Learned Publishing, 27: 185–194
doi:10.1087/20140304

Evaluating information seeking and use in the changing virtual world: the emerging role of Google Analytics

D.J. CLARK, David NICHOLAS
CIBER Research

Hamid R. JAMALI
Kharazmi University, Tehran, Iran

ABSTRACT. The paper identifies changes that have occurred in the Web environment over the last decade which have gradually rendered server logs, once the pre-eminent source of intelligence on usage and information-seeking behaviour, an ineffective, impractical, and uneconomic resource. It also looks at the implications of these changes for information professionals and publishers who have come to rely on this data to understand the behaviour of clients and customers in the virtual environment. Ubiquitous and expanding, Google Analytics generates statistics about a website's traffic and traffic sources, albeit from a marketers' perspective, is evaluated as a possible replacement; something which might plug the user knowledge gap which is worryingly opening up, or maybe even, put us in a better position overall. The paper is built on the knowledge and experience of evaluating server logs for more than a decade, mostly for publishers and libraries, and also on two recent projects where server log analysis was supplemented with Google Analytics.


D.J. Clark


David Nicholas


Hamid R. Jamali

Introduction

The World Wide Web (WWW) has not only provided broad access to a huge variety of online information, it has also enabled an unprecedented gathering of information about online usage – web analytics. For more than a decade the analysis of HTTP access logs has provided information professions and scholarly publishers with data on the information-seeking and usage behaviour of digital consumers: abundant data with a reach and specificity never seen before. However, in today's more complex online environment such analyses are becoming no longer economical or effective. This results from dispersed and individual use rather than institution-based access, and complex web applications that defeat simple measures of content and access. The game is changing: access logs of online activity no longer yield the useful information they once did. Indeed, there is a risk that the supply will dry up and we will be in the unenviable position of knowing less about more users. Given the sparseness of the published literature on the subject it would seem that few people have woken up to the seriousness of the situation that we are facing.

This paper chronicles the changes and challenges that confronts the web analyst and gauges the ability of Google Analytics to overcome the difficulties being faced. The research and knowledge upon which this paper is built comes from a number of current CIBER Research projects that have confirmed the diminishing returns that can be expected from deep log analysis, including an investigation of 'turnaways'[1] and two studies featuring both log analysis and Google Analytics, an evaluation of Europeana,[2,3] the gateway to Europe's cultural heritage and CVCE, an interdisciplinary research and documentation centre dedicated to the European integration process.[4]

The goal and challenge of capturing user information

For any provider of online information and web analyst there are in essence three points to can 'take the pulse' of the digital consumer: incoming (access) logs, internal accounting, and tracking and tagging. Internal accounting will give the best information about users, but it is confined to 'your users', those people who sign-up, give demographic data willingly, and log-on. Access logs capture everything, including much clutter and noise, but for some time were useful, at least in part because Internet access used to be institution based and therefore could yield some information about users, as well as usage. Now cookies and trackers are ascendant, providing feedback on what has been consumed. But, automatic and generally surreptitious, they can be resented by users concerned about privacy and the ethical boundaries of shared information. Yet, unadorned, they count web-consuming devices (smartphones, tablets, etc.) not users. The problem of identifying real people remains. Thus we find a common theme in cloud services, synchronization, Facebook, Twitter, Google, Amazon, eBay: all inducements to sign-in and thereby provide that valuable property of user identification.

cookies and trackers are ascendant, providing feedback on what has been consumed

We argue here that there is an essential tension between an ideal of free and open access to information and the desire of all information providers (in public and private spheres) to know who their users are and how they behave online. Access to online information was once wholly restricted by pay-walls and log-on credentials, but the WWW changed much of that, creating a new online environment of free access and open data. The early web offered free information, but within the context of access mediated by large organizations, most notably academic institutions. There was for a limited time a golden age for analysts, in which the user could be identified, if not as an individual, at least by affiliation to institution, country, etc. Efforts to shepherd information resources into walled-gardens, by and large, are being frustrated, certainly in the scholarly world, by a powerful open access/data lobby. The loss of these 'border controls' have made the web-user anonymous; that is to all but the fortunate web provider that can convince users to part with personal data, and they are highly unlikely to release this to a wider public. We have gone from a world in which there was no access to anything without some sort of sign-up and log-on, through a period of pleading for members for 'myWebsite', to the present social media appeal for followers and their likes.

there was for a limited time a golden age for analysts

Today, then, the challenge for information professionals and researchers wanting to understand virtual-information-seeking and reading on the world stage can be best summarized as how can digital behaviour be ascribed to a class of users.

