search. Most are searches on variants of the domain name 'europeana': this is using Google in place of bookmarks or remembering the web address. Having the site indexed by search engines is important, but, for all its volume in the log record, it is secondary activity; we set it aside to pay attention to what really matters: real people using europeana as a multilingual point of access to networked digital content.

Identifying Googlebot is easy, it declares itself to be the UserAgent in every request sent to the server, so we can say with certainty that 50% of all page views are by the Googlebot. Though much smaller in volume the same certainty attaches to the identification of other search engine robots: *Yahoo* (11%), *MSNbot* (11%), *Yandex* (0.9%) *Baidu* (0.1%). Generally, we apply the rule that if the text sent to the server declaring the UserAgent says 'spider', 'bot', 'crawl', 'validator' or 'robot' then it is what it claims to be: an automated program retrieving web pages for machine processing, and therefore not representing the behaviour of a person interactively requesting a 'page view' from the server and performing some cognitive and physical action at each request.

Our quest for the underlying 'real user' is assisted by the converse of robot identification; ordinary people use plain everyday web browsers: *Internet Explorer* or *Firefox*, which declare themselves as such in the UserAgent text. Common search engines and browsers account for over 95% of all our page views. Once we account for less common browsers: *Safari*, *Opera*, and mobile devices such as *iPhone*, *iPad* etc. we have a firm identification for almost every form of user agent. A qualification needs to be introduced however, whilst we take a declaration to be a robot as *prima facie* true, claims to be an ordinary user with an ordinary browser cannot be entirely trusted. Additional tests are applied which have the effect of shifting 0.5% to the robot category. Overall the Page View category breaks down to around 85% robot, 15% user.

Outliers

We analyse logs from a user perspective; that implies a real live person viewing a web page, thinking about its content, following the next link. A linked chain of thought and page views. In the case of Europeana, our 12 month log of 150 million server requests has been processed to reveal around 10 million Page Views. Those ten-million page views should tell us something about who these people are, how they use Europeana, to what result. Can that result be considered a success for the user, the Europeana project, the content providers? Except that at this point we encounter a problem that is a notable peculiarity of Europeana: the 'real users' are not homogeneous, and some behave so unlike 'real users' that we suspect they may be robots.

This is best defined by example. In May 2010 it appears at first sight that there were 970,000 page views in that one month. This is much higher than the previous month but could be plausible. Until the more detailed pattern is examined: there are 479,000 pages viewed using the Opera browser; the normal figure for Opera usage is less than 2.5%. Of those, 476,000 can be traced to a single IP address (effectively a single user) on the CNR-PISA network. All these were requests for a full-doc (record) page, never the home page or a search result (brief-doc), all occurring in a period of less than 24 hours. Clearly this is not plausible as the behaviour of our putative 'real user'. This is an extreme example and it would be possible to refine our definition of robot to place it it that category. But taking into account factors such as frequency of use and unusual page transitions creates a complex definition of robot, that is to hard maintain with consistency.

We prefer to apply an 'outlier' classification to such extraordinary cases. These are special cases where it is possible to clearly and precisely identify a non-user or pseudo-robot. It is an incident, confined to a short period and localised. Such cases occur rarely, no more than one in each monthly dataset.

Real Users

We analyse logs from a user perspective; real live people doing ordinary things with an internet browser. Users who account for approximately 9 million page views in 12 months.

The Europeana users are not homogeneous, even having removed robots and pseudo-robots. Our page view data is characterised by its anomalies, odd peaks of activity, unlike outliers not always traceable to particular internet addresses, user agents or server requests. We are looking at a minestrone of activity records, a composite of several patterns. It is difficult to discern clear trends overall, we need to define sub-categories of users. This is a work-in-progress: so far we can see some trends relating Providers and Collections to the location and language of users, there are seasonal and daily patterns of use that also vary by location. We can also define large institutional users, and networks of predominantly educational use.

Confidence Ratings

We are concerned with the quality of data and analyses that it generates. Sometimes, factors beyond our control mean that the highest absolute levels of quality cannot be met. This may be due to issues relating to the quality of metadata or third party data (certainly an issue in the case of Europeana), problems over definition, or other technical or computational issues. For this reason, we offer confidence ratings in many of our reports. The ratings used in this report to support each table and graphic are given below.

*** Evidence that meets the highest standards of rigour and transparency and in which we have complete confidence.

** Data that we believes to be highly representative but which fails to meet the very highest standards of statistical validity. These are still pretty robust.

* Evidence that for reasons of sampling, metadata or other technical issues is believed to be broadly indicative. This can be regarded as a safe bet until more complete data becomes available.

Summary of key findings

1. Analysis of usage by hour of day, taking into account time zones, reveals significant differences in the daily pattern of consumption even between different countries within the EU. [Fig 2]
2. At the daily level, Europeana usage is characterized by enormous volatility that makes analysis of the kind offered by Google Analytics open to misinterpretation. [Fig 6]
3. From the vantage point of a full year's data, usage of Europeana shows a strong seasonal rhythm of a kind very familiar from other studies. This could possibly be a function of a large user base in schools or colleges; other possibilities are that it represents bursts of developer activity ahead of deadlines, or press, conference and similar publicity events. [[Fig 6]
4. Overall growth in Europeana has been sluggish in the past year (compound growth equivalent to 0.9 per cent per annum). However, this one year's data is all we have; a more interesting result may be expected when we can compare year on year data under conditions of active marketing. [Fig 7]
5. At the national level, rates of growth vary by a wide margin. Poland is the fastest growing user country (compound growth of 1.5 per cent per annum) and along with France, a major user, provides the motor for much of the growth in usage in the past year. Some countries (e.g. Denmark and Belgium) are very sluggish and the reasons for this might be investigated. [Table 4-5 and Figure 8]
6. In absolute terms, France, Germany and Poland are Europeana's largest consumers within the EU-27. However, in terms of usage per million capita, Luxembourg emerges as the single most intense user country. [Table 5, Table 15]
7. Users tend to exhibit a marked preference for collections created or curated in their own countries. This is a general finding but one that is particularly notable in the cases of France and Poland. [Table 6-7, Figure 11]
8. Multimedia content is a spectacular success with consumers: they are more than ten times more likely to select video material when viewing thumbnails than could be accounted for by chance. Static images are of less interest than expected using a simple probabilistic model. [Table 9]
9. Mobile devices accounted for less than one per cent of Europeana page views in September 2010, but there are signs of rapid growth in this form of access. [Fig 12]
10. Europeana appears to generate a very high volumes of searches but many are in fact pre-formed and embedded in static links to sign-posted content and exhibitions. Very few people use advanced search. [Fig 13]

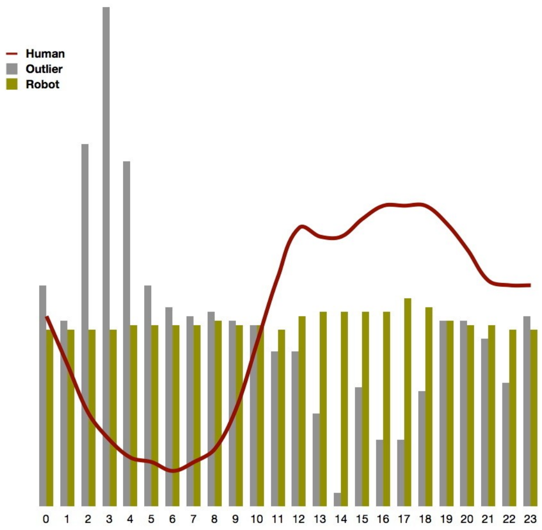
Patterns of use

With one year of log data we can be reasonably confident that the patterns we see are stable over time. However usage is quite low and it remains to be seen what patterns may emerge as year-on-year comparisons become available over the next twelve months.

Hourly patterns

Figure 1 shows the distribution of usage over 24 hours This does not take account of time zone differences, however, as the majority of usage is European the overall pattern reflects that of a time zone within a hour of UTC+01. Real use follows a pattern one might expect: rapid growth during the morning, peaking initially at 11am and then building up and reaching a plateau in the evening.

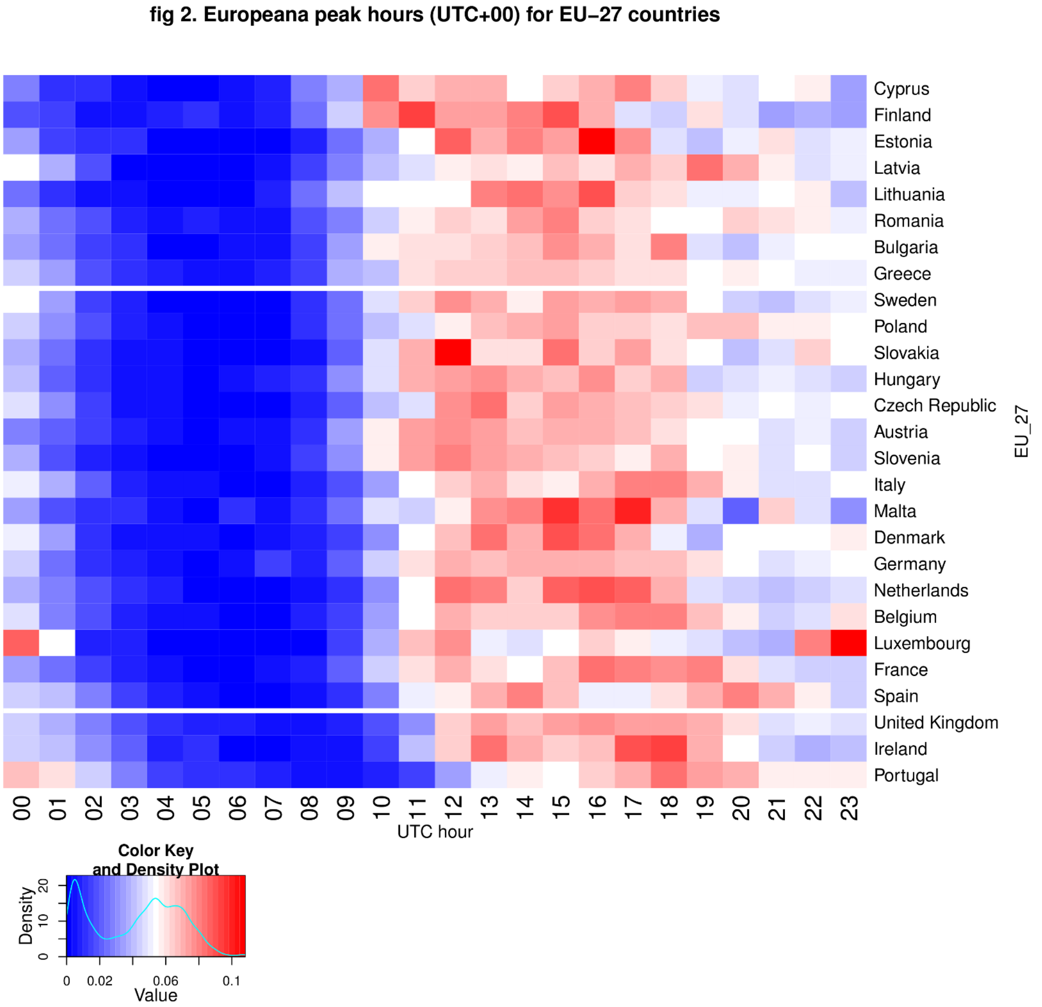
Figure 1: Hourly pattern of use; whole world***



