ECP-2008-DILI-528001

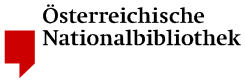
EuropeanaConnect

D3.1.3 – Publishable report on best practice and how users are using the Europeana service

Deliverable number/name	*D 3.1.3*
Dissemination level	*PUBLIC*
Delivery date	*2011-10-31*
Status	*1.0*
Author(s)	*DJ Clark, D Nicholas, I Rowlands*

e*Content*plus

This project is funded under the *e*Content*plus* programme,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.



EuropeanaConnect is coordinated by the Austrian National Library

europeana
connect

D3.1.3 Report on best-practice and how users are using the Europeana service

Europeana: an evaluation of users, usage, and information-seeking behaviour derived from the web-server log-files of europeana.eu (October 2009–October 2011)

co-funded by the European Union

The project is co-funded by the European Union, through the ***e*Content*plus*** programme

http://ec.europa.eu/econtentplus

Österreichische Nationalbibliothek

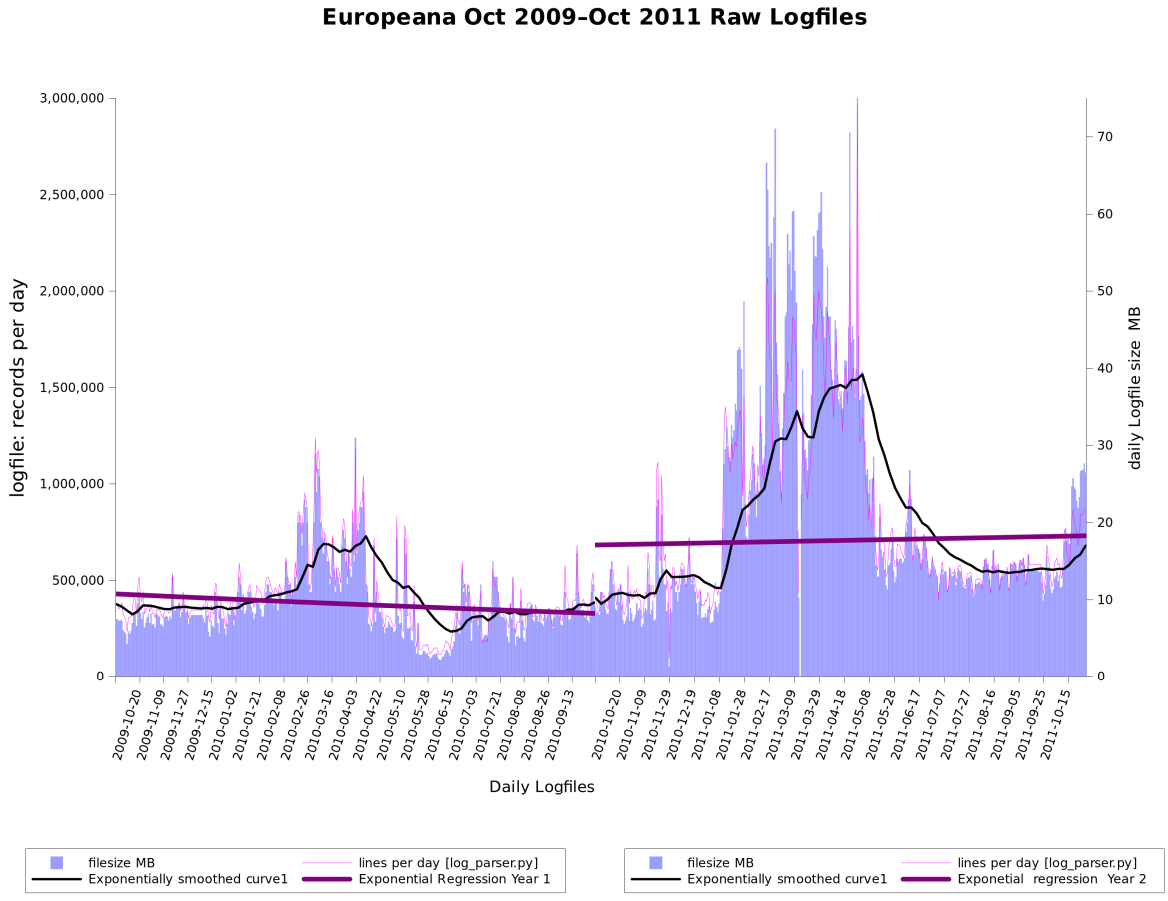
EuropeanaConnect is coordinated by the Austrian National Library

Table of Contents

Overview

Growth

The Europeana.eu website was launched in November 2008 but analysis of the log files as part of the EuropeanaConnect programme did not commence until the following year. This study is based on the continuous series of log data over 2 years from 3 October 2009 to late September 2011. The first of these years (Oct 2009–Sept 2010) was characterised by volatile patterns of use within an overall flat growth trend. The project was in an advanced development stage where it was difficult to identify genuine external use of Europeana.eu from the activity by the many development and content provider partners.



The second year of this study (Oct 2010–Sept2011) was witness to significant growth and the emergence of some notable patterns and trends. Changes to promote Search Engine Optimisation (SEO) which commenced with the Rhine release in Autumn 2010 can be credited with a phenomenal growth in site traffic during the first four months of 2011. This fell back in May 2011, though the level of activity over the summer has been both stable and two or three times greater the corresponding period in the previous year. At present (late-October 2011) there are signs that activity levels are beginning to rise again.

We believe that the observed pattern contains several components: an overall annual growth trend as a result of more active marketing; a step-change since January 2011 as SEO brought in many new referrals, in particular from Google; and a perennial pattern related to academic usage

which peaks in March of each year. In the academic pattern heavy use of the service by schools in France during the February–March period is very noticeable in both 2010 and 2011.

In addition, in April 2011 technical problems closed the site to indexing robots for a short period which coincided with changes to Googles page-ranking of aggregator sites. These are exceptional, one-off, perturbations whose overall contribution to the pattern is hard to assess.

In autumn 2011 the prospect is of a fourfold increase in users when the full year 2011 is compared with 2010 and a twofold increase in page-views. For the future we expect user numbers to rise from over three million in 2011 to five and a half million in 2012.

Mobiles

Three years ago Europeana.eu was prescient in considering the mobile user in its development plans. But since then the iPad has upset the easy definition of 'mobile' user. Thus, from October 2011, users of 'tablet' devices such as the iPad will no longer be presented with a cut down version of Europeana.eu intended for mobile 'phones— a small screen and restricted data transfer. During the period of this study we have observed both a steady growth in the use of mobile internet use, and the transformational introduction of the iPad in 2010 which succeeded where previous attempts to promote tablet devices had foundered. In October 2009 mobile use was an insignificant 0.3% of all page views,, 0.7% of users, and we believe that internal system testing may have accounted for much of that. Now, the most recent verified data shows a fourfold growth of users and five times as many pages from a diverse user base.

Mobile use is qualitatively different from desktop use, we see more short 'one-shot' visitors, more use in the evening and at weekends, but surprisingly there appear to be fewer referrals from blogs and social media.

We expect to see the number of users accessing Europeana.eu from 'phone and tablet devices rise from 3% at present to 17% by the close of 2012.

Users

We can identify clusters of users: 'bouncers' who view only a single page; 'checkers' who perform a single search operation and spend less that two minutes on the site, and a small minority (6%) of 'explorers' who have relatively long visits of ten minutes or more and may view many pages. About 'Bouncers'—more than half of all users— a single page-view says little. We do know that SEO vastly increased their number, it is far harder to say if the visits were satisfactory from the user point of view. For those users who do pause to view a few pages there is potential for a deeper cluster analysis. In particular to better identify institutional users, probably multiple users of kiosk-mode browsers, and characterise 'the ones that get away': users who are referred by sites other than search engines, do browse around the site, but neither revisit nor follow through to visit a provider site. Audio-Visual material has a high appeal, users are ten times more likely to select video material when viewing thumbnails than could be accounted for by chance, whereas text is less popular.

We have noted above the heavy use of Europeana.eu by French schools in the first four months of the year. Poland is another country that logs significant use particularly from libraries. More significantly, users tend to exhibit a marked preference for collections created or curated in their own countries. This applies to all EU countries but is particularly notable in the cases of France and Poland

Blogging sites are beginning to make a very significant impact, accounting for around one referred visit in ten between January and April 2011, up tenfold from the equivalent period in 2010.

About this report

This report is presented is several parts; the document before you forms the main narrative report with a detailed description of methods and results. “Culture on the Go”, first presented at the EuropeanaTech conference (Wien, October 2011) presents our main findings in a more populist format with particular emphasis on the implications for Europeana of the rapid growth of internet access from mobile and tablet devices. Our analysis regenerates a dataset of over700 tables each month; a selection of the most significant are included in a separate document. Likewise the maps and charts presented in small-scale in this document are also made available separately.

Introduction

An evaluation needs to consider, who are the users, and for whom is value 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?

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

In our first-year report on the Europeana Prototype (M3.1.4) we noted the high 'noise levels' in the log data; the difficulty of pointing to reliable and useful conclusions when significant levels of logged activity may represent internal use by developers and irregular bursts of testing by partners. The situation changed In the latter half of 2010 when major upgrade of the site ('the Rhine release') and a programme of search-engine optimisation (SEO) preceeded a fivefold increase in traffic to the site in the first four months of 2011.

Now, as the EuropeanConnect programme nears completion in October 2011, we are in a position to evaluate two full years of logged user activity. The unfeasibly exponential growth of early 2011 fell back in May and remained flat over the summer, the result of several factors: the rhythm of the academic year, changes in Google's page-ranking of aggregator sites, a temporary closure of the site to indexing robots. It is difficult to assess the relative weight due to these factors but in October 2011 activity appears to be on the rise; the coming year will prove how big a part the seasonal factor plays. In autumn 2011 the prospect is of a fourfold increase in users when the full year 2011 is compared with 2010 and a twofold increase in page-views. For the future we expect user numbers to rise from over three million in 2011 to five and a half million in 2012.