The old way of evaluating digital usage: HTTP-access-log analysis

The original WWW was simple text-file based: a web-page retrieved by a web-browser mapped directly to a text-markup (HTML) file on a server. The innovative step was the HTML: Hypertext Markup Language linked one page to another, the link as displayed by the browser was clickable, one page led direct to another, not just in one local filestore, but using HTTP (Hypertext Transfer Protocol) anywhere 'out there' a worldwide web of knowledge was spun. These HTTP transactions were logged by default by system administrators and programmers; for their own purposes, to see how well things were working, to lookout for problems. Then, because these logs tracked in some detail the usage of this new medium, the log data, almost as a by-product, became of interest for market research, for analysing digital-information-seeking behaviour.

The default information provided by the access log is in fact rather sparse: IP address, date and time, URL requested, referrer URL, and user-agent (this last field usually some indication of device, operating system, and browser used). Other standard fields are either normally unused or not very useful. Within a chain of referrals and requests the time can give some indication of the length of visit. Such is the theory because there can be no visit timing without a beginning and an end; the single page view (and most visits are of that kind) has zero duration. The IP address can serve as an imperfect user identification; it can also provide an indication of location and of institutional use. Once-upon-a-time anyway: where once an IP would often track back

to a university department, today, it is likely to be mobile and transient.

The log format can be modified – just adding a session cookie to the record will greatly simplify reconstruction of a visitor's journey – but HTTP access logs were never an ideal way to track web usage. They were never designed for that purpose. But even a data-mine with poor yield can be a rich information resource if it provides a lot of raw data for very little effort. And it did, and hence the attraction.

However, a great deal has changed in the past decade or so. Web-pages are no longer simple text files with a few ancillary pictures and scripts. The modern web-page is better thought of as a complex and hybrid confusion of program code and content that is processed by the web-browser virtual machine. That mobile devices use discrete 'apps' rather than an omnibus browser to serve the equivalent content is a recognition of this reality. And the situation on the server is similar: no longer a HTTP request for a file on the server, the link from page to page is now more likely to involve an Application Programming Interface (API) and a Content Management System (CMS). In plain language, an exchange of program instructions from one program to another; and much of it automatic and obscure, difficult to attribute to user interaction. The result is that the HTTP access log is no longer a cost-effective and reliable insight into the stream-of-consciousness of the online user.

Of course, it has never been easy to extract useful data from the logs. In reality HTTP access logs provide traffic not transaction analysis and too much of that has always been either noise or irrelevant to user studies. Web crawlers and other robots often account for 90% of all website traffic. The ancillary 'page furniture' such as images and stylesheets add volume to the log file but rarely information. But today's traffic flow is far more diverse and fragmented: the content that makes up the user's experience of a 'page' may be assembled from a variety of servers making the single log file incomplete. Asynchronous and autonomous scripting means there is no strict temporal sequence to the log record. Nor can we be sure which requests were of the user's own volition.

Another fundamental problem of the HTTP access log is that it records the request sent to the server, not what the server has delivered to the user. That did not matter much so long as there was a simple mapping between URL (which was logged) and an HTML file (the page displayed). But now, making sense of HTTP requests will always require some reverse engineering of the CMS. Slow, effortful, and generally inefficient. Better, far better, would be to tap directly in to the CMS; that way much of the noise is eliminated, it would mean counting goods-out rather than orders-in. But that takes us back to the old 'accounting and stock control' methods. Whereas the access-log format tends to be universal and gathered for free on the back of normal server operations, extracting information from within the particular systems of an organization requires the analyst to be on the inside. Building logging into a system from the start depends on knowing in advance what will be useful to know. All of which tends to slow progress and stifle insight.

web crawlers and other robots often account for 90% of all website traffic

Goods-out are not necessarily the same as orders-delivered, what needs to be discovered is what happens 'out-there'. This has led to a mass of cookies, tagging, and tracking: an acknowledgement is sent when the page is displayed to the user, cookies maintain state between page-views so visits can be identified and also repeat visits. This solves some problems but not all – in particular, if the required scripts and cookies are blocked there will be no record. An incomplete and in some cases biased sample remains. And, for the independent analyst, the problems of knowing what to record, or reverse-engineering the site structure and of being in a position to set up tracking in appropriate places.