October 2009 to September 2010, user page views, (normalised to UTC time)

An alternative visualisation of Europeana use over a 24-hour period is shown in Figure 2, a heat-map for the 27 members of the European Union. For each country a row shows the daily usage profile: each hour as a percentage of the whole day. Using a scale in which dark blue represents the lowest and red the highest values we contrast night and day. The times shown are normalised to UTC+00 but by arranging the countries is a sequence that reflects both differences in time-zone (Cyprus UTC+02 to Portugal UTC+00) and location, East-West and North-South, differences other than timezone begin to emerge.

Figure 2: Hourly page-views for EU-27 Oct 2009–Sep 2010***



There are national differences in this profile, even when the drift rightward as we work down the time zone shifts is taken into account. People in Cyprus and Portugal clearly have very different information seeking rhythms. Usage in Cyprus shows peaks in the morning and afternoon but is very low by 9pm local time (19:00 UTC+00). By contrast, Portuguese usage begins in the afternoon and is maintained through the evening.

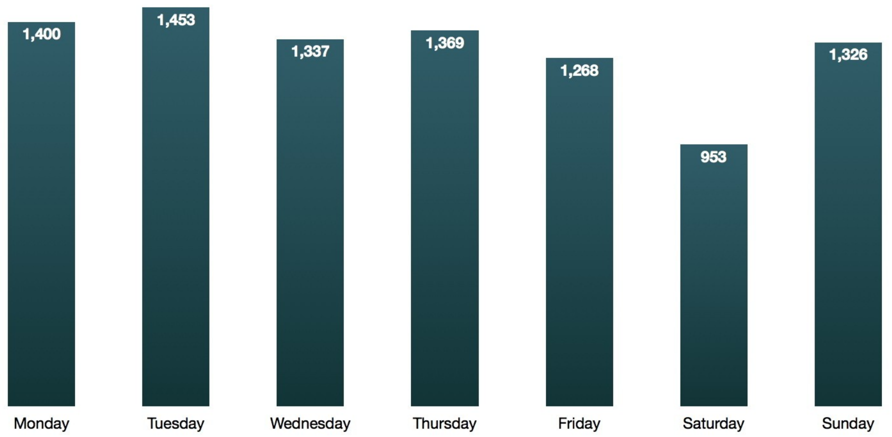
Weekly patterns

The daily distribution over the average week shows clearly that Europeana is least used on Saturdays. The level on Saturday is less than two-thirds of the weekday peak on Tuesdays. By contrast Sunday is not significantly different from any working day of the week and might indicate a higher level of home or leisure use when compared with patterns typical of academic journals.

Table 1: Weekly pattern of use; whole world ***

Week	Total	%	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Users	9,107,879	13.5%	1,400,314	1,453,420	1,336,696	1,369,414	1,268,152	953,468	1,326,415
Outliers	945,651	1.4%	158,023	573,679	6,165	4,383	155,907	38,453	9,041
Robots	57,209,587	85.1%	7,694,221	8,233,703	8,183,225	8,247,581	8,722,117	8,127,750	8,000,990
Total	67,263,117		9,252,558	10,260,802	9,526,086	9,621,378	10,146,176	9,119,671	9,336,446
Users			**15.4%**	**16.0%**	**14.7%**	**15.0%**	**13.9%**	**10.5%**	**14.6%**

Figure 3. Weekly pattern of use, whole world ***



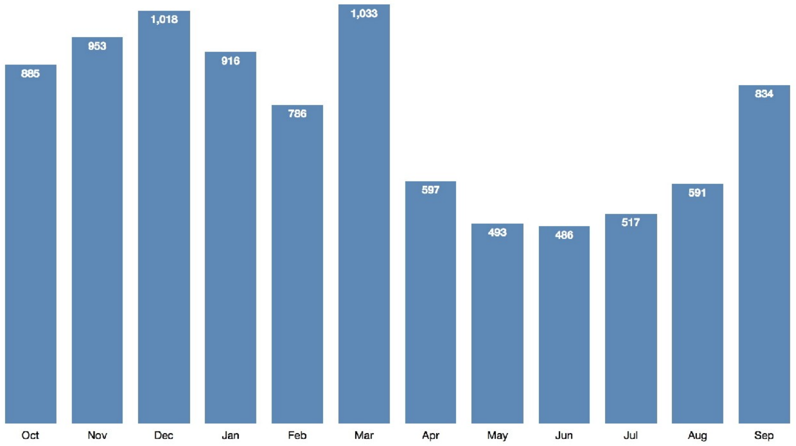
Monthly patterns

Looking at Europeana logs over a full 12-months, we begin to sense a seasonal rhythm and gain some early insight into the growth trends. Table 2 and Figure 4 offer a summary of month-by-month use with peaks in December and March and a lull over the summer holiday period, which suggest that Europeana does not at present appeal to the tourist.

Table 2: Monthly pattern of use; whole world ***

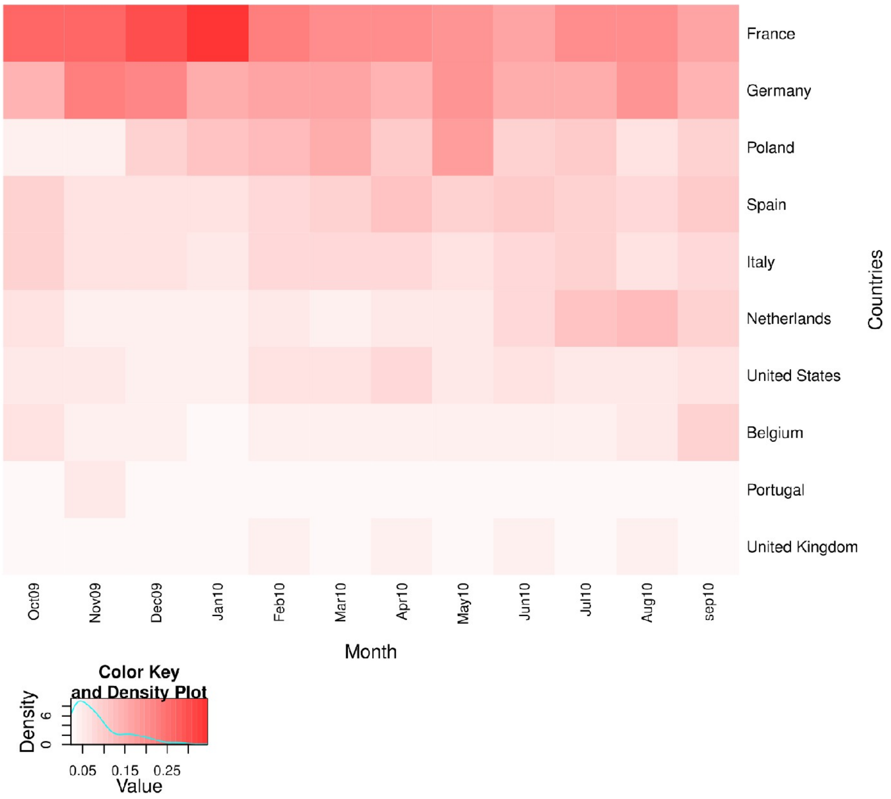
	2009			2010								
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Users	885	953	1018	916	786	1033	597	493	486	517	591	834
Outliers	94	0	0	209	0	37	0	479	0	0	4	29
Robots	2758	2954	3283	5481	8142	11673	9328	2543	1841	3752	2522	2933
Total	**3737**	**3907**	**4301**	**6606**	**8928**	**12743**	**9925**	**3515**	**2327**	**4269**	**3117**	**3796**
Users	9.7%	10.5%	11.2%	10.1%	8.6%	11.3%	6.6%	5.4%	5.3%	5.7%	6.5%	9.2%

Figure 4: Page views: 12 months, whole world ***



In Figure 5, a heat map shows monthly use for ten heaviest using countries. These ten counties account for 75% of all Europeana page views. The percentages, represented by the strength of the red tint, are calculated relative to the whole year in all countries, so we can clearly see how significant users from France are to Europeana. Distribution over the year is variable, but overall the peak months appear between November and May.