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 CIBER-research.eu. This automated transfer of the complete files has been in operation since October 2009. Thus now, in October 2011, we are able to present a report covering 2 years of stable operation of the europeana.eu web-site. One-year's data sets out a template, two years suggests seasonal patterns.

Data processing

Since January 2011 there has been a significant increase in the size of log-files and the peak was reached on 29 April when three million hits were recorded in a single day. 'Hits', each of which generates a log-file record, do not translate directly into web pages viewed or counts of unique visitors: a web-page is composed of many components, some visible such as images, others unseen by the user such as style-sheets and javascript. We also remove records of error pages; server hits that do not result in data presented to the user. The result is a set of records of pages viewed, from which we extract relevant information about the page, its content and the viewer.

The result is a very large database table: more than 150 million page-views, 4.5 million visitors, since October 2009. For each of those rows we can identify a multitude of attributes, but only a few hundred, among several thousand, occur with sufficient regularity and with the stable range of values that permit effective data mining. The aim of data-mining is not just to summarise these records in convenient tables, it must also find the hidden patterns and connections, cell to cell, within the whole table.

Log-analysis

A fundamental logging choice is between using an existing facility or designing a specific logging capability into the application. There are considerable advantages to making best use of existing mechanisms such as the server logs of the Apache server. They are readily available, the format is well understood, they can be created and processed without incurring significant development work. The disadvantage is that such log files have their origin as a tool of system administration; a format designed to monitor server performance and security, may not be ideal as the basis for market research and user-testing. On the other hand, not being designed to a purpose can be an advantage for our research; standard log files may be considered neutral: they were not designed to record only what we think we need to know.

However originated, in analysing log-files there are three basic approaches. The first is user-centric: observe how the web site functions interactively and correlate user actions to log records. The second is based on understanding software; what actions within the program mechanism generate a logged event. The third is to start with the logs themselves; data-mining the logs to discover association rules and hence predict patterns of behaviour. We employ all three but data mining is central to our approach. Not least because it is best suited to the analysis of existing standard log-files.

For europeana.eu we have had the advantage of access to software sources and developers; we could entertain the option of specifying a custom logging function. But although some changes to the log-file format, the recording of additional data readily available to the server, would be desirable, and were recommended (our report M3.1.Recommendation on use of logging analysis

tools in Europeana v1.0 June 2010) these have yet to be implemented. Consequently all our analysis is derived from the standard format log-file. It is a limitation, one that adds difficulties, but as noted the great advantage of standard files is that we can proceed independently of site specifics.

Attribute Analysis

Analysis of the first year's log data was largely been confined to attribute analysis. Each line in the log-file records a request for an item (html file, image, stylesheet,script) from the server. It identifies the user's IP address, and UserAgent (Browser), the date and time, the URL requested and usually the referrer (ie. the previous web-page visited that contained the requested link). It may be convenient to visualise the daily log-file as a very large table: each line divided into columns for IP, date, time, etc. As is now usual for a website of any size, the file component of the URL invokes a program, the 'web page' is held not as a single html file but is composited on request by the server. The important consequence of this for log analysis is that analysis by 'page viewed' tells us very little. To understand what is being requested and viewed we need to analyse the query string in the URL. Hence we need to decompose the URL into its various components including the query string, and the query string in turn is then further divided into its components (field=value pairs). Sometimes the values of the fields may themselves be composite and require further decomposition. The referrer column is also in the format of a URL and is likewise decomposed. A similar process can be applied to the UserAgent string. The result is that after processing one day's log file has become a table of perhaps one million rows and, in the case of europeana.eu, over four-thousand columns. In data-mining the columns are usually known as attributes; the first task is to identify useful attributes.

The utility of an attribute, and the value of the information it may yield, takes into account its reliability, ubiquity, frequency and discrimination. And, more subjectively, relevance to the purpose of the analysis. An attribute present in every row is more useful that one that is present only rarely. An attribute that takes a few well defined values is better than one that is almost, but not quite, unique to each row. Some attributes merely restate what we already know. As a result of this assessment a set of useful attributes can be selected. For europeana.eu there are currently around a hundred that in our judgement should be regularly monitored. Changes to the standard log-file format such as adding fields create additional attributes. It entails development work to implement these changes on the server, an increase in the volume of data to be transferred, and additional work to analyse the logs. We need to justify changes as a worthwhile addition to the current set of one hundred useful attributes and not an increase in the four-thousand varieties of noise.

Classifiers & Clustering

Having identified useful attributes we can combine them: we make connections between attributes, we group similar values and instances, we look for patterns and make the connections between patterns in the data and trends and communities in the world at large.

Pages, 'Users' , and Time

Logs are analysed 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

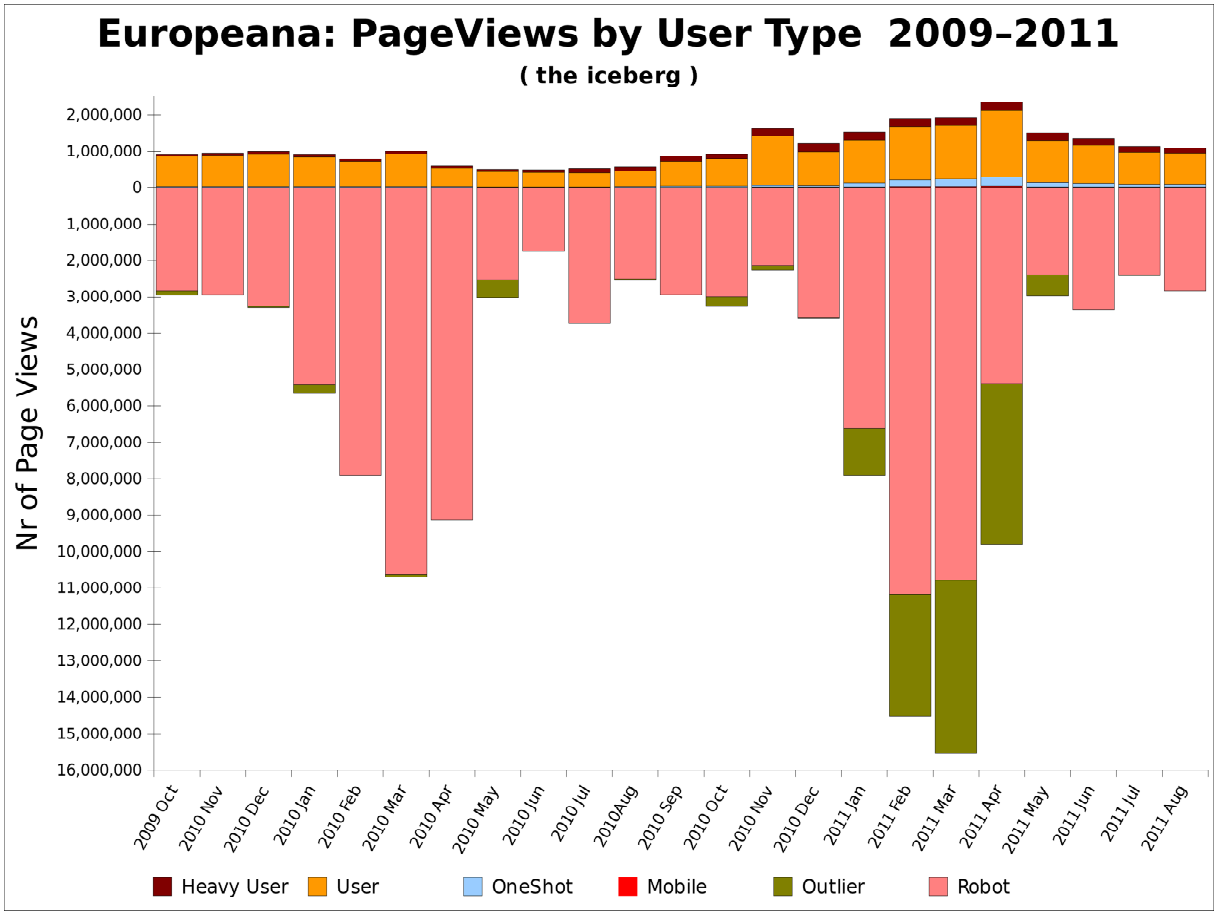
therefore a 'page view' of the europeana.eu site (and would not count as such for advertiser oriented 'page impressions' counters such as Google Analytics), however we are able to record these and they are included in our page view counts as 'redirect'. Additional analysis of 'shownAt' (the link text '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. Clicking on the Picture (shownBy) is far more popular: the popularity varies with type of collection and image but 60–80% of redirects are from clicking on a picture.

Visits (or sessions) are a sequence of page views that we can ascribe to one user, at one location, with an implicit continuity from first page to last. Although the Europeana.eu site uses session cookies these are not recorded in the log files, hence our visits are defined independently of the sessions defined by cookies. This provides some flexibility in the post facto definition of a visit. We are thus not constrained by the conventions of advertiser driven analytics; our visits seek to capture an Aristotelian unity of action, time and place. That is, a visit has one actor, begins with a referral from another site, follows a chain of links from one Europeana page to another, and lasts no longer than one day.

Traditionally time metrics such as session-time, are used to show site 'stickiness' as a surrogate for interest and satisfaction, the supposition being the longer the better. This is the advertisers web-view: more time on site means more time to view an advertisement. But we have no way of knowing if the user was viewing the site: a page was requested, some time later another page was requested, that is all we really know. So it is questionable whether these metrics do demonstrate interest and satisfaction. In the case of a gateway, portal or search-centric site like Europeana the opposite may be argued: the faster people move through the site the better and more efficient the indexing and navigation.