The new way of evaluating digital usage: cookies, tagging, and tracking

There are very many products and techniques based around tracking, i.e. usage at the client side generating information that is sent back to the site operator or, more often, a third-party analytics provider. Usage is based not on access, nor on requests processed, but on the acknowledgement of data received. At the heart of these methods are cookies, tags, and trackers. Cookies are variables stored by the client browser or 'app' that accompany requests to the server. Fundamentally they overcome the limitation of HTTP that it is

a stateless protocol; cookies, exchanged and stored by client and server, join page-views into visits. If enough information is available, then those visits may be ascribed to a visitor. Tags are the snippets of code – in the case of a web browser usually JavaScript – that, often dependent on cookies, collect information on usage. Tracking code is yet more scripting that sends that information back to be analysed.

Among these trackers, Google Analytics, introduced in 2004, and developed from an earlier product known as 'Urchin',[5] is probably the best known and widely used. Google Analytics tracks web usage not by inferring page-views from page requests, but by sending a tracking tag signalling that the page has been displayed. The recording is at the end-point rather than the start of the usage cycle. And, because the tracking is collected by Google rather than the individual website, this gives both Google and, to a lesser extent, the website operator – in theory anyway – a more complete picture of user behaviour.

the latest developments in Google Analytics have greatly extended its power and ease of use

It is important to bear in mind in any assessment of Google Analytics that it is very much bound to a marketing perspective of online behaviour. Analytics is primarily there to measure usage for essentially two purposes: the effectiveness of advertising in sending traffic to a site, and the effectiveness of a website's 'conversion funnel'. That is, the web as a world of billboards, shop-windows, and checkouts. This narrow view of what online usage is for, of what the both the user and site operator needs and desires, may limit the relevance and effectiveness of the service when applied to the study of other kinds of online usage, such as that connected to scholarly information.

However, the latest developments in Google Analytics – Universal Analytics and Tag Manager – have greatly extended its power and ease of use. Google Tag manager was introduced in October 2012. Tags are tiny bits of website code that can help provide useful insights, but they can also create challenges. Too many tags can slow page loading and presentation; incorrectly applied tags can distort measurement; and it can be time-consuming to add new tags. Google Tag Manager consolidates website tags with a single snippet of code and management is from a web interface. Thus it is possible to add and update tags, without direct access to the site code.[6]

In October 2013, auto-event tracking was added to Tag Manager. This extends the recording of usage to 'events' – in-page activities such as form submission, timed intervals, clicks and mouse movements, downloading files, or playing a video.[7] Events can record clicks that lead to external links so that, in some cases, it is possible to say how the user left the site, something not normally recorded by logs.

Universal Analytics has tracking code in three varieties: analytics.js a JavaScript library for websites; Google Analytics SDK tracks mobile app; and the Measurement Protocol handles other digital devices, such as game consoles and information kiosks.[8] The new tracking code offers the ability to track a user's behaviour among different devices and is able to import data from online and offline devices. The Google claim that this new version will work better in a multi-device world needs to be treated with some caution: it depends essentially on the user having in some way logged on to a site and from more than one of their devices.

Whereas access logs can record all usage of a website, Google Analytics is limited because it depends on the co-operation of the user. Or, perhaps, it would be better to say the unwitting acquiescence of the user. The blocking or deletion of cookies or disabling of scripting will defeat or degrade the data available to Google Analytics. On the other hand by counting only pages displayed (implicitly viewed) the task of filtering irrelevant data is simplified. Web-crawlers and other robots ignore scripts; therefore, they return no usage data to Analytics. Because the Analytics script is included only in pages that are to be tracked the page ancillaries that clutter logs – images, style sheets, etc. – are ignored. The tracking code can also be elaborated to count not just page views but clicks on particular links and other user actions within a page. It will not always be possible to obtain a direct correlation of log and tracking data, sometimes we will be comparing apples with oranges. Neither source should be considered as more definitive than the other; the context matters. Data gathered via Google Analytics will give, for most purposes, as fair and accurate a representation of usage as any other method.