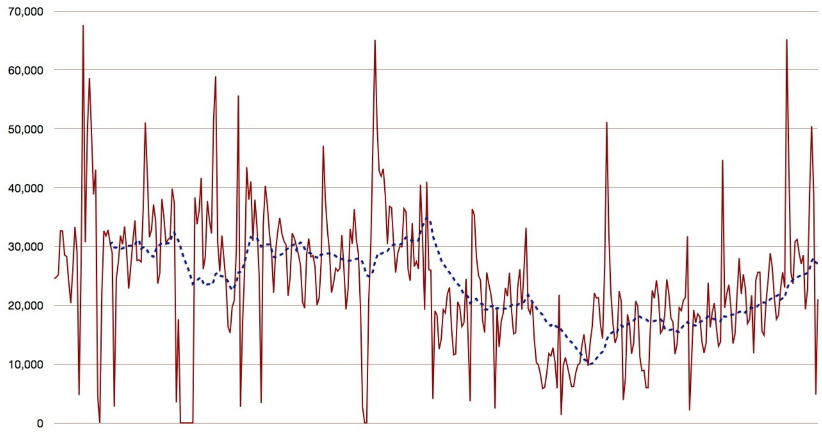
fig5. 75% of Europeana users are from 10 countries



Annual patterns

Figure 6 contrasts a seasonal trend with the volatility of day by day use. The seasonal trend of the dotted blue line is a 28-day moving average. A drop in the spring and a late summer recovery is apparent. But the daily range is very wide in relation to the average, and we are uncertain how to interpret the data at this level. Occasional bursts of activity are not unusual for any website and perpetual volatility would be considered normal for a news-based site whose popularity varies according to the sensation of the day, but for an educational and reference site a satisfactory account seems wanting.

Figure 6: Seasonal trends in use; whole world***



October 2009 to September 2010, user page views (red), 28-day moving average trend line (blue)

Some of the spikes could be accounted for by periods of intensive testing by Europeana developers (Table 3). For many of the developers, contributors and partners it is possible to track use based on the allocation of network addresses to the institution. The sample is small and biased toward large institutions with substantial networks. It does however suggest some atypical use during the summer months. A similar segmentation based on net-blocks highlights the activity of large institutions such as universities, schools and libraries.

Table 3: Use by Europeana Connect, and its contributors and partners**

October 2009 to September 2010, user page views (000s)

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Contributor	2.2	2.5	1.7	2.3	1.6	3.6	2.8	2.3	5.4	5.4	2.5	3.9
E-Conect	1.2	1.0	2.9	1.2	0.8	1.0	2.3	0.9	1.1	0.6	1.4	1.0
Partner	11.8	13.0	6.7	11.2	11.5	13.0	8.9	6.6	8.3	7.5	8.3	12.6
volume User	198.3	327.0	204.4	238.3	189.4	194.5	99.7	76.9	80.0	96.0	116.5	131.3

Figure 6a A seasonal pattern is evident for academic users

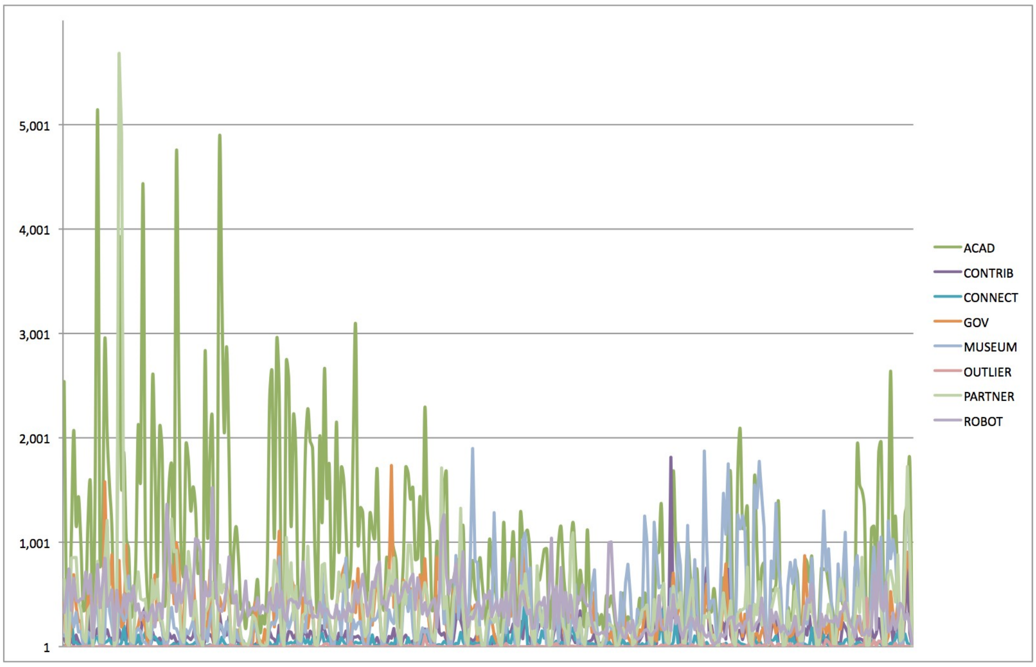
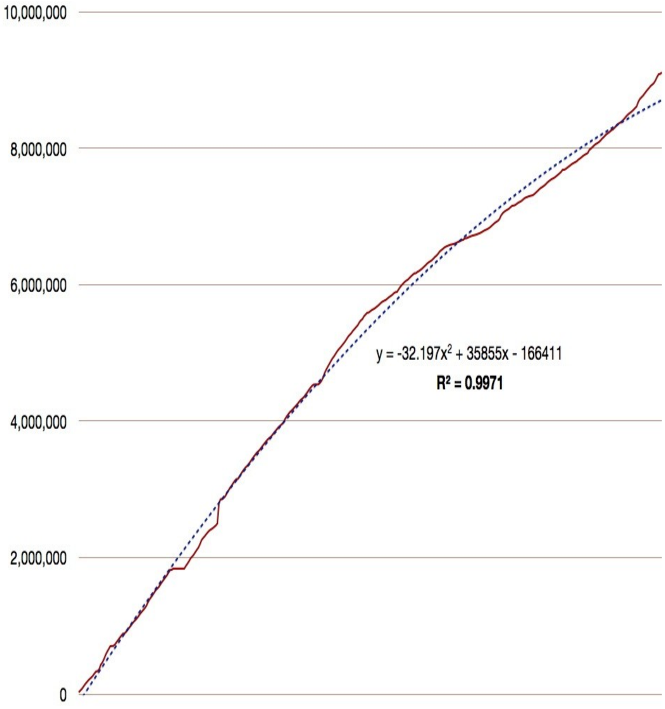


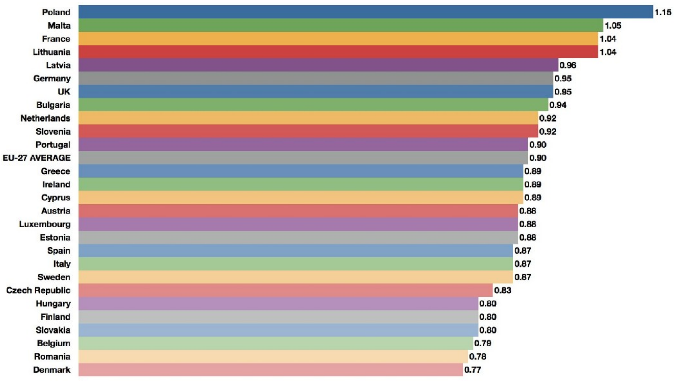
Figure 7: Cumulated annual use; whole world***

In Figure 7, we have accumulated each daily page views from September 2009 to October 2010 (the red line). At a first glance his appears to be a straight line—this would imply no growth—however, the line is not quite straight, looked at more closely we can see evidence of the seasonal patterns noted earlier: the summer lull for example. The best-fit trend line is not straight but a `third order polynomial' function: there is growth but not much. The shape of the red curve in Figure 7 suggests that overall usage of Europeana is equivalent to a compound annual growth rate (CAGR) of about 0.9 per cent per annum.



We have taken the data apart and estimated the individual compound growth rates for each of the EU-27 members (table 4). Given that we only have twelve months of data, some within a year of launch, these should be taken as being merely indicative of some early trends. Although the signals are rather weak, it appears that the main engines of growth in usage derive from users in Poland and France, two major users of Europeana. The overall picture though appears sluggish at this stage, especially at the bottom end of the table. The table is visualised in Figure 8.

Table 4 & Fig 8 : Comparative annual growth rates by EU-27 country*

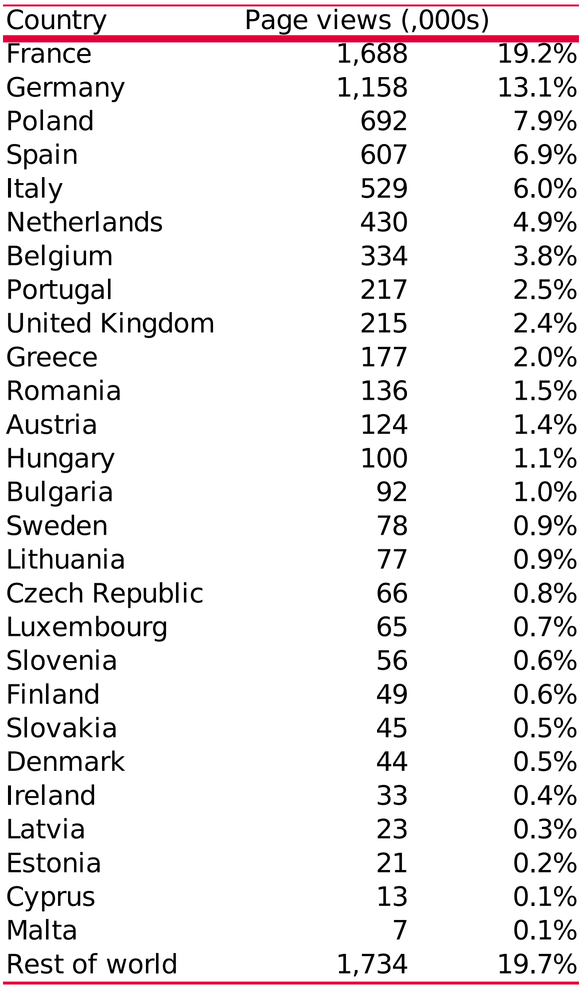


Country	rate
Poland	1.15
Malta	1.05
France	1.04
Lithuania	1.04
Latvia	0.96
Germany	0.95
UK	0.95
Bulgaria	0.94
Netherlands	0.92
Slovenia	0.92
Portugal	0.90
EU-27 average	0.90
Greece	0.89
Ireland	0.89
Cyprus	0.89
Austria	0.88
Luxembourg	0.88
Estonia	0.88
Spain	0.87
Italy	0.87
Sweden	0.87
Czech Republic	0.83
Hungary	0.80
Finland	0.80
Slovakia	0.80
Belgium	0.79
Romania	0.78
Denmark	0.77

Use by country

We can tell a user's location by their IP address. The two big EU-27 users are France and Germany, with a very strong showing from Poland (Table 5).