Robots, outliers, and real users

A user is defined by the combination of internet address and user agent string. Although not formally a unique identifier, in practice this proves sufficient to distinguish one personal user from another. In the case of the shared use of a browser in kiosk mode (as might be encountered in a library) it is possible that several visits may be merged, but this would also be the case when relying on session cookies.



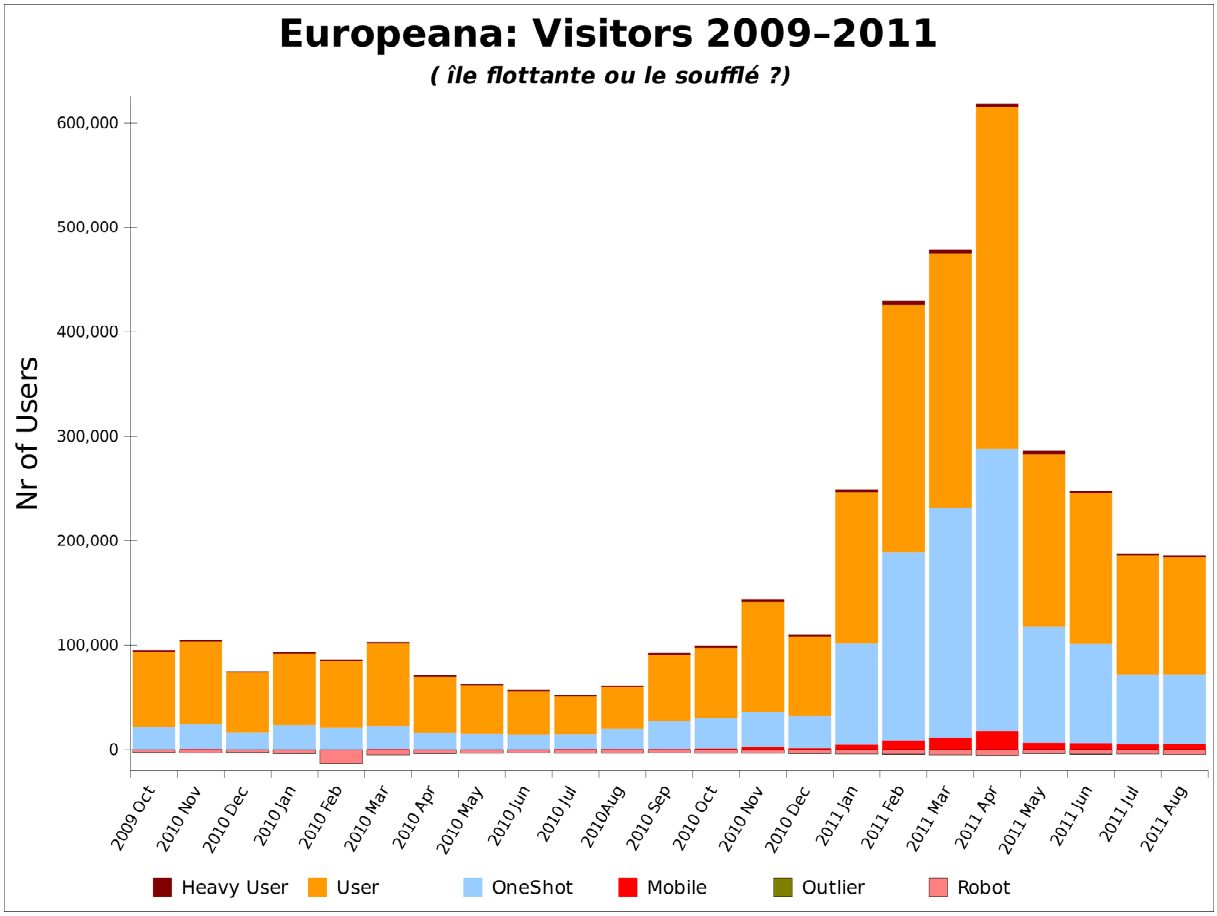
The distinction between 'user' and 'robot' is a separation of interactive use from automated data gathering. It is important: like an iceberg with nine-tenths of its bulk out of view, robots constitute 90 per cent of logged page views. For the most part this is clear cut: much robot activity can be clearly identified from the User-agent alone and is obviously distinct from a person seeking to find specific information or generally browsing. Such personal users are also clearly identified by their use of common browsers such as Internet Explorer or Firefox. The user/robot partition is thus reliable but nonetheless there are some ambiguities, particularly when our intent is to study how well a website serves the needs of normal conscious interactive Users. The effective study of user behaviour requires more than the binary distinction of user/robot.

Services such as "Google Web Preview" this will show a thumbnail image of the site's page alongside the search result presented by Google. In such cases the end-user has not visited our site but the Google agent has. The user has had a preview of the site: should that count as something akin to visiting the site or is it just another robot?

Some apparent users are in fact robots with malicious intent (e.g. DTS which gathers email addresses), others are benign (e.g. ECIS/Documentum Federated Search Services) but by acting a proxy for a search will misrepresent the user's activity. We also need to consider that sometimes there is considerable activity logged by developers and others close to the project. A careful study of user behaviour therefore requires more than the binary distinction of user/robot.

Of the Robots there are the obvious, overt robots: Googlebot, Yahoo, Yandex, msnbot. But there is another class, not declaring themselves as robot user agents but consuming resources on a scale far in excess of the heaviest of users. This category, the 'Outliers', are cases where use from a single IP address involves hundreds of thousands of pages usually at a very high and

steady hourly rate for a few hours or days. Although the user-agent is declared to be an ordinary web browser the behaviour has all the characteristics of a robot or crawler. An outlier may not be malicious or intentionally deceptive, it may originate from testing or development activity as part of the Europeana project, but it is not normal usage and is set aside in our analyses. During the course of the two-year eConnect project there have been around two dozen instances of this classification. Notably they were a significant feature in the Europeana logs in the first half of 2011. However since May 2011 this significant drain on the site's resources has not been active.



In addition to the Robot/User classification which largely relies on identification by user-agent, we use the location and institutional information that can be derived from the network address together with patterns in the timing and frequency of access to sub-set the 'User' category. For Europeana we have settled on four classes of 'real user' and two classes of 'robot' Because of rapid growth over the past year, a distinct page-format and implied information need we treat Mobile users as a special category. Heavy users are probably based in institutions and may well be associated with the Europeana project as a developer or provider. The remainder, accounting for the majority of page-views and over half of all visits we simply class 'Users'.

First, we identify `**OneShot**' users. 'Bouncer' is a label often used in web analytics, superficially it seems descriptive and intuitive: bouncers follow a link to a website but follow no links from the landing page, they view only one page, they leave little trace. Probably our site was not what they were looking for and they moved rapidly on, but we can only guess. But there is a problem to identifying 'bouncers', they are the elusive atomic particles of web-analytics. We know bouncers exist because a page view is recorded in the log. But we do not see the bounce because our logs have no record of where they went next. Unless they return to our site. In which case they are not

bouncers. So the fundamental problem is how long should we wait before concluding that they will not return?

Standard analytics practice is to allow 30 minutes for a 'visit' to time-out; one page landing, no return within 30 minutes, equals one bouncer. If we can identify repeat visitors (eg within 6months) then we might also class as 'singletons' those bouncers who are not returning visitors. But both the reliable identification of an individual user and visit timing depend on cookies which are not currently recorded in the Europeana.eu logs. In addition we believe this definition, established by convention and bound to an advertisers view of of the function of a website, is not best suited to capturing significant web-behaviour patterns. So in place of the bouncer we use an operation definition of 'OneShot' user: those who viewed one page in one visit and who appear not to have returned to Europeana.eu since we began looking at the logs in October 2009. This is more restrictive than 'bouncer': all OneShot users are bouncers, but not all —conventionally defined— bouncers are to be found in our OneShot category.

The remaining users are split into three further categories. Mobile users are easily identifiable because of the operating systems that they use. We then classify the remaining users (excluding OneShots and mobile users) into `heavy' or `normal' by page views.

Heavy users are genuine users, not robots, nor outliers, but over the past two years they have each viewed thousands of Europeana pages. In many cases this activity is based in institutions associated with the Europeana project: it will include both development activity from within and general use from public kiosks. Group use by schools and colleges is another common use case. The heavy user category numbers less than a thousand, mainly institutional users selected by internet address. The criteria used to identify heavy users are continually reviewed as the data accumulates. The objective is to select the heaviest users and set the lower bound at a level which captures the majority of internal (Europeana) and significant instances of institutional use (museums, libraries, etc.).

One shot or heavy, these are significant users, we need to know them better,and by applying a 'top and tail' filter to the user category, setting them aside for particular study, we also clarify the middle ground: the millions of anonymous Europeana users.

Because of rapid growth over the past year, a distinct page format and implied information need we treat **mobile users** as a special category. Selected on the basis of user-agent string, mobile users currently account for about 3 per cent of use. As the numbers are relatively small and the localisation of internet address less reliable we do not top and tail this category as we do in the case of general users. We can sub-divide this category into 'phones and 'pads' but with numbers small and the trend volatile we must beware of over-fine categorisation. For the future, we need to develop a reliable and relevant classification to distinguish both format constraints ('phone v 'pad) and mobility (roaming v tethered).

In the 12 months September 2010 to August 2011, Europeana had around three million unique users. Our `heavy user' category where many users may share an institutional connection, represent less than 1 per cent of all users, but these account for a far higher proportion of visits and page views. They will score higher in terms of `engagement' but because this category includes internal use by the Europeana project, such `engagement' will not be typical. Mobile use accounts for 2.3 per cent of users over the most recent 12 months. The remaining use can be split between `OneShot' users (41 per cent) who show no evidence of engagement and the rest (56 per cent) to which we can apply an engagement metric.