Piwik (piwik.org) is another example of a tracker, notable for being open-source and

making a direct appeal to those concerned by both user privacy and sharing data with a third-party analytics provider; hence its some adoption by libraries. Against this, however, must be set the 'network effect' advantage of using a more ubiquitous solution such as Google Analytics.

Key information behaviour metrics

Can a tracker-based approach portray digital information seeking and usage as well and fully as the log-based methods? In particular in regard to key performance indicators such as unique visitors, returnees, bouncers, and, critically for libraries/publishers, core users. This is a question that is far from straightforward to answer as the traditional methods were far from problem-free and arguably no longer as relevant as they once were. Let us take these established metrics one at a time.

Unique visitors

Unique visitors are, probably the gold standard usage metric, and widely quoted. The problem is that ever since the web broke the comfortable association of visitor and logged-on user there never has been any certainty about number of users. With access logging, a huge discount had to be made for robots. On the other hand, proxies and Network Address Translation (NAT) tended to under-count unique visitors. Cookies and tracking solve these problems, but, as mentioned earlier, cookies are often cleared and private mode browsing may be used. So, overall, tracking will tend to undercount unique visitors. But it gets worse: today there are so many web-consuming devices: phones, tablets, PCs at home and at work, that any count of unique visitors can only be considered a relative figure of audience level. Even then some allowance has to be made for the growth of the multi-device world: if many more people have smartphones, then that will show as a growth of visitors but may not mean more new or unique visitors, just the same visitors returning with new devices. 'Unique visitors' should of course read unique browsers. We know that many people use several (unique) browsers, finding the number of unique users is simply not doable, yet it remains a key metric with many information providers.[9]

Returnees

If it is not possible to count unique visitors with confidence, it follows that looking for their return will fare no better. Yet this is something which, for websites, signifies those desirable characteristics of stickiness, loyalty, and engagement. HTTP is by design a 'stateless' protocol, which means that there is nothing built-in that identifies a returning user even from one page request to another: thus, in practice, there is no such thing as a visitor, let alone a return visitor. From a raw unmodified log the best that can be done is to stitch individual log records together based on the IP address. This works as long as it can be assumed that each IP address represents a single device with a single user. Even then we need to be careful in assuming the persistence of the IP address. Hence the convention that a visit 'times out' after 30 minutes of inactivity. One user per IP, even for a limited time, can be difficult to justify: NAT is widely used to enable several devices to share a single IP on a home broadband connection, for example. In some cases matching both IP and user–agent string can resolve such ambiguity, but, in practice, it is rarely effective; if there are many browsers masquerading behind a single IP it is quite likely to involve the sort of corporate network in which all have identical user agents On the other hand in the most common case of a home network only the most heavily trafficked sites are likely to be accessed by more than one person at any one time. And, it needs to be stressed, none of this identifies a visitor or user; at best it may identify a browser or a device.

The 'stateless' design of HTTP is such a fundamental barrier to any form of transaction processing (i.e. any notion of a visit or series of linked page-views) that a means of working round this limitation was essential: this proved to be the 'cookie'. Cookies, even if not so called, pre-date HTTP and the WWW. A unique identifying token – a session cookie – accompanies each request from the browser. Stored by the server, subsequent requests from the same visitor can be reliably linked. Requests are chained together into a session: it is possible to build web-applications that involve a dialogue between browser and server rather than just isolated and unrelated requests. If the cookie is persistent, i.e. it is

cookies, even if not so called, pre-date HTTP and the WWW

stored by both browser and server for longer that a single visit, then it may serve to identify a visitor across subsequent visits. It is important to stress that the 'visitor' in this case is the browser rather than a real person.