Table 5: Europeana use by EU-27 country **



Country	Page views (,000s)	
France	1,688	19.2%
Germany	1,158	13.1%
Poland	692	7.9%
Spain	607	6.9%
Italy	529	6.0%
Netherlands	430	4.9%
Belgium	334	3.8%
Portugal	217	2.5%
United Kingdom	215	2.4%
Greece	177	2.0%
Romania	136	1.5%
Austria	124	1.4%
Hungary	100	1.1%
Bulgaria	92	1.0%
Sweden	78	0.9%
Lithuania	77	0.9%
Czech Republic	66	0.8%
Luxembourg	65	0.7%
Slovenia	56	0.6%
Finland	49	0.6%
Slovakia	45	0.5%
Denmark	44	0.5%
Ireland	33	0.4%
Latvia	23	0.3%
Estonia	21	0.2%
Cyprus	13	0.1%
Malta	7	0.1%
Rest of world	1,734	19.7%

Figure 9 shows the distribution of Europeana usage across the globe taking no account of population density. Figure 10 maps the same dataset but shows page views relative to population and thus reveals Europeana's reach to be rather more Eurocentric.

Figure 9: Absolute use mapped across the whole world***

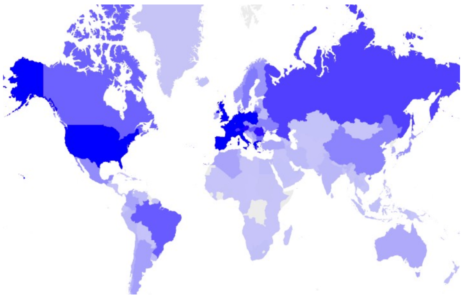
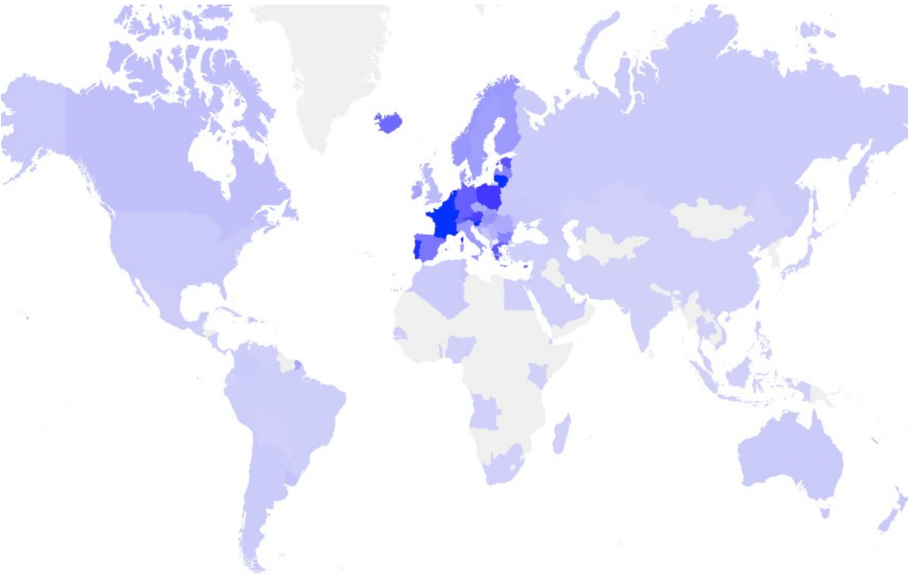


Figure 10: Per capita use mapped across whole world***



Since we can identify the locations of users,some closer geographic analysis is possible. For example (table 6), we can pose the question: `Do users tend to be attracted primarily to their own national collections, or do they range more widely across Europeana?'.

For this we need to identify not only the location of the user but also to attribute a 'nationality' to the collection; this is not always easy to decide but we believe ambiguous cases are not significant to the overall result. Also, we can only do so in cases where the page view can be attributed to a collection; hence, the calculation is based solely on views of the full-doc (record).

The table here presents a few of the most significant countries for both content curation (rows) and number of users viewing a full record (columns).

Table 6: National preferences for national collections; selected EU-27 countries**

	FR	DE	PL	IT	ES	GB	NL	BE
France	**231,225**	27,036	8,426	23,022	18,824	34,170	13,348	17,806
Germany	13,240	**83,387**	9,021	11,236	6,111	3,538	10,366	4,609
Poland	2,504	10,500	**78,383**	1,776	1,150	937	1,049	650
United Kingdom	11,370	13,624	4,717	5,982	6,043	**16,105**	8,685	3,742
Spain	2,210	2,061	878	2,358	**20,918**	580	3,644	2,006
Italy	2,779	1,884	1,036	**15,564**	2,672	542	1,157	769
Slovenia	1,435	2,516	1,516	1,791	905	289	801	577
Belgium	1,389	552	79	316	184	245	1,265	**12,212**
Austria	845	3,284	762	730	648	399	1,269	599
Romania	1,323	1,522	361	652	581	242	678	394
Netherlands	655	1,293	429	1,138	927	516	**1,113**	737

The pattern is more readily appreciated when Table 6 is presented as a heat-map (fig 11). In this format we display all 27 EU countries. For fig 11a (red tint) the values are percentages calculated by column, this emphasises the curatorial home of the collection. Heavy use of collections from France, Germany and UK is clear.

fig 11a. Europeana Collections and their markets

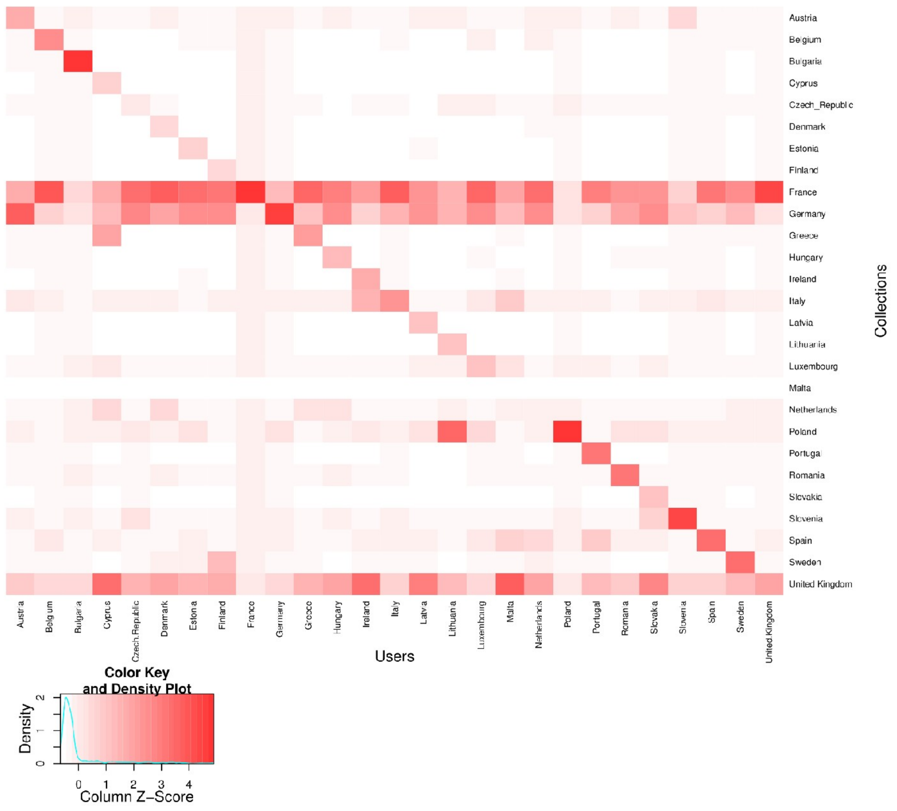
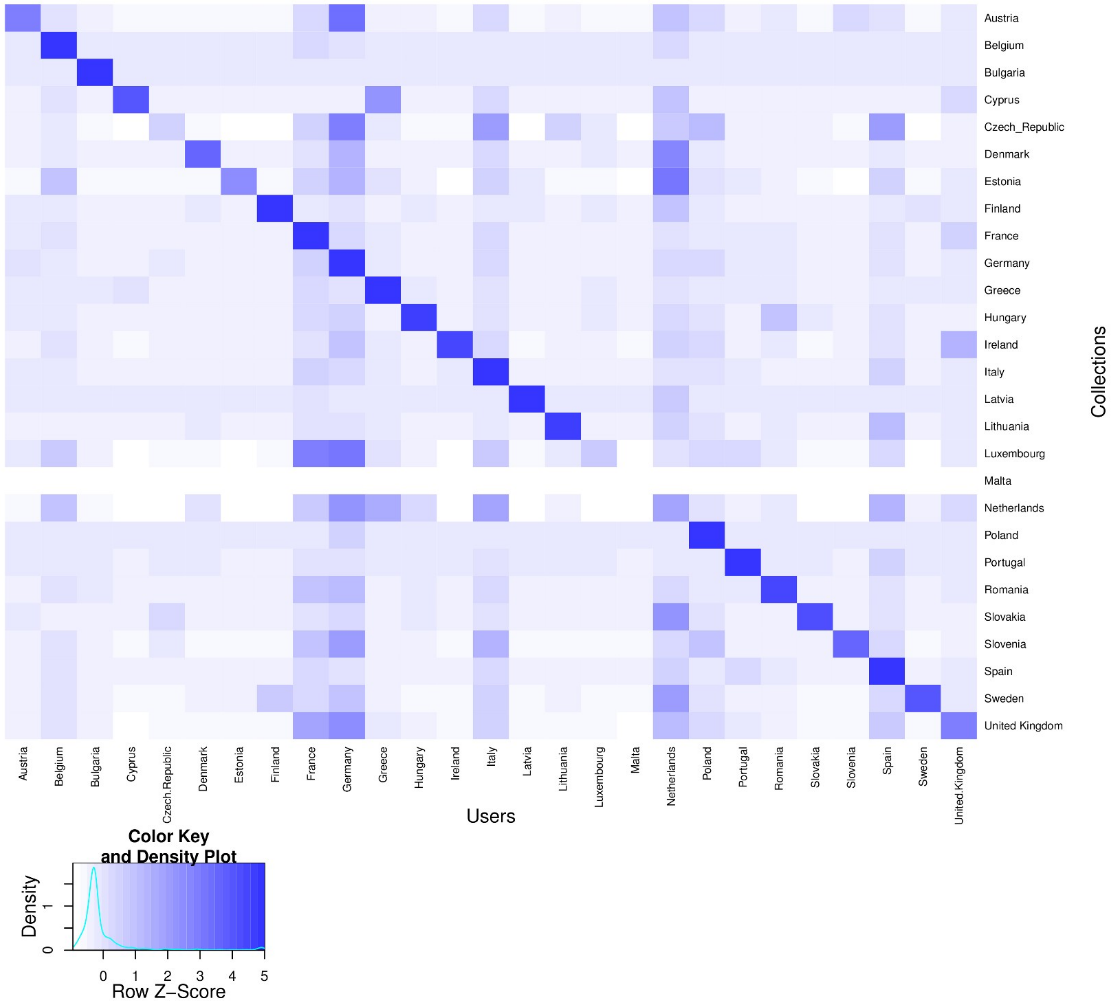