Measuring 'satisfaction'

Click-through: this is as close as we get to a success or satisfaction metric for Europeana. It is perhaps equivalent to a download in publisher platform terms, or maybe in e-commerce terms a 'conversion', it is implemented as a 'redirect', an instance of Europeana sending traffic to a provider/collection site. As Europeana is a portal/advertiser of collections then this is a hit for the provider and a 'sale' for Europeana. A click-through involves the viewing of two pages. From the user viewpoint page 1 is the Europeana Record and page 2 is the Collection Provider site.

We cannot measure **engagement** in the case of bouncers, and in all cases we can only guess the context or motivation that brought someone to Europeana. It is however reasonable to suppose that the nature of the page view will be a significant factor. A single view of a page such as 'aboutus' may provide a satisfactory answer, on the other hand the Europeana homepage, offers little to engage the user who goes no deeper into the site. A notable effect of the search engine optimisation in early 2011 was to greatly increase the number of bounce visits going to a record page rather than the homepage. But Mobile visits are more likely to go to the homepage.

Mobile visits are nearly twice (1.95 times) as likely to be bouncers than is the case for normal users, and more than ten times as likely for heavy users. They are 25% more likely to view the homepage, 25% less likely to view the record. We do not entirely know why this is the case, although there is much variation between Europeana partners in the extent to which they have adapted their offerings for mobile users.

A note on Visit (session) timing

Calculating time metrics

1.It is not possible to calculate times for Europeana visits of just one page and 65% of visits fall into this category.

2.There is no one-true-definition of a `visit' or 'session'. Using alternative definitions changes the total number of visits, the maximum duration, the number of pages, and the proportion of one page (non-timed) visits.

3.Google Analytics (GA) sessions are defined by session-cookies. Without these session-cookies we cannot perfectly reproduce the session data used by Google; these cookies are not currently logged so we calculate sessions from the bare log-file. Our calculations may not be strictly comparable. (We requested the cookies be included in the log-files with our M3.1.1 Recommendation on use of logging analysis tools in Europeana v1.0 of Sept 2010).

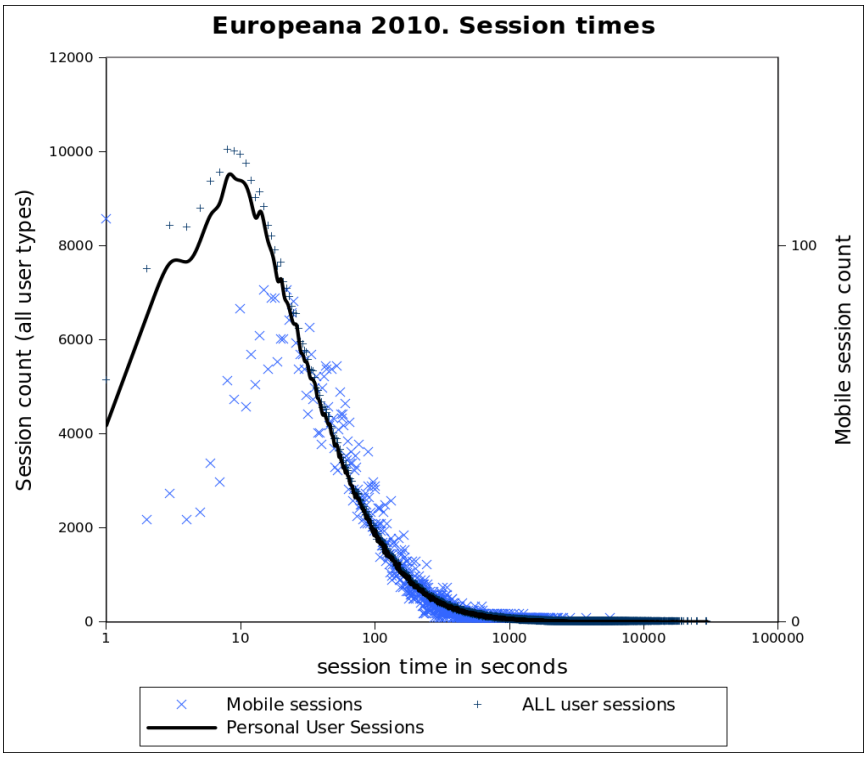
4.Usage time data are strongly skewed and therefore a simple mean (average) calculation as Google Analytics and many similar services provide, can be very misleading. The arithmetic average is a very poor statistic to use.

Europeana visit time typically follows a log-normal distribution, in other words there are lots of very short sessions and a few extremely long sessions. Many visits (65%) involve just a single page view and cannot be timed at all. Of those we can time, two-page visits of around 8 seconds represent the peak of the curve. However, a small number are very long: the longest time recorded is more than eight hours, the greatest number of pages in a single visit nearly 6,000. The presence of such extreme values causes the 'average', which is usually associated with a normal distribution, to give a misleading impression of session length.

However, if we plot the same visit time data using a logarithmic scale for the vertical axis, we get a classic bell-shaped curve (hence the description log-linear). By using this transformation, we can calculate a much more meaningful average, as we shall see.

The GA Google Analytics session figures we have (for May-Jul 2010) are around the 4 minute mark (3:56, 4:12 and 4:14). Using a definition of `visit' that attempts to reproduce the effect of the GA cookies, we calculate the arithmetic average (mean) session time to be 4 minutes 20 seconds. Our guess is that the GA average is a simple arithmetic mean with no sensitivity to the way the underlying data is distributed. It differs considerably from the median value, which we calculate to be 1 minute 26 seconds.

The simple arithmetic mean is much longer than is reasonable as a picture of the duration of the 'typical visit'. An 'average', whether median or mean, needs to be qualified by the context of its distribution. We calculate the arithmetic mean by averaging the natural logs of the data (because we are dealing with a lognormal distribution). Using this method, the mean and medians converge to near identical values. Since the natural logs of the data are normally distributed, we can go a stage further. If we add three standard deviations to that mean, we can identify an upper bound on what might be considered normal behaviour (capturing 99.7% of what might be considered normal human behaviour). Finally we note that session time varies significantly by user type: 70.4 seconds (mobile users), 80.9 seconds (ordinary users) and 63.4 seconds (heavy users).



Search Engine Optimisation (SEO)

In the later part of 2010 changes were made to the structure of European.eu that made the indexing of the site by robots, in particular Googlebot, far more effective. As a result from January 2011 there have been far more users of the Europeana site. And it has changed the way the site is used: these new users go direct from search engine to content (record page) rather than to the home page. We see far more users but they are bouncers.

There were two components to this change. Firstly a change in the format of the URLs made individual items on the site (records) much more visible to search engines.

Old Style: full-doc

`http://www.europeana.eu/portal/full-doc.html?query=Scotland&qf=YEAR:1900&tab=&start=4&startPage=1&uri=http://www.europeana.eu/resolve/record/00401/248AE744BF9E19BA04FCDBDCF9EE9ACADA5AEB78&view=table&pageId=bd`

The 'query string' component of the URL, everything after the '?', is usually ignored by search engine robots, consequently the full-doc page appears to be one page with ever changing content.

New Style: record

`http://www.europeana.eu/portal/record/09405a/EDDB8C4DD7DF41EE9F5A5569C03F70E15AF2199F.html?query=modes+OR+fashion&qf=TYPE:IMAGE&start=13&startPage=13&view=table&pageId=brd`

The record identifier is now part of the resource id, each record is seen by the robot as a distinct page and is indexed.

A similar problem with the search result thumbnails

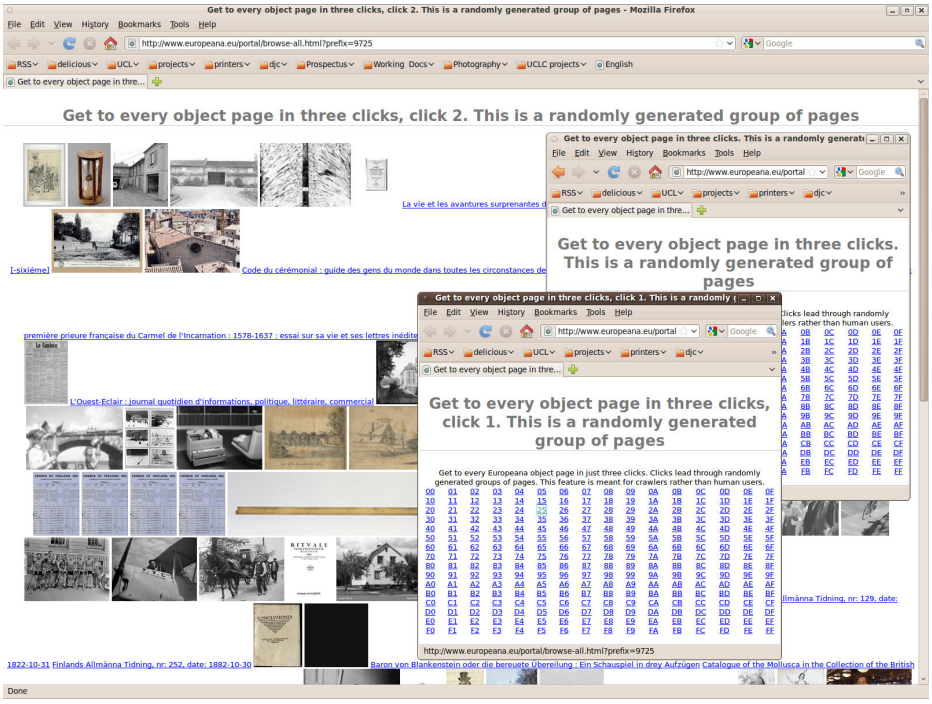
`http://www.europeana.eu/portal/search.html?query=sole+bay`

Robots ignore the important part, so the search results appear as one page with never the same content twice.