Cookies have limitations, the most significant being you cannot rely on their being set and even less on their persistence beyond a session. Third-party cookies, sent to a server at an address other that the one visited are often automatically rejected, others are routinely deleted when the browser is closed. The result is that a visit from a single device can usually be tracked from page to page by cookies, but the return of a visitor after more than a few hours is far less certain. This becomes even less certain in a multi-device world where the same user may use a number of devices (desktop, laptop, tablet, and smartphone) to access the same resource.

core users can be considered as those who download content rather than those who only look at abstracts

Google Analytics cannot get around the fundamentals of this; it depends on cookies. The only real solution here is a return to the old pre-web world of logged-on users. Data in 'clouds', bookmark synchronization, Facebook, Twitter, Google+, Amazon, eBay: all in various ways invite you to log-on and stay logged-on and thus give a handle on the identity of the user. If the user is logged on to these services, if they 'like' or 'follow', if they click on advertisers links and do not routinely block third-party cookies, then some demographics may accrue. But mostly this information will only be of limited value to a website operator, the real value accrues to the advertising business that funds these services. As always in this context, the user is not the customer, the user is a product sold to advertisers.

Core users

Clearly everyone is interested in who their core users are, but we need to think clearly what is meant by the description. What, then, is a 'core user'? There is a need for caution when counting a returning user, cautious even when asserting precisely what we mean by a 'user' – so what can we say about a core? The definition cannot be based on the individuation of 'users', so it has to be an aggregation of characteristic usage or behaviour pattern. If it is not possible to say who core users are we may at least be able to say what they do or how. The core user can be defined as one whose usage is in some way what the site operator intended. If a digital library, like its physical counterpart, exists to match books to readers, then core users can be considered as those who download content rather than those who only look at abstracts.

This is an area where Google Analytics does offer some advantages because of the facility to track 'events': the possibility of measuring how much of a long (below the fold) page is read, the files that are downloaded, videos played, and forms filled.

Bouncers

The problem of bouncers, and these are the vast majority of visitors, is that they vanish without trace. Because there is only a first page, which is also a last page, any estimation of time on site is impossible from a raw log of page-requests. Google Analytics offers some prospect of illumination. Timing events enable an estimate of 'dwell time'[10] per page, so not only can bouncers be timed but it is also possible to have an 'adjusted bounce rate'. Intuitively there is a difference between a visitor who lands briefly on a page and one who spends time reading it. If that one page visit is exactly what the visitor wanted it makes no sense to discount it just because it is also a bouncer. Event tracking can also measure interactions with a page such as scrolling. Putting this together, viewing behaviour on a page can show how much of a long below-the-fold page has been read and the time taken. Bouncers can be redefined according to the degree of interaction with the content.

Event tracking also makes it possible to record outbound links so if the bouncer uses a single page-view as a springboard to another site there is the possibility of distinguishing that 'stepping-stone' pattern from a dead bounce.

Mobile users

A lot of development of the new 'universal analytics' focuses on providing analytics not just for web-browsers, but for Android and iOS (i.e. Apple) apps, i.e. smartphones and tablets. This is important as such users will soon constitute the majority of web users, but it also poses several challenges.[11] For all analytics whether reliant on HTTP-logs or Google Analytics,

there is a need to obtain some form of inside access to the website; it just is not effective to hope to interpret logs by reverse engineering the website; nor can Google Analytics be deployed effectively without the ability to set event tags (the new Tag Manager is really essential here); and mobile use increasingly means 'apps' rather than browsers. Which means there is a need to look at more than one form of presentation, of more than one set of 'logs', more than a single back-end process.

Demographic (user) information

This is a difficult area, and goes back to the beginning, the recurring user. Country does not mean very much in a cloud world; for HTTP-logs the IP is of limited use. Given the vast traffic flow through its servers worldwide it is reasonable to assume that even constrained by the limitations of IP addressing as the base, Google probably knows all there is to know. Browsers and even more 'apps' can send location data but increasing concerns over privacy mean such data will be partial and, possibly, unrepresentative. Regarding institutional data, it is just not there, unless we are dealing with old-style journal subscriptions, in which case the data has to come from publishers' internal data: it will not be in the logs or in Google Analytics.