Fig 11b (blue tint) is the same dataset but showing a percentage by row; the emphasis here is on the location of the user. The heavy commitment to Europeana by French users is revealed by the vertical banding. But the strongest signal, visible in both versions, is the diagonal step: a strong national interest in national collections is clear to see.

fig 11b. Europeana Users and national collections



In Table 7 we construct an `insularity rating' for each of the EU-27 members. This index (which may be more familiar to economists as the Herfindahl index) ranges from 0 to 1 and it shows the degree to which users in that country concentrate their attention on national collections. A value of 1 would be obtained in the highly unlikely scenario that Finnish users only ever viewed material curated by Finns. Similarly unlikely, a value that was near zero, would mean that users ranged equally across all national collections with no bias whatsoever. From this it would appear that users in Luxembourg and Austria view the widest range of content and those in France the least.

Table 7 Europeana insularity rating selected EU-27countries

Country	Insularity rating
France	0.782
Poland	0.606
Bulgaria	0.45
Belgium	0.314
Spain	0.307
Germany	0.245
Italy	0.232
Greece	0.231
Portugal	0.227
UK	0.212
Romania	0.191
Sweden	0.189
Lithuania	0.183
Austria	0.166
Luxembourg	0.16

User preferences

In this section, we look at some more revealed consumer preferences.

Page type

Around twenty types of user pages can be found on Europeana. However, not surprisingly, most views are to Europeana's content: brief-doc (thumbnails) and full-doc (record), which display content within a standard frame. Table 8 shows the relative incidence of page types viewed.

Table 8: Most popular page types viewed; whole world***

Type of page	Page views	Column%
brief-doc	3787330	41.6
default_to_homepage	1441143	15.8
full-doc	1151275	12.6
Redirect	928926	10.2
Record	362370	4
Bob	243045	2.7
Index	180260	2
Login	170710	1.9
Aboutus	152332	1.7
Myeuropeana	125348	1.4
year-grid	106141	1.2
Communities	96226	1.1
Partners	80151	0.9
thought-lab	65053	0.7
new-content	41338	0.5
Register	36534	0.4
using-europeana	16909	0.4

Media format

When we focus on user preferences, as expressed by making a ?tab= selection on a Europeana thumbnail page (brief-doc) (Table 9) it becomes clear that users show a strong preference for multimedia content.

Table 9. User media format preferences; whole world **

Tab selection on brief-doc	odds-ratio
Images	0.69
Text	1.04
Sound	9.8
Video	10.7

The data in this table are `odds ratios'. We know, from Europeana metadata, how many records there are in each of the media formats above. We also know how many individual decisions were made at the level of the thumbnail click. The odds ratio expresses the likelihood that a user will select a particular format type. If users were viewing images, say, in exact proportion to the numbers in the system, the odds ratio would be 1. Higher than 1, and they are using images more than expected, less than 1, fewer times than expected. As can be seen, consumers are voting massively in favour of video and audio material rather than static images or text.

Most popular providers and collections

Tables 10 and 11 show content providers and collections that have so far proved the most popular with users across the world. This data is subject to some cautions: it only counts views of the full-doc/record page, there are biases introduced both by the featuring of content in pre-formated 'searches' and testing activity, and finally the process of identifying collection and provider from the log record requires refinement.

Table 10: Most popular content providers*

Provider	record views
Culture.fr/collections - Ministère de la Culture et de la Communication	268,467
Bibliothèque nationale de France	135,476
Biblioteca de Catalunya	119,446
Europeana Local Poland	91,774
Scran	76,515
Institut national de l’Audiovisuel	34,361
Europeana Local United Kingdom	29,376
Bayerische Staatsbibliothek	25,062
Biblioteka Narodowa	21,721
Cervantes Library	19,876
digiCULT-Verbund - DigiCult Museen SH, University of Kiel	18,713
Narodna in univerzitetna knjižnica	18,412
The British Library	16,754
Hispana	15,722
Fondazione Federico Zeri	15,419

Table 11: Most popular collections*

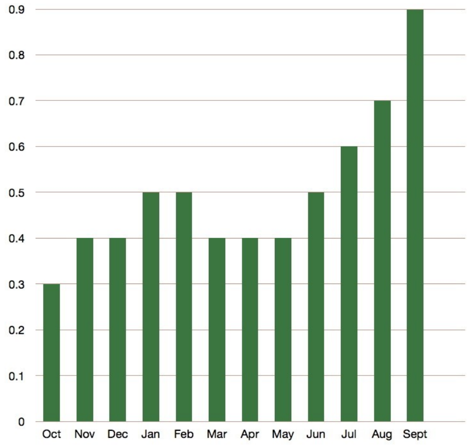
Collection	Provider	record views
RMN	Culture.fr/collections - Ministère de la Culture et de la Communication	135,259
Stadtgeschichtliche Museum Leipzig	Stadtgeschichtliche Museum Leipzig	95,476
Federacja Bibliotek Cyfrowych Poland	Europeana Local Poland	91,774
Gallica- Monographies	Bibliothèque nationale de France	83,970
Art and Design, Business and Computing, Craft, Design and Technology, Geography, History, Music, Science and Nature	Scran	76,515
Joconde	Culture.fr/collections - Ministère de la Culture et de la Communication	57,512
Public tv and radio programmes from 1920 to 2007	Institut national de l'Audiovisuel	34,361
Memoire SAP	Culture.fr/collections - Ministère de la Culture et de la Communication	24,658
Gallica- Images	Bibliothèque nationale de France	22,089
Digital Library Polona,	Biblioteka Narodowa	21,721

Browser and platform

The last few months have seen an explosion of interest in mobile access to the internet with the launch of the iPad. By the end of September 2010, mobile agents accounted for just under one per cent of Europeana page views (around 820,000), but a monthly analysis reveals this was more than three times the level at the beginning of the period. This may be a strong area for future growth, although incidental factors such as user testing of new iPads may mean that there is a high level of use within the Europeana project.

Figure 12: Growth of mobile access; whole world**

October 2009 to September 2010, mobile page views as a % of all page views



Search and navigation

The underlying technology of Europeana is that of a search engine and portal (although this not obvious to the first- time visitor). Its front page, with a very prominent search box, has obvious echoes of Google. There is also a hint of 'social media' interaction with 'People are currently thinking about'. But the practicality of Europeana, as currently implemented (Rhine release, autumn 2010) , is that every interaction generates a search. Browsing through a virtual exhibition such as 'Art Nouveau', amounts ultimately to a display of thumbnail images, an invitation to 'Click here to view object in Europeana'. An object in Europeana means in essence a library catalogue entry, a description, a small but larger-than-thumbnail image and an invitation to 'View in original context'. Original context leads to the opening of a new window on the site of the content provider; that may present a larger image, a more detailed catalogue and description, or present more of the same now dressed in the provider's livery.