So to recap, the thumbnail display 'search.html' leads to www.europeana.eu/portal/record/... , which leads to the provider site (via redirect.html). But to a search engine all the search and record pages appeared as a single page with ever changing content. The search engine robot is caught in a maze with no clear navigation. The content [record and providers site] is never properly indexed.

Browse-all

The second important change was the 'browse-all' page designed for search engines and bypassing the brief-doc page. The search engine can now use browse-all as a gateway to the full content of Europeana, each indexed as a distinct identifiable and retrievable page.



Since January 2011 there has been a huge increase in referrals from Google (The only search engine that counts: whatever their share of the search engine market, we see negligible referrals from other search engines —0.2% Yahoo, 0.1% Bing against 68.5% Google.). They go direct to

the record rather than the home page. And from the record there is link to the provider site. So we are seeing an increase in 'conversion rate' driving traffic to provider sites, making content more widely available.

Here are some key findings:

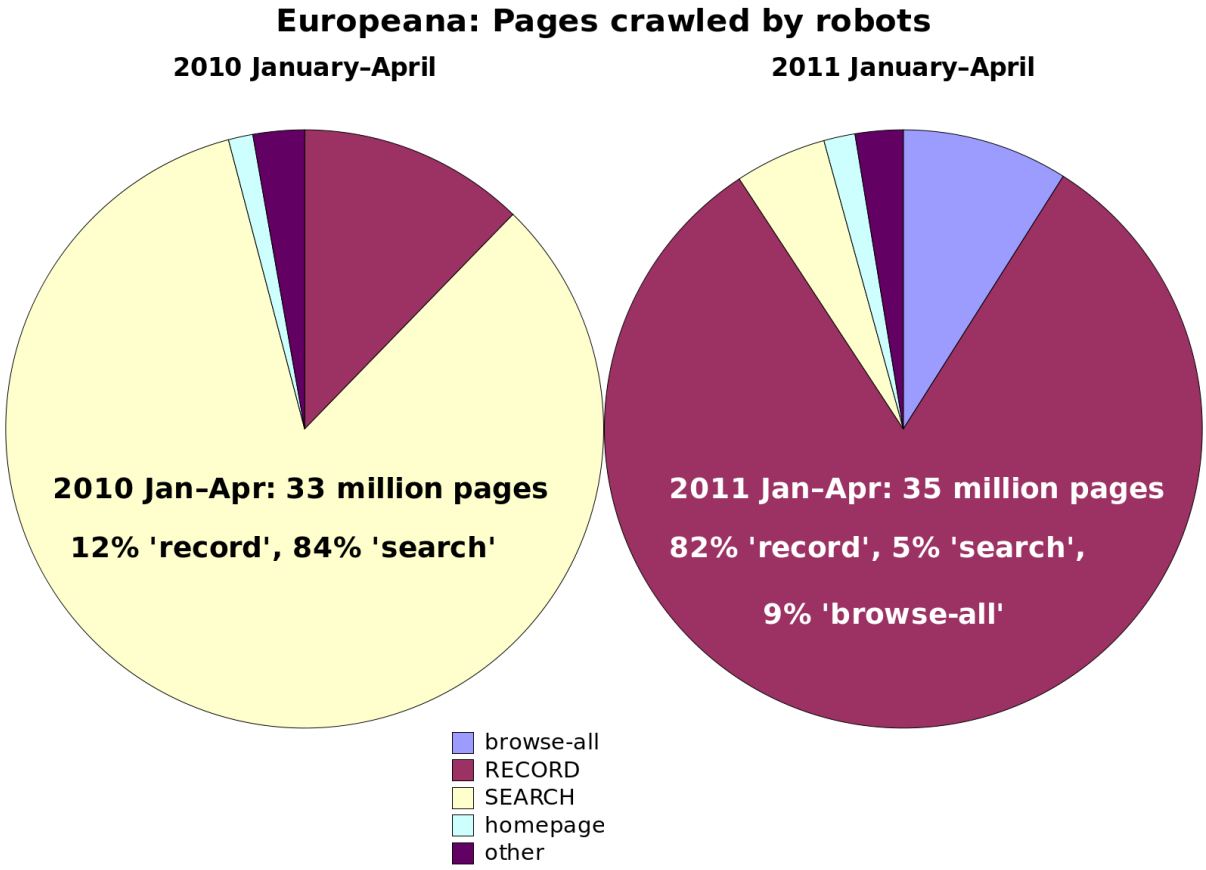
• Four key pages account for 94% of all page-views in 2010: Home-page(5%), brief-doc or thumbnails (50%), full-doc/record (36%) and redirect or click-through (3%)

• Jan-Apr 2010: Google was responsible for 23% of all Visits (150,000 out of 656,000); 69% were to homepage (104,000), 8.6% to record [full-doc] (25,000);

• Jan-Apr 2011: Google was responsible for 57% of all Visits (1,775,000 out of 3,141,000); 5.5% were to homepage (97,500) and 94% were to record (1,673,500).

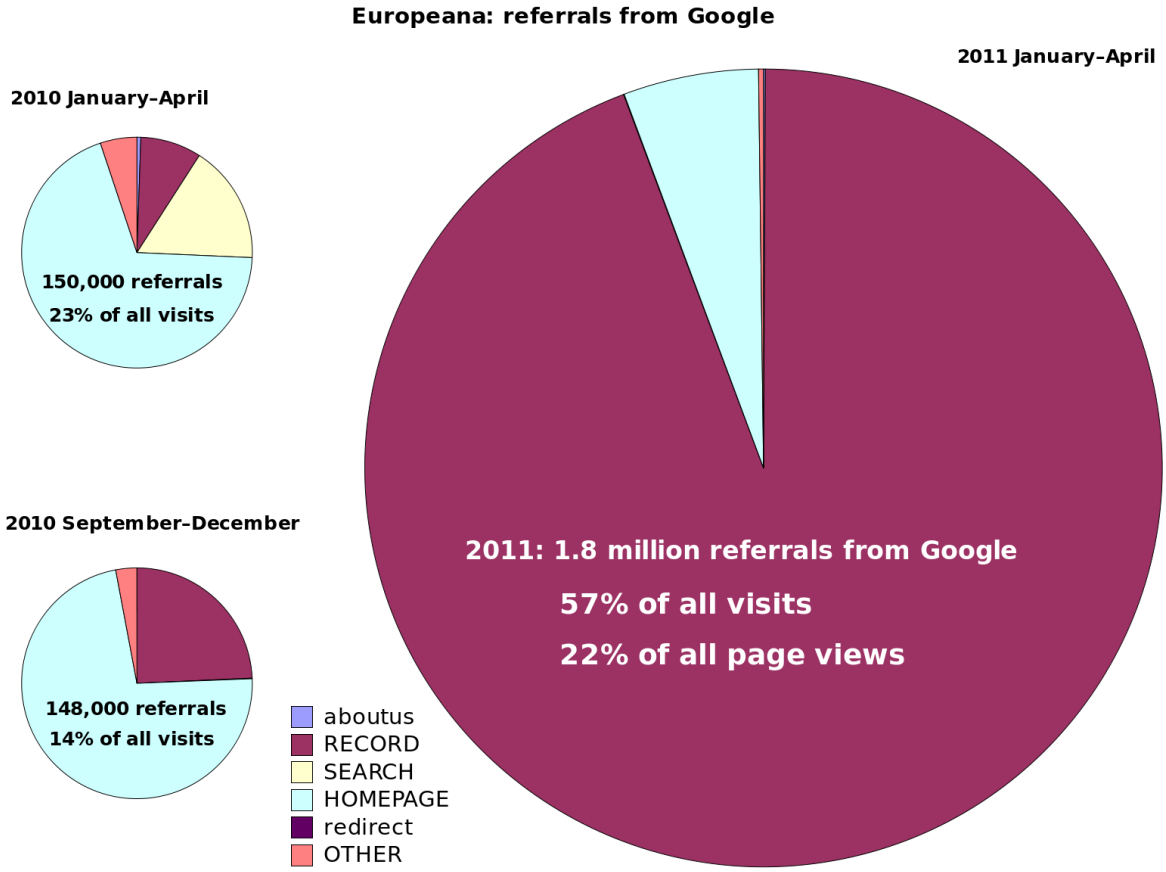
Robot use

Now that Europeana records are being indexed, a lot more robots (in fact the usual team of Googlebots) are following browse-all and going straight to the record, by-passing brief-doc, as the following graphic shows. The impact of search engine optimisation is very dramatic.



Human use

As a result of this indexing at the record level, Google referrals have increased tenfold in terms of absolute numbers of visits. There are a lot more users, a lot more bouncers, but also a lot more 'conversions'. A small slice but a much bigger pie.



Forecasting numbers of Europeana users

We offer a forecast of the numbers that Europeana may expect to see during the twelve months ending December 2011 and December 2012. The mathematical models that underpin these forecasts fit the historic data very well and we can be quite confident that they will generate reasonably accurate short term forecasts.

We predict that five and a half million individuals will visit Europeana during the 2012 calendar year, compared with just under three million in the latest set of figures: the 12 months ending in August 2011. Mobile visitors are the fastest growing category of Europeana users and their influence will be felt increasingly, and our estimate that they will comprise around 17 per cent of visitors during 2012 is somewhat conservative. Enhancements to the mobile Europeana experience, which is currently quite limited (see later in this report) is likely to change this proportion considerably, since it is likely that the installed base of internet-ready mobile devices will actually overtake that of desktops and laptops around 2013.