Google Analytics integrates with DoubleClick, a subsidiary of Google which provides advertising services, to provide demographics: in principle it could show age, gender, and, possibly 'interests'. The problem is that, firstly, this data depends on the DoubleClick third-party cookie, and this may be rejected by many browsers. Secondly, the data is implicit, derived from which ads appear to have been viewed, etc. Finally, although it is hinted that this may be supplemented by data gathered from other sources such as Google+ profiles, it is uncertain how much of this data may be legitimately used given concerns over privacy, or indeed, if such data as users volunteer can be relied upon.[12]

In effect Facebook, Google+ Twitter, even Amazon and eBay, can be seen as attempts to privatize the web and encourage users to log-on, because that is the only way that the prized 'demographics' that everyone (especially advertisers) would like to possess can really be obtained. They are, of course, unlikely to ever pass this information on to anyone else.

New analyses and approaches

As mentioned earlier some of the old analyses were never very reliable, but a lot better than nothing; now they are practically impossible using the old methods. Logs still have a place but for many purposes they are no longer practical or cheap to obtain and analyse. Tracker-based methods have been in use for many years now, Google Analytics for a decade, but developments in the past year or so have tipped the balance in their favour.

Tag Manager is a really important development

Universal Analytics is still in 'public beta' but will replace the older version over the next year or so. The important feature is that it is designed to track usage, not just in web-browsers, but also by 'apps' in mobile devices. The challenge is that in order to understand the information gathered, researchers may find themselves not just having to reverse-engineer a web-page CMS, but an 'app' version as well.

Tag Manager is a really important development. Without it there is no way of setting up 'event' tags unless the analyst has direct access to web-site code. If there was such access then all the existing problems of reverse engineering would disappear. But life is not like that, the authors of this paper have found that just obtaining a simple change to a log configuration to record session cookies can take months. Obtaining a level of access that will allow the setting of tags may not be simple but it is at least a realistic, achievable goal.

Once deep diving into logs could provide insights at low cost. Now a new era is being entered in which it is practical to have deep access to the behavioural core of a website. Many things could be measured before, but the effort of implementation was far too great to justify the yield.

As already mentioned above (see 'Bouncers'), dwell time and scrolling are an example of the sort of thing that becomes practical and more robust when timer and click events can be set through tag manager. Dwell time, the length of time a user spends on a document is a strong indication of engagement and relevance. Link and timing tags can provide sub-categories of bouncer, distinguishing dead-ends from transits, the page that was just

scanned from the page gleaned. For long text, or a blog roll with multiple articles, setting event tags to record scrolling will show how much of the page has been read. Counting such events as page interactions produces an adjusted bounce rate. Perhaps, in many cases, the 'bounce rate' of 70–80% we often see is not really as bad as it first appears.

As to Analytics as a whole, although the 'real-time reporting' is possibly something of a gimmick, the ability to monitor activity and adjust the 'logging' as a situation develops is likely to lead to changes in the way that analysts work. They no longer have to spend months extracting log-files, which are outdated when they arrive. The mapping of the current state of the website to the observed usage is much easier if you can see both at the same time rather than work with historical records of a site since redesigned.

setting event tags to record scrolling will show how much of the page has been read

Summary case studies: Turnaways, CVCE, Europeana

Over a period of three years since 2011 the experience of working on a number of usage analysis projects has convinced CIBER that reliance on HTTP access logs as a primary source of web analytics data is no longer viable. During the same period the use of Google Analytics has become an industry norm and CIBER have had several opportunities to evaluate Google Analytics as a supplement to access logging. Three projects in particular have contributed to this perception. Although we highlight here each project with a particular insight, it should be stressed that all these factors are present in some degree in all cases.

Turnaways (2011–13), a study of scholars denied access from publisher platforms, confirmed that big corporate logs would never again be an easy informal source of data. But this was a more complex project in analytical terms and thus confirmed a growing sense that today's web services are too complex to make sense of from access logs alone. There is a 'reverse-engineering' problem, you have to infer from a log of incoming requests what would have been delivered to the client; you have to puzzle again over what the user did with that content based on subsequent requests. If you knew how the server processed the requests perhaps it would not be necessary to guess. But then, if you had that level of access to the server's CMS why not just build the logging and analysis into that CMS? The attraction of access logs was always that of an open resource, sitting there untouched by any preconception of what should be recorded; imperfect, unrefined but free of contamination and possession.