Clearly then Europeana's performance as a search engine will be instrumental in determining its future success. Developers have provided numerous routes to multimedia content from the home page; indeed this page is crowded with various search, browse and navigating enticements all jostling for the user's attention. Thus at the very least we have:

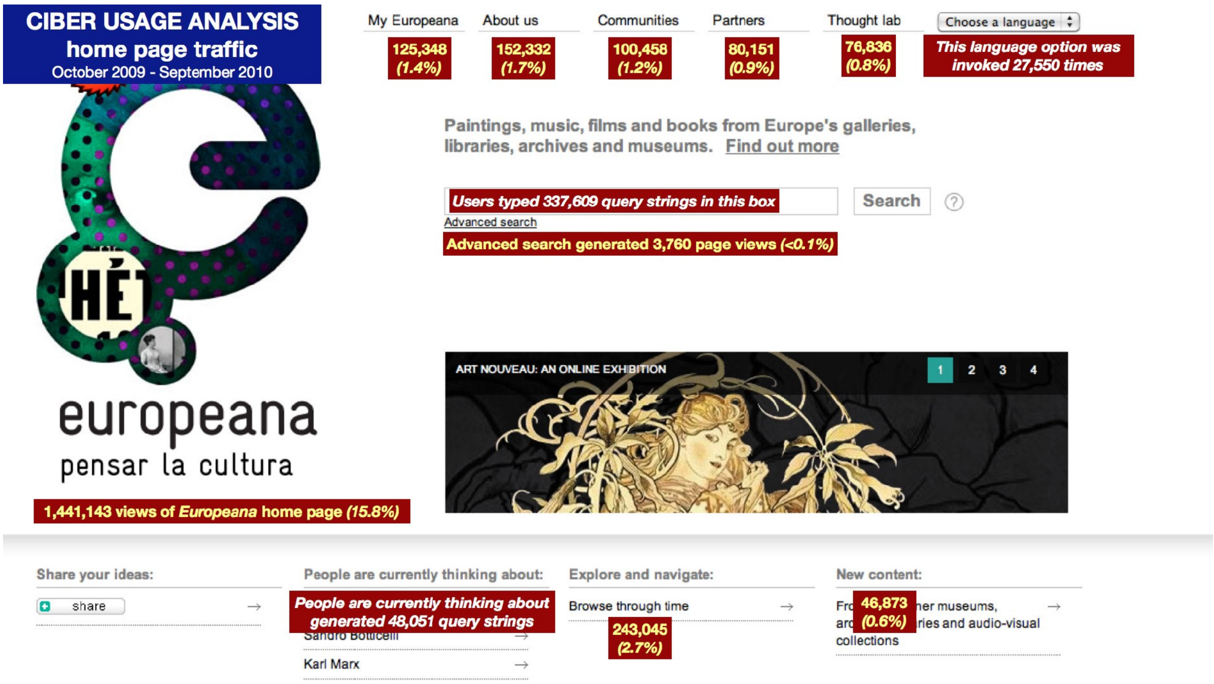
- Search
- Advanced search
- Animated invitations to exhibitions or other pre-selected content
- People are currently thinking about—a user feedback approach
- Explore and navigate—a browsing approach based on time
- New content—another invitation to pre-selected content

The main contest has to be between the Google-like big empty search box and the animated exhibitions box. A clear priority for any evaluator is then to establish how people search and navigate Europeana. The next three figures attempt to do this, although at this stage we are only offering an elementary analysis which maybe will generate discussion as to where it should go next.

The home page

It can be seen from Figure 13 that the search box is well used and that very few people use advanced search. The browse through time option has some significant use.

Figure 13: Analysis of home page traffic



CIBER USAGE ANALYSIS
home page traffic
October 2009 - September 2010

My Europeana 125,348 (1.4%)

About us 152,332 (1.7%)

Communities 100,458 (1.2%)

Partners 80,151 (0.9%)

Thought lab 76,836 (0.8%)

Choose a language: This language option was invoked 27,550 times

Paintings, music, films and books from Europe's galleries, libraries, archives and museums. Find out more

Users typed 337,609 query strings in this box

Advanced search generated 3,760 page views (<0.1%)

1,441,143 views of *Europeana* home page (15.8%)

People are currently thinking about: People are currently thinking about generated 48,051 query strings

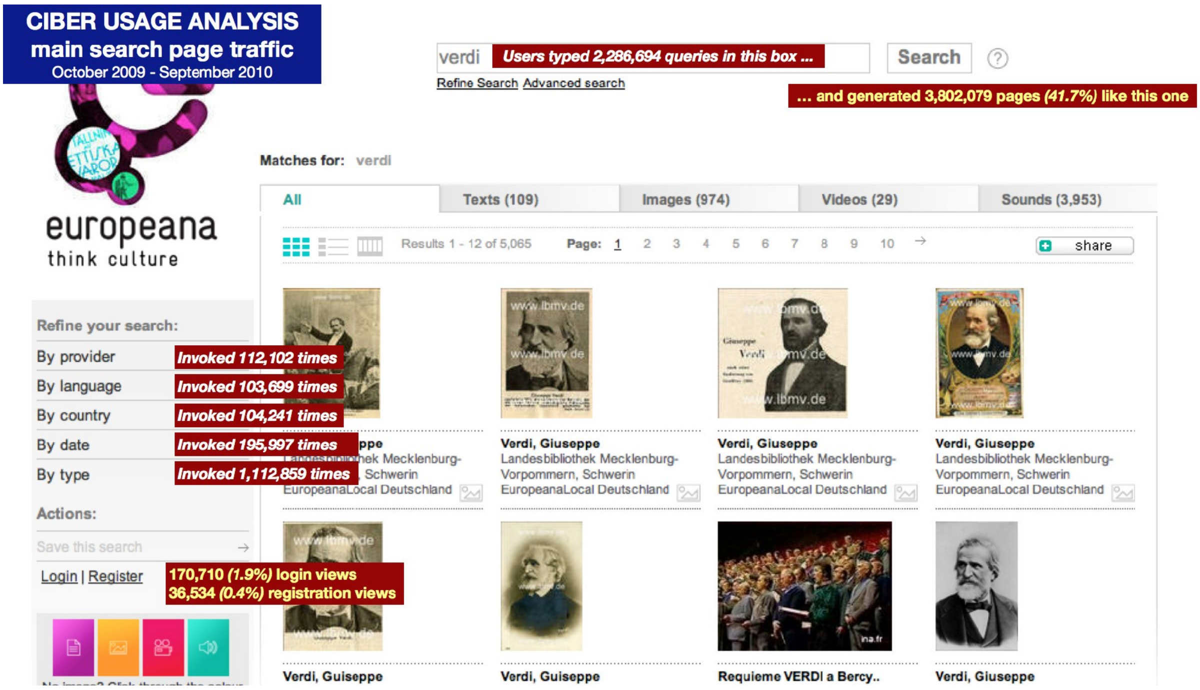
Explore and navigate: Browse through time 243,045 (2.7%)

New content: 46,873 (0.6%)

Thumbnail page (brief-doc)

The apparent popularity of the search box—huge volumes of queries—should be interpreted with caution; each time the page is redisplayed, whether to select a tab, refine a search, or to move on to the next set of twelve thumbnails, a search will be recorded. (Figure 14). In consequence it is difficult to separate genuine 'typed-in' searches from the myriad of actions that prompt a page display containing a search query string. (The Rhine release of summer 2010 introduced some changes which may improve this analysis in the future.)

Figure 14: Analysis of thumbnail page traffic



CIBER USAGE ANALYSIS
main search page traffic
October 2009 - September 2010

Users typed 2,286,694 queries in this box ...

... and generated 3,802,079 pages (*41.7%*) like this one

Invoked 112,102 times

Invoked 103,699 times

Invoked 104,241 times

Invoked 195,997 times

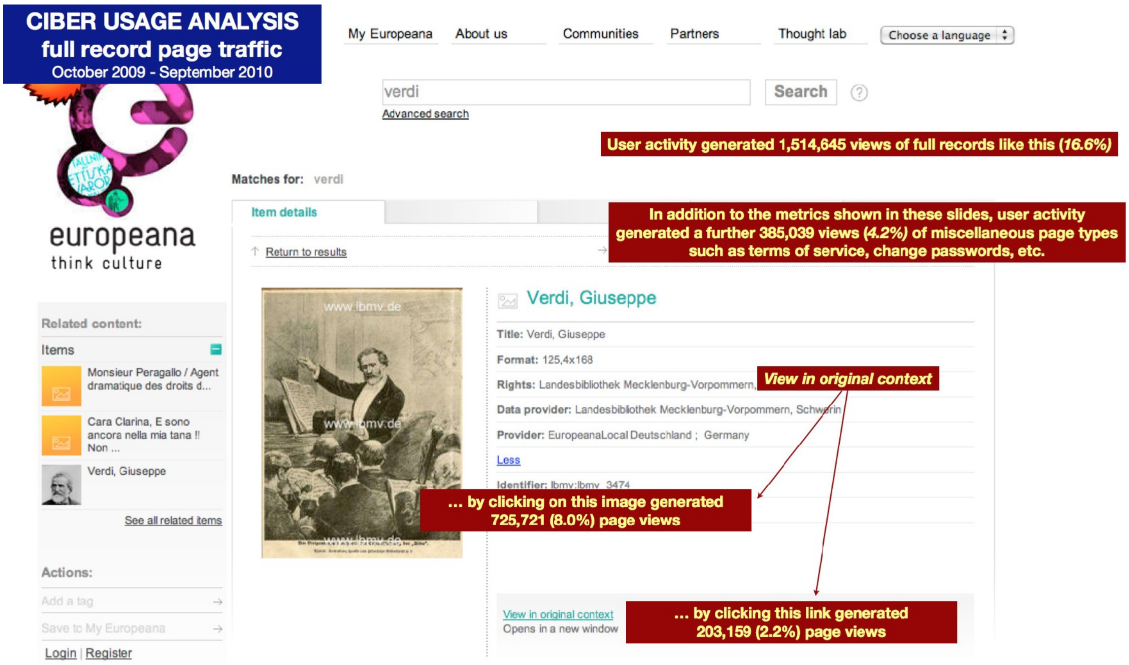
Invoked 1,112,859 times

170,710 (*1.9%*) login views
36,534 (*0.4%*) registration views

Main content page (full-doc/record)

Over a million and a half full-doc/records were downloaded.