	ACTUAL	PROJECTED	PROJECTED
User type	Unique users 12 months to August 2011 (000s)	Unique users 12 months to December 2011 (000s)	Unique users 12 months to December 2012 (000s)
Heavy	18	22	45
Normal	1,663	1,879	2,710
Mobile	69	134	977
OneShot	1,225	1,352	1,818
All users	2,974	3,387	5,550
Mobile as % of all users	2.3%	4.0%	17.6%

Site Navigation

User navigation within Europeana

As noted earlier, the classic pathway that one might expect a user to take through Europeana would be homepage -> search ->[more searches] ->record ->redirect (i.e. the provider site). We have already noted, however, that opening up the site to Google indexing has changed the game, with many more bouncers going straight to a record.

Page Transits

Evidence that the classic mode of navigation is now less relevant than it was can be seen in the table of page-transits. It shows how users actually navigate their way around Europeana in terms of transits from page (rows) to page (columns). The data are for the period May–August 2011, post-search engine optimisation, and the numbers in the table are percentages of all page-to-page transits. Around twenty types of user pages can be found on Europeana. However, not surprisingly, most views are to Europeana's content: search (thumbnails) and record, which display content within a standard frame.

from Page (row)\to Page (col)	#_Total	search	record	homepage	redirect	aboutus	login	OTHER
#_Total	3983970	1341002	1229739	563517	558986	100617	32831	157278
search	1490878	1075507	399718	5647	0	1571	2383	6052
_OFFSITE_LINK_	1167532	47999	551921	486810	35698	17044	3170	24890
record	895946	78153	267125	5330	522385	1611	2579	18763
homepage	228924	114022	1635	48306	781	25055	6229	32896
aboutus	75618	3300	3665	3687	0	43765	5119	16082
timeline	24307	14153	3892	639	0	229	406	4988
OTHER	24717	7536	1005	3839	6	984	1117	10230
login	21487	2	0	3638	0	6459	7683	3705
communities	17630	0	331	596	0	1448	1270	13985
thoughtlab	13942	0	0	695	0	656	966	11625
partners	6951	0	0	717	0	895	934	4405
register	5765	123	0	2013	0	17	258	3354
usingeuropeana	3490	0	0	257	0	199	207	2827
myeuropeana	2862	200	291	472	0	478	298	1123
rr	2361	0	0	808	0	73	147	1333
contact	1038	0	0	54	0	132	63	789
browse-all	380	0	156	6	0	0	0	218
redirect	118	0	0	2	116	0	0	0
new-content	24	7	0	1	0	1	2	13

Nearly a quarter of all page-views are now referred into the Europeana site direct to a record: predominantly by referral from Google. This is twice the number of visitors who commence their

visit by starting at the home page (where a leading referrer is Blogspot). Many of these visitors will be bouncers; they do not view any other pages. Where the visitor does view multiple pages the predominant flow is much as expected: from homepage to a search, possibly multiple views of the search result thumbnails, then to a record, and from a record to a redirect to the provider site. But since such as flow represents a minimum of four page views, few visitors stay the course.

Media format

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

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.

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. But the practicality of Europeana, as currently implemented , is that every interaction generates a search. 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.

Most frequent referring sites

The top referring sites, for the period May to August 2011 are shown below. As noted, Google is the most frequent referring site by some margin, accounting for 57% of all externally-generated traffic to Europeana. The units are numbers of visits. There is significant traffic from various blogs hosted by blogspot and directed to Europeana from PIONIER online, a catalogue of the Federated Digital Libraries of Poland.

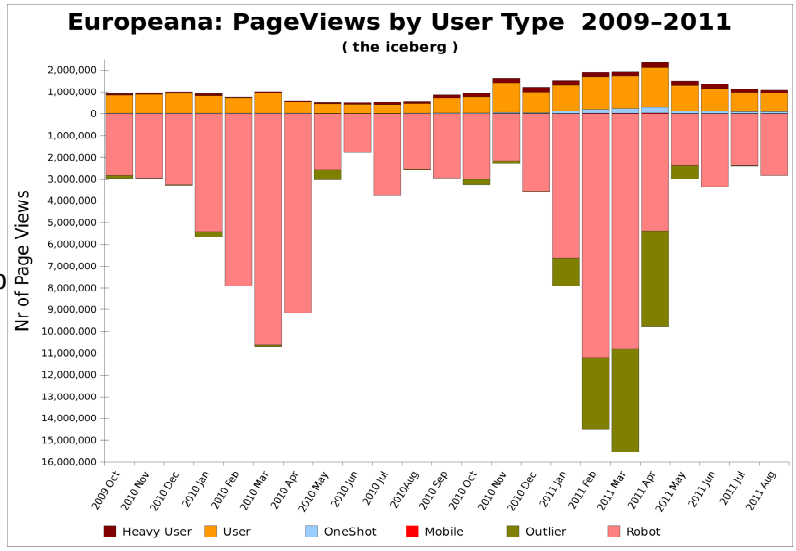
Referrer to Landing Page	#_Total	record	homepage	search	aboutus	redirect	OTHER
#_Total	715415	351401	301810	28623	12933	10565	10083
GOOGLE	318697	277534	39301	477	465	61	859
-	260764	45220	188636	13405	6255	1637	5611
_KNOWN_OTHER_	90728	22713	42270	12851	3261	6700	2933
BLOGSPOT	16938	66	16598	91	8	162	13
FACEBOOK	3048	1403	822	369	426	7	21
GOOGLEUSERCONTENT	2971	624	250	596	269	859	373
BING	2696	1861	817	1	11	0	6
WIKIPEDIA	2538	212	2055	205	55	1	10
www.emob.fr	2384	0	2384	0	0	0	0
YAHOO	2052	1128	700	74	71	5	74
www.kb.nl	1996	0	96	0	1900	0	0
www.bnf.fr	1646	0	1646	0	0	0	0
EUROPA	1406	0	1401	0	4	0	1
WORDPRESS	1066	155	307	28	8	557	11
www.heise.de	743	0	743	0	0	0	0
LIVE	580	113	284	68	52	21	42
TWITTER	492	65	235	31	139	0	22
roai.mcu.es	482	81	381	20	0	0	0
www.netvibes.com	479	5	464	0	8	0	2
www.nytimes.com	471	0	471	0	0	0	0
www.elgrancapitan.org	419	0	0	0	0	419	0

BLOG	418	40	277	9	1	0	91
www.bn.org.pl	395	0	395	0	0	0	0
www.digmap.eu	361	0	361	0	0	0	0
scd-sfx.u-strasbg.fr	269	0	0	269	0	0	0
www.deutsche-digitale-bibliothek.de	233	0	233	0	0	0	0
www.service-public.fr	179	0	179	0	0	0	0
www2u.biglobe.ne.jp	177	177	0	0	0	0	0
www.culture.gouv.fr	170	0	167	0	0	3	0
www.e-book.com.au	155	0	155	0	0	0	0
www.lecdi.net	142	0	21	121	0	0	0
doucetpiquante2.canalblog.com	128	0	1	0	0	127	0
BLOGS	125	1	111	5	0	0	8
app.e2ma.net	31	3	19	3	0	0	6
tek.sapo.pt	27	0	27	0	0	0	0
www.photo.rmn.fr	6	0	0	0	0	6	0
winfuture.de	3	0	3	0	0	0	0

Most popular providers and collections

We have analysed click-through activity to detect the most popular collections and the sources that generate traffic to both Europeana.eu and to the provider sites. Although we give some results in Part II (tables) the results are as yet subject to some caution: it only counts views of the record page, there are biases introduced both by the featuring of content in pre-formatted 'searches' and testing activity, and finally the process of identifying collection and provider from the log record requires refinement.

As already noted, in terms of page views Europeana use is dominated by Robots: like an iceberg nine-tenths of pages are hidden from view, they go unrecorded by GoogleAnalytics but they are nonetheless essential, as we have seen above, effective search engine optimisations essential to the visibility of Europeana on the world wide web.



What may be less useful however are the 'outliers'. In the first three months of 2011 three IP addresses allocated to an ISP in Spain accounted for 20% of all page-views. This level of activity is not credible as coming from a lone genuine user and there is no evidence that this is a NATed or proxy connection. In the case of genuine search engine robots (e.g.Googlebot's 40% of pageviews) the benefits are clear, but in the case of these outliers it is not. Certainly we do not see an increase in visitors from Spain that we might expect as the result of such intensive localised search engine activity.

Another pattern is seen if we consider not page-views but Visitors. While insignificant in terms of page views, One-shot users are far more visible when we count users. Most of these are the result of Google referrals and there is some evidence that French users, already well represented in Europeana content and usage, are also the most significant of the One-shot visitors. It seems there may be a very high seasonal use from French schools in January–April each year, much of it mediated by Google search.

There is a challenge for Europeana in this: how much of the growth in One-shot use in particular can be converted to a sustained interest in Europeana. Search-engines index pages bring in visitors, but what makes a site 'sticky'. What sort of growth have we seen these past four months: Île flottante or le soufflé?