CVCE (2012), a project to develop new, robust key performance indicators and associated logging and reporting procedures for the Centre Virtuel de la Connaissance sur l'Europe, was not primarily a log analysis project; thus the resource expended on log analysis brought attention to the effort required to interpret complex multimedia web applications. This was not a project that presented problems in obtaining the raw data, nor were we overwhelmed by its volume – the assistance from knowledgeable personnel within the organization was excellent. Yet even with inside information available to understand content management a great deal of reverse engineering was necessary to make sense of access logs. After several months work it was clearly an ineffective and uneconomic approach: we settled for a less detailed view gained from Google Analytics.

Europeana (2009–13), a project providing an analysis and evaluation of users, usage, and information-seeking behaviour for the portal to European culture, was a case study in the growth of Google Analytics. At the project's inception a case was made for using Google Analytics but at the time the information was too basic and did not offer the possibilities of detailed analysis that could be derived from log files; besides as partners in the Europeana Connect project there was the potential to influence the format of the logfiles and obtain additional data from the CMS. Google Analytics was added to the Europeana site in early 2011, it provided marketing with easy to use graphic reports and basic usage statistics that matched the results of CIBER's log analysis. It also revealed one of the pitfalls of ready-made results: figures for duration of visit were disappointingly low. The analysis of log data found the cause: the dwell time had a log-normal distribution, the simple average presented by Google Analytics had overemphasised the huge number of fleeting visits over the long tail of longer but less frequent visitors. Just what counts as a visitor or visit is a complex

issue and access logs are not best placed to provide an answer but in this instance a combination of experience, insight, and detailed analysis was able to counter the temptation to naive interpretation of an attractive presentation. By 2012 the organizational obstacles to timely delivery of full logs were growing and we looked again at Google Analytics. The potential was there, but obstacles remained. In particular although 'event' tracking would allow much more detailed analysis than simply counting page-views, there was the difficulty of inserting the tracking code into the web pages. Our original hopes of a facility to influence the format of log files and logging within the CMS proved overoptimistic; the same problem presented itself in the case of event tracking: without access to the internals the analyst is passive and can only work with what is provided.

The introduction of Tag Manager has the potential to resolve this last difficulty as it enables events to be tagged as required without access to the full extent of the underlying content and code. The problems of negotiating the corporate maze and of authorized access remain, web properties grow ever more complex and thus require a considerable investment in reverse engineering by the independent analyst, but if event tags can be set and usage tracked from within the analytics interface, then many problems of effective and economic specialist analysis may be solved. It will go some way to restore the principal attraction of the old access logging: that like Google Analytics today it was usually there by default, even if not exploited; discovery could be deferred until the site was up and running; a skilled interpreter could take data for free and turn it into something worth knowing.

Conclusions

The big problem that faces us all, librarian and publisher, is that the old digital usage methods that served well enough for so long are just no longer effective or practical. Trying to analyse usage by working from a dump of pre-existing data – the basis of the raw log method – does not yield ready results the way it once did. The task of filtering out all the noise – bots and crawlers, fragments and frameworks, images and styles – and reconstructing a browsing history has become overwhelmingly complex and consequently unreliable. The CVCE project mentioned earlier, which sought to generate key performance indicators from usage logs, proved that this was not cost effective, even with the active assistance of skilled and knowledgeable people within the organization; after several months the effort to interpret logs was abandoned in favour of Google Analytics.

Google Analytics, the most readily available alternative to logs, does offer ease of use, smart presentation, 'real-time' views, and accessibility. It provides an easy way to put a toe in the water of analytics. If technical skills and domain knowledge are brought together it is a powerful tool for deep access to observe complex online behaviour. It has already been successfully used by researchers for the study of user behaviour,[13] website effectiveness,[14,15] and web traffic[16] and has been recommended by all these researchers. The only naysayers we have encountered in the scholarly sphere, who published a paper entitled 'Why Google Analytics cannot be used for educational web content' appear to come to this conclusion on the dubious grounds that educational use typically involves page views longer than 30 minutes.[17] But there is no reason in principle for tag-based methods to be so limited, and in practice Google Analytics allows session timeout to be up to 4 hours. Also, in our long experience in the field, we have established that hardly anyone spends more than 30 minutes on one page in the virtual environment. So the evidence against Google Analytics is flimsy.