Figure 15: Analysis of full-doc(record) page traffic



CIBER USAGE ANALYSIS
full record page traffic
October 2009 - September 2010

User activity generated 1,514,645 views of full records like this (*16.6%*)

In addition to the metrics shown in these slides, user activity generated a further 385,039 views (*4.2%*) of miscellaneous page types such as terms of service, change passwords, etc.

Verdi, Giuseppe

Title: Verdi, Giuseppe
Format: 125,4x168
Rights: Landesbibliothek Mecklenburg-Vorpommern,
Data provider: Landesbibliothek Mecklenburg-Vorpommern, Schwerin
Provider: EuropeanaLocal Deutschland ; Germany
Identifier: lbmv:lbmv_3474

View in original context

... by clicking on this image generated 725,721 (8.0%) page views

... by clicking this link generated 203,159 (2.2%) page views

Search terms

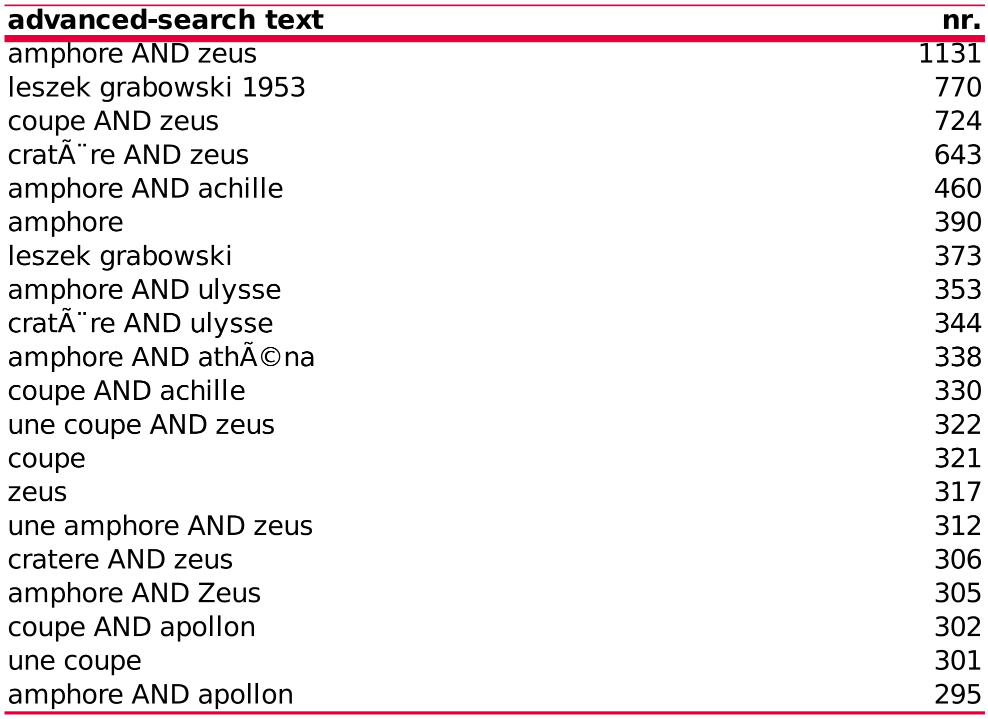
As mentioned above in relation to the brief-doc page, practically every page request on europeana.eu includes a search query string. It is difficult to separate genuine type-in queries from those that have been pre-selected in the link, or fully discount repetitions of the same search string through the multi-page displays of thumbnails. Thus in table 12 *most—probably all—of these top twenty search terms are 'canned' searches*.

Table 12: Most popular search terms from home page**

search text	nr.
postcard OR carte postale OR postkort OR Postkarte OR cartolina OR pocztówk	1836
:	1646
Wolfgang Amadeus Mozart	1429
map OR carte OR kaart OR karte OR mapa OR mappa OR kartta	1388
mozart	1034
picasso	694
goethe	629
mona lisa	623
da vinci	468
descartes	467
louvre	454
van gogh	443
chopin	439
freud	398
vase grec	390
leonardo da vinci	382
bible	379
rubens	378
dante	350
darwin	346

A different problem afflicts analysis of 'Advanced Search'; the feature has very few users, in addition we note that most likely these queries originated from an assisted search engine: *EMC Documentum Federated Search Services* (www.emc.com).

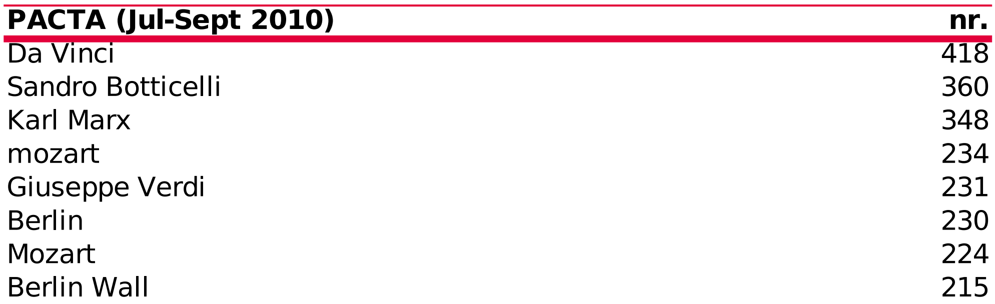
Table 13: Most popular search terms: advanced search; whole world**



advanced-search text	nr.
amphore AND zeus	1131
leszek grabowski 1953	770
coupe AND zeus	724
cratÃ¨re AND zeus	643
amphore AND achille	460
amphore	390
leszek grabowski	373
amphore AND ulysse	353
cratÃ¨re AND ulysse	344
amphore AND athÃ©na	338
coupe AND achille	330
une coupe AND zeus	322
coupe	321
zeus	317
une amphore AND zeus	312
cratere AND zeus	306
amphore AND Zeus	305
coupe AND apollon	302
une coupe	301
amphore AND apollon	295

The Rhine release of summer 2010 introduced some changes which may improve this analysis in the future. Firstly features such as 'browse through time' set a query field 'bt=pacta' which will enable us to more correctly identify 'canned' searches.

Fig 13b People are Currently Thinking About...



PACTA (Jul-Sept 2010)	nr.
Da Vinci	418
Sandro Botticelli	360
Karl Marx	348
mozart	234
Giuseppe Verdi	231
Berlin	230
Mozart	224
Berlin Wall	215

Secondly the Rhine release introduced search suggestions when queries are typed into the search box. This feature sends a series of queries to the server as the characters of the search string are typed, those requests are logged and may enable us to build a better analysis of genuine 'typed-in' searches.

Table 13c 'typed-in' searches

terms' query string = typed-in search
la cuestión de límitese las eficaces virtudes
la madonna del parto di piero della francesca
Landesbibliothek Mecklenburg-Vorpommern, Schw
les mystères de la franc maçonnerie par léo t
lettera del ministro baccelli istruire quanto
manifiesto sobre la construccion de las dos a
Mitteilungen der Reichsforschungsgesellschaft
Monarquia Religión triunfante de los sofismas
Muñoz "el idioma francés al alcance de todos"
Nobreza feminina da corte de Luiz XVI na Fran
Notitia Ecclesiastica Historiarum Conciliorum
obras de la pintora italiana del siglo xlx va
pintores y tapicistas del siglo xlx, italiano
pinturas do seculo XV e XVI da ilha damadeira
pliometrie et qualités fonctionnelle basketba
Por Quanto siendo tan repetidos los embarazos
portrait du generfext:de gaulle AND date:1945
psychopathologie de l'enfant et de l'adolesce
recherche portraits peints mignatures par ISA
Registre d'ordres du maréchal Berthier pendan
Relazioni Degli Ambasciatori Veneti Al Senato
repertoir des danses folklorique algerienneab
resolucion del rey conde de valdeparayso 1756
retable des7 douleurs de la vierge - DURER -

Most frequent referring sites

Table 14 shows the most popular referrer hostnames that are responsible for sending traffic direct to Europeana.

Where the referrer is 'europeana' we are logging movement within the europeana site: the second and subsequent pages of each unique visit. This accounts for some 60% of all page views; thus it would seem the average visitor may view three pages: a landing page from a referrer and two pages within Europeana. However there are, as already noted, many irregularities and outliers in our data. We should therefore be cautious in supposing that most visits follow a canonical path such as from home page, to brief-doc, to record. From table 8 we can see that the brief-doc page is by far the most used; which suggests that for the average three-page-visit two of those pages would be of thumbnail views. An analysis of typical paths through the sites is something we hope to address in more detail in a future report.

Google is by far the most popular referring site but we need to distinguish between Google pointing to europeana as a response to a query on a particular topic and an increasingly common use of Google as a substitute for typing a domain name: some two-thirds of all references from Google are in response to the search word 'europeanna' and many others may be considered linguistic variants, misspellings, or descriptive etc. There are many variants of google domain, eg www.google.fr, www.google.com etc. putting them all together we can see that Google accounts for around 10% of all referrals, no other domain accounts for more than 0.4%.

table 14a Referrer Names

Referrer Names	Page Views	% of Pages	% Visits
Total:15402 names	**9,080,735**	**100.0%**	
www.europeana.eu	4,737,565	52.2%	
europeana.eu	720,416	7.9%	
www.europeana.com	22,171	0.2%	
www.europeana.org	12,381	0.1%	
Total: 49 EUROPEANA domains	**5,509,472**	**60.7%**	
Implied visit (session) count	**3,571,263**	**39.3%**	**100.0%**
Pages with unnamed referrer	**1,114,874**	**12.3%**	**31.2%**
www.google.fr	72,846		2.0%
www.google.com	54,456		1.5%
www.google.de	46,068		1.3%
www.google.es	22,348		0.6%
www.google.pl	21,533		0.6%
www.google.it	18,509		0.5%
www.google.co.uk	13,878		0.4%
translate.googleusercontent.com	13,028		0.4%
Total: 344 GOOGLE domains	**374,890**	**4.1%**	**10.5%**
YAHOO domains	13,115	0.1%	0.4%
ulises-itaca.blogspot.com	11,962	0.1%	0.3%
www.bnf.fr	11,120	0.1%	0.3%
www.emob.fr	8,986	0.1%	0.3%
app.e2ma.net	5,837	0.1%	0.2%
doucetpiquante2.canalblog.com	5,553	0.1%	0.2%
ec.europa.eu	5,193	0.1%	0.1%
europa.eu	4,959	0.1%	0.1%
www.facebook.com	4,844	0.1%	0.1%