Temporal Patterns

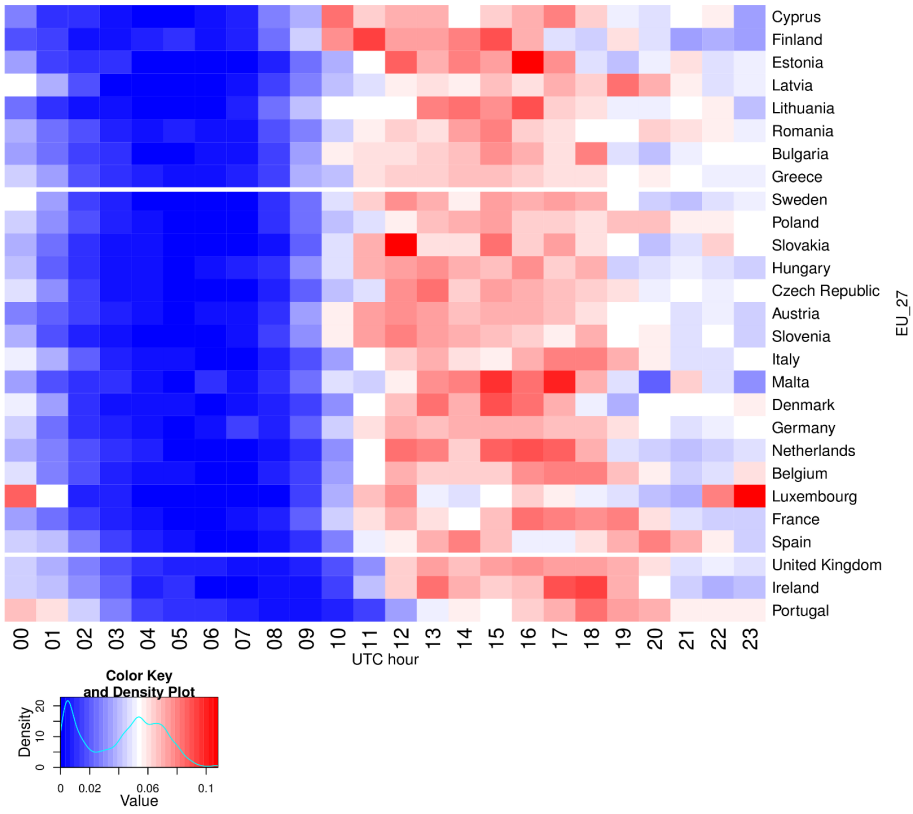
Hourly patterns and Timezones

During the average day, visits to Europeana peak in the late afternoon but activity never ceases with a significant amount of `night time' traffic much from outside the EU time zone.

A visualisation of Europeana use over a 24-hour period is shown in [fig n] 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.

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.

fig 2. Europeana peak hours (UTC+00) for EU–27 countries



Weekly patterns

With regard to temporal patterns at the weekly level, usage shows a distinct peak on Thursday and lower but still significant levels over the weekend, at about three quarters of the level seen during the working week.

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.

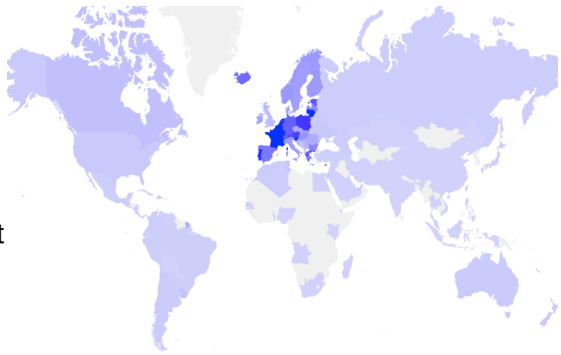
Monthly patterns

Looking at Europeana logs over a full 2 years, we begin to sense a seasonal rhythm and gain some early insight into the growth trends. In summary use rises strongly from December to March and a lull over the summer holiday period, which suggest that Europeana does not at present appeal to the tourist.

Local Patterns

Based on number of visits France is the largest single user of Europeana, accounting for 16% of all visits. The next highest destinations are Germany (14%), the USA (10%), Poland (7%) and Spain (7%). These five countries now account for over half of all (54.1%) of Europeana usage; the top ten countries for more than three-quarters (75.8%) of visits.

Mapping the worldwide distribution of Europeana users relative to population (per capita) reveals Europeana's reach to be rather Eurocentric.



Further analysis shows that users tend to focus their interest on collections maintained within their own country. (Austrians looking at Austrian collections, Slovenians looking at Slovenian collections, etc.).

In this case 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 record.

The pattern is more readily appreciated when presented as a heat-map. 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 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 11a. Europeana Collections and their markets

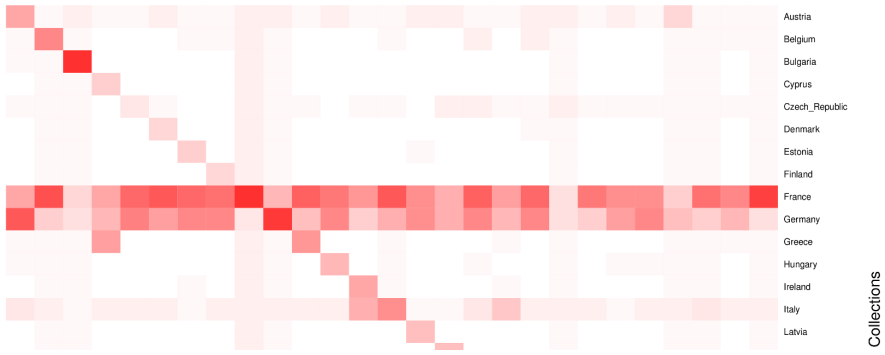
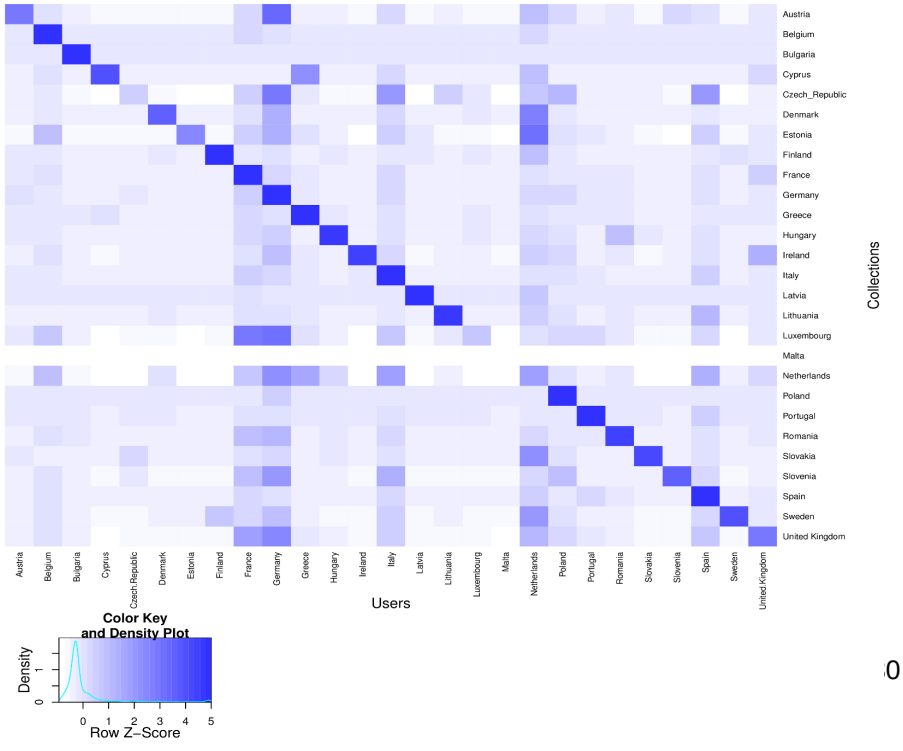
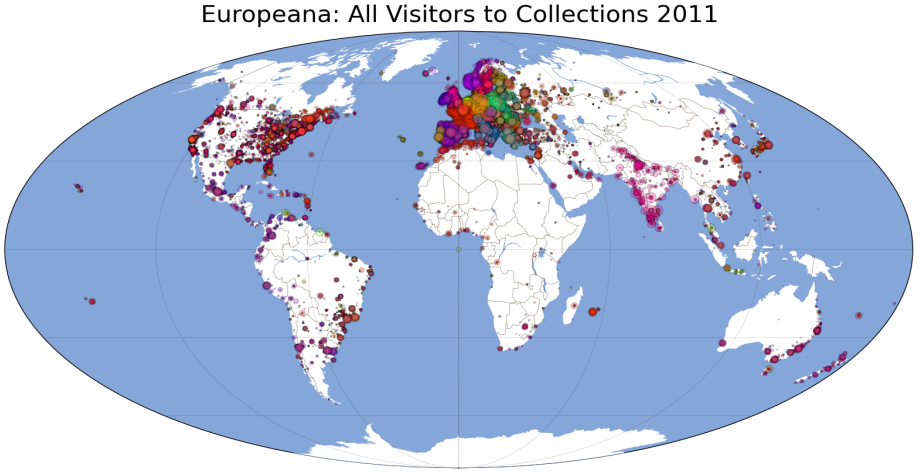


fig 11b. Europeana Users and national collections

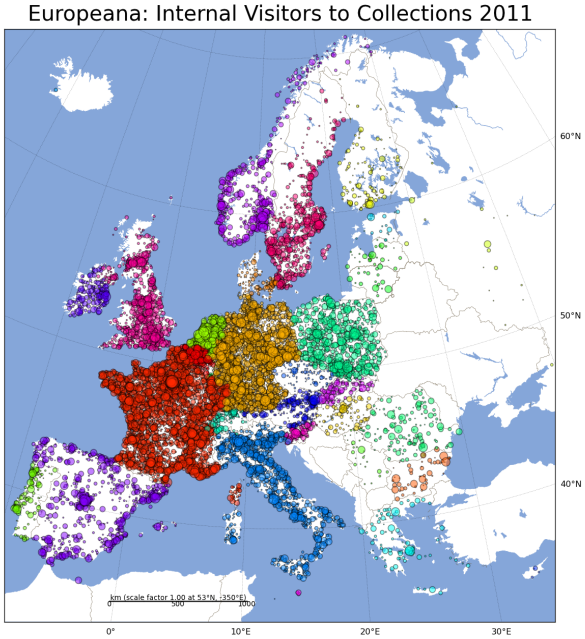


Europeana: All Visitors to Collections 2011



The analysis of this phenomenon can be extended by plotting the location of users and the collections they access on a map. In each of these maps circles represent the location of users, the size of the circle is proportional to the number of visits, and each colour a different national provider. [Additional maps in appendix show greater detail]

Europeana: Internal Visitors to Collections 2011



Most visitors view very few pages, and only a very few view the `record' page. So this is an analysis of a subset of visitors: those who viewed at least one record page. But with so many

users now diving straight into the record page from a Google referral that is a significant though perhaps not fully representative sample. In the few cases where more than one record was viewed in a visit, the country is the one most frequently occurring per visit.