the old digital usage methods that served well enough for so long are just no longer effective or practical

Overall there is no less information than could be obtained from raw logs and a lot less noise. The possible exception being that once upon a time there was an assumption that institutional usage could be tracked because people were using institutional resources to access the 'Net'. But that always was something of an anachronism: it only worked for a transitional period when old online access overlapped with new personal computing. Today affiliations can be tracked but they tend to group according to social networks. We are back to a world of logins but they are not as coherent or attributable these days.

Google Analytics provides an opportunity to evaluate really detailed behaviour, and that has always been difficult in the past using tra-

ditional web log analysis.[18] The benefit of the latest versions of Google Analytics is that they provide the tools to track behaviour (i.e. 'events'), to track on a variety of devices, and – with some reservations about accuracy and representation – to supplement this with demographic data. As important as all this, however, is the facility, through Tag Manager, to insert tracking code into a website without requiring extensive and direct access to the code. It is this possibility that opens up an opportunity for CIBER to combine domain expertise in libraries/publishing/academia, skills in data analysis, and knowledge of online technologies to provide a deep and rounded approach to understanding the digital consumer.

the limitation of Google Analytics from an information professional's point of view is that it sees the world in a perspective of e-commerce and advertising

The limitation of Google Analytics from an information professional's point of view is that it sees the world in a perspective of e-commerce and advertising. It has inbuilt assumptions about why people are online and their goals that may not really align to academic views, etc. It should not matter: dwell-time, a reference followed, serendipitous links – such elemental forms surely apply to all online behaviour. We can adopt the language of commerce and talk of digital consumers, our students can be customers, education and certification can be a product. Behaviour, acquisition, conversion; 'touch points with the brand', 'paths to purchase':[19] the language of marketing is beguiling but rarely nuanced.

Finally, while this paper comes to the conclusion that Google Analytics can pass muster in the scholarly communications world, it comes with an important qualification that this is only the case if it is deployed with some sensitivity and skill, and not treated simply as the analytical equivalent of a fast-food takeaway. This is particularly so in the academic and cultural heritage contexts where the audience and purpose do not always sit well with goals and metrics, which are largely intended for a web of advertising and e-commerce. The aforementioned usage research for Europeana proved conclusively that without the level of access needed to define 'events' and set tags it is not possible to use Google Analytics for proactive and investigatory research.

References

1. Turnaways. http://ciber-research.eu/CIBER_projects.html
2. Nicholas, D and Clark, D. 2013. Social media referrals on a multi-media platform: case study Europeana. *Journal of Documentation, Information & Knowledge*, 0(6): 9–22.
3. Usage, loyalty and sharing. http://ciber-research.eu/CIBER_projects.html
4. Key performance indicators for the CVCE. http://ciber-research.eu/CIBER_projects.html
5. Fang, W. 2007. Using Google Analytics for Improving Library Website Content and Design: A Case Study, Library Philosophy and Practice, Paper 121, http://digitalcommons.unl.edu/libphilprac/121/
6. http://analytics.blogspot.co.uk/2012/10/google-tag-manager.html
7. http://analytics.blogspot.co.uk/2013/10/no-code-required-auto-event-tracking.html
8. http://www.jeffalytics.com/introducing-universal-analytics-google-analytics/
9. http://www.theregister.co.uk/2014/01/22/reading_this_headline_you_and_947m_million_others/
10. Liu, C., White, R.W., and Dumais, S. 2010. Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR ‹10). ACM, New York, pp. 379–386. http://doi.acm.org/10.1145/1835449.1835513
11. Nicholas, D. and Clark. D. 2013The second digital transition: to the mobile space — an analysis of Europeana. *Learned Publishing*, 26(4): 240–252.
12. https://support.google.com/analytics/answer/2700409
13. Crutzen, R., Roosjen, J.L., and Poelman, J. 2013. Using Google Analytics as a process evaluation method for Internet-delivered interventions: an example on sexual health. *Health