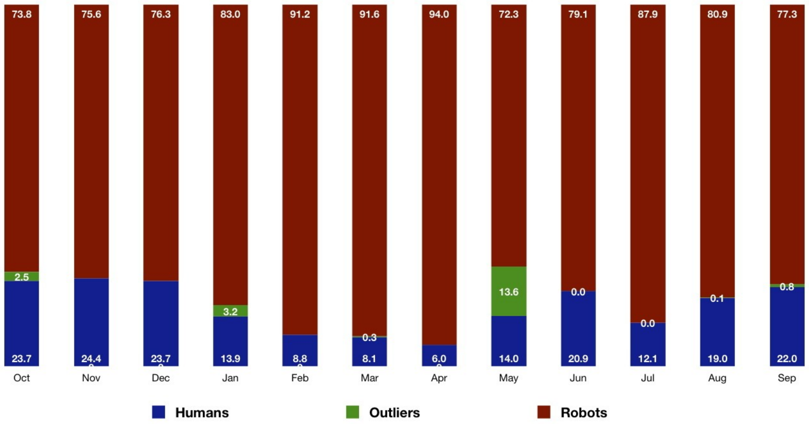
table 14b Google search terms

Google referrals	Page Views	%
Total: 66118 Google search strings	**340,689**	**100.0%**
*EUROPEANA*	230,251	67.6%
europÃ©ana	2,245	0.7%
bibliothÃ¨que numÃ©rique europÃ©enne	1,267	0.4%
europeanna	1,051	0.3%
bibliothÃ¨que europÃ©enne en ligne	544	0.2%
eurpeana	506	0.1%
europiana	502	0.1%
europena	493	0.1%
europana	372	0.1%
biblioteca online	366	0.1%
europÃ¤ische bibliothek	346	0.1%
biblioteca europea	343	0.1%
european library	299	0.1%
european digital library	287	0.1%

Robots

As noted above much of the log content is noise, in Figure 19, we can see that real use comprises only a small portion of all activity. Of course, the value of robots is fundamental; they make possible the discovery of huge amounts of information that would otherwise remain invisible to the user.

Figure 19: Monthly robot and human activity compared; whole world***

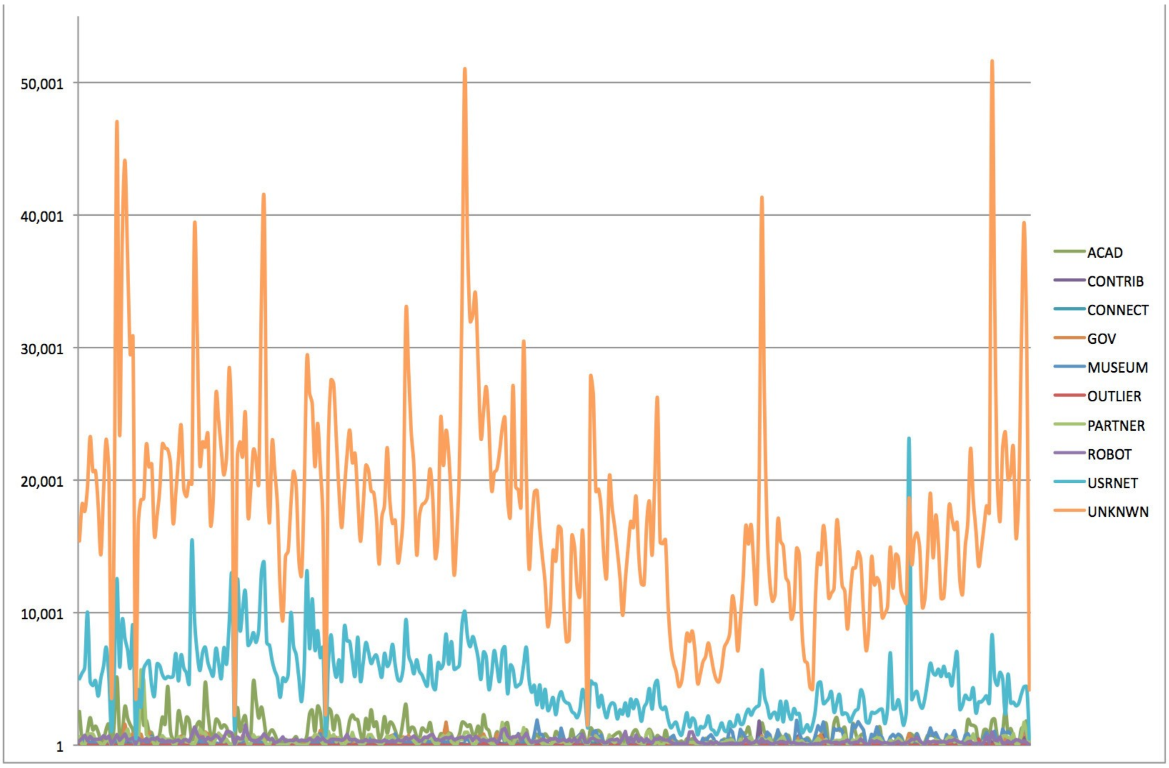


Future analysis of Europeana users and information seeking

- Produce more accurate data about *real* users. At the time of our last report in March 2010 we expressed a view that further refinement of our classification of users into types was required. In particular a need to differentiate between those associated with the development and content provision to Europeana. We also noted the presence of large institutional users, particularly those associated with French education. Subsequent work has shown that separating out these types of user still leaves many erratic features on the landscape. We will need to employ more advanced data mining techniques to find hidden patterns. There are clusters of attributes that identify what appear to be several communities of users.

-

more signal less noise



- Essentially Europeana is a search engine and we need to obtain a more detailed understanding of how people search and navigate Europeana. There is considerable potential in tracing how people move around the site, how much turning round and round in the brief-doc maze people do before either leaving the site or taking a route out to the provider.

- A refinement of our analysis of search terms is desirable. As noted above changes introduced with the Rhine release will assist this task.
- We are also working to produce a 'click-through' report, principally of use to providers, showing the most highly sought content as rated by the incidence of searches leading to a view of the provider's own content.
- Our analysis needs to take account of the anticipated growth in the use of mobile devices.The numbers currently observed are small, and the difficulty of separating genuine use from the many varieties of noise (see above) are considerable.

Annex: Global reach of Europeana

Table 15 shows the relative intensity of interest in Europeana by providing a simple ratio: page views per million head of population.

Table 15: Europeana: per million capita use; whole world

October 2009 to September 2010, page views per million population

Country	Page Views per million population	Rank
Luxembourg	131,816	1
Belgium	31,606	2
France	27,748	3
Slovenia	27,310	4
Netherlands	26,593	5
Lithuania	23,628	6
Portugal	20,828	7
Poland	18,275	8
Austria	17,686	9
Malta	17,445	10
Cyprus	16,470	11
Estonia	16,140	12
Greece	15,841	13
Andorra	15,540	14
Switzerland	15,411	15
Germany	15,233	16
Iceland	14,520	17
Spain	13,399	18
Bulgaria	12,207	19

Latvia	12,040	20
Hungary	10,216	21
Norway	9,330	22
Finland	9,207	23
Italy	9,031	24
Slovakia	8,600	25
Sweden	8,433	26
Denmark	8,060	27
Guadeloupe	7,748	28
Ireland	7,473	29
Croatia	7,047	30
French Polynesia	6,857	31
French Guiana	6,725	32
Romania	6,436	33
Czech Republic	6,377	34
Martinique	5,848	35
New Caledonia	5,240	36
Serbia	4,946	37
United Kingdom	3,540	38
Reunion	2,926	39
Canada	2,890	40
Moldova	2,615	41
Montenegro	2,447	42
Uruguay	2,379	43
Macau	2,336	44
Bosnia and Herzegovina	2,113	45

Dominican Republic	1,815	46
Israel	1,799	47
Albania	1,735	48
Tunisia	1,735	49
Taiwan	1,551	50
Ukraine	1,500	51
United States	1,350	52
Djibouti	1,291	53
New Zealand	1,275	54
United Arab Emirates	1,222	55
Australia	1,187	56
Chile	1,164	57
Belarus	1,137	58
Qatar	1,049	59
Morocco	1,036	60
Trinidad and Tobago	1,018	61
Argentina	948	62
Algeria	918	63
Hong Kong	870	64
Lebanon	839	65
Russian Federation	818	66
Mauritius	814	67
Colombia	756	68
Kuwait	664	69
Puerto Rico	576	70
Armenia	556	71

Mexico	554	72
Singapore	534	73
Brazil	522	74
Panama	448	75
Japan	436	76
Costa Rica	432	77
Turkey	421	78
Georgia	411	79
Peru	392	80
Jordan	277	81
Togo	255	82
Venezuela	244	83
Azerbaijan	229	84
Kazakhstan	213	85
Guatemala	201	86
El Salvador	191	87
Paraguay	186	88
Bolivia	167	89
Senegal	166	90
Ecuador	160	91
Iran	154	92
Nicaragua	152	93
Saudi Arabia	107	94
Malaysia	105	95
South Africa	101	96
Thailand	95	97

Egypt	93	98
Cote D'Ivoire	91	99
Cuba	86	100
Syria	73	101
China	51	102
Madagascar	49	103
Angola	43	104
Philippines	41	105
Vietnam	38	106
Afghanistan	30	107
Kenya	20	108
Indonesia	18	109
Pakistan	14	110
India	11	111
Nigeria	5	112