In 2010 76% of visits did not result in a record view. In 2011 this was down to 56%. This can be interpreted as a sign that visitors are going deeper into the site; drilling down from search results to the collection record; but we need also to consider that most visitors are bouncers, viewing only a single page. Since 2011 that single page has most often been a record. The big increase in record views can be attributed to the deeper indexing of the site that is sending Google users in particular straight to the record. It appears that this trend toward direct access to records has been most notable among users located in France: in 2010 they were not the most frequent bouncers (Germany, and USA being slightly ahead), in 2011 there were 360,000 single page visitors from France, thirteen times the number a year earlier. By contrast single page visits from

Germany numbered 144,000 in 2011 a less than fivefold increase over 2010. Nonetheless the range of provider countries has grown: in 2010 99% of content came from 10 countries, by 2011 that 99% of content was provided by 19 countries.

It appears that French users (including DOM-TOM) are over represented relative to material in general. But as we observed in past reports we do see a strong preference for video from all users and it appears that France contributes the majority of the Video Content. It is difficult to classify content type from the log data so we cannot say at present if the majority users of video are French, and we will take a further look at this for a future report.

Clustering

People visit Europeana for a wide variety of reasons, from simply finding themselves there as the result of a Google search to a planned and intensive research session.

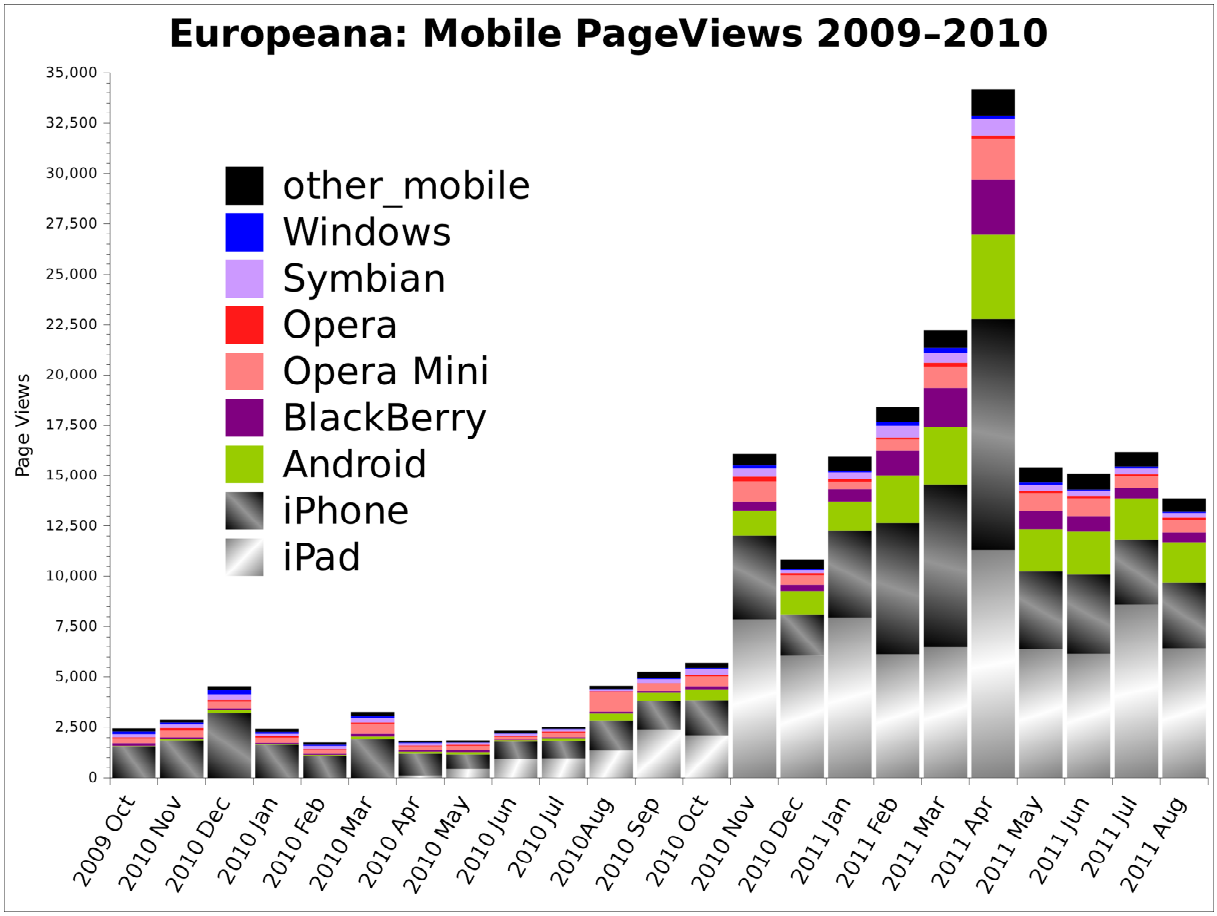
Many visitors (52 per cent) are `bouncers' who only view a single page, very likely having been swept there courtesy of a general search engine like Google. A possibly high proportion will never return, but that is not to say that they may not have extracted valuable facts or information from that visit. A large minority of mobile users make relatively brief visits of just under two minutes and engage in real interaction with Europeana, typically conducting a single search and viewing several pages of content. A small but relatively high proportion of these visits are referrals from social media or blogging sites (a third more referrals than expected) and this indicates interesting potential for the social media plus mobile use combination. We are provisionally associating these kinds of visits with a form of `checking' behaviour - they appear to be fact-finding or checking in nature, short and sharply focused. This leaves a small minority, around six per cent of visits, that are characterised by considerably longer duration (around ten minutes) and much higher degrees of interaction with Europeana software and content. This is the kind of behaviour that one would associate with a need for more in-depth research or perhaps users who are simply exploring the website to see what Europeana can offer them.

Media

Culture on the Go: Mobiles and other iThings

The availability of mobile platforms such as the iPad and iPhonefor viewing web content has expanded dramatically and this reflects in a rapidly growing share of Europeana page views. In 2010 mobile use was 0.5% of all page views, in the first four months of 2011 1.75%

The growth of pages viewed on mobile devices is estimated to be of the order of 191% per annum (compound growth).



As of April 2011, the fastest growing segments were the iPhone with 33.6% of all mobile page views and the iPad, with 33.0%. Apple devices have therefore captured two thirds of all Europeana mobile use.

Internet use via mobile phone and tablet offers a different user experience from the desk-bound PC. This is not just a growing platform: mobile user interface designs are beginning to influence the look and feel of desktops. And the growth of a market for 'apps' suggests it may be possible to find users willing to pay for content.

Not all mobile devices are the same and there clearly is a difference between 'phones and tablets, yet the latest 'smartphones' now offer screen resolutions far higher than was normal on a desktop a few years ago. The assumption that 'mobile' means low resolution and restricted bandwith cannot be relied upon. Changes to the Europeana.eu site introduced in October 2011 mean tablet users will no longer be presented with the 'mobile' interface designed but a few years ago for small screen phones. This study can only report on the situation as it appears at the end

of summer 2011. The situation is rapidly evolving and even the most up-to-date data is already of only historical interest.

Europeana is proving exceptionally popular for users with mobile devices. Because we are starting from a very low base (in January 2010, for example, there were fewer than 3,000 mobile page views) it is difficult to predict the future with certainty. However, over the 12 months from August 2010 to July 2011, page views from mobile devices grew at a rate than four times greater than from fixed devices, with the fastest growth coming from the iPhone.

Mobiles are a very fast growing market segment for Europeana, still small but it has quadrupled in the past year. The real change for Europeana has not been in smart-phones but in tablets. The iPad has achieved a breakthrough making the tablet (big touch-screen, un-encumbered by wires or peripheral devices) a popular platform where previous attempts have failed.

It redefines the consumer 'personal computer' experience; in fact it is an 'interweb' access-device rather than a computational machine. It makes apparent the difference between telephone/internet access and PC as office machine (even if that office is in the home). Tablet-oriented interfaces are influencing design of PC interfaces e.g. Gnome3, KDE4. The iPad has shown the way to go and is now being chased by rivals such as Android.

Mobile (smartphone and tablet) use is personal use, happens at evenings and weekends; occurs in the home or 'anywhere but the office'. It is about consuming content not creating it. Social networking, courtesy of the mobile, may be creating contacts and networks but it is not content as envisaged by those who suppose 'content is king'

Three years ago Europeana was prescient in considering the mobile user in its development plans. But since then 'Pad' has changed the way we need to conceive the 'mobile' user. Where once there was a clear difference between mobile and PC the differentiation that is opening up is between Office and Personal. The Office is the desktop and laptop, keyboard and mouse, work and study, documents and organisation. The Personal is 'Pad and 'Phone, touch-sensitive and wireless, conversation and affiliation, in a word mercurial.

Annex "Culture on the Go"

see separate document

corrigenda

"So, by definition, there can be no bouncers within the OneShot category." should read "So, by definition, all in the OneShot category are 'Bouncers'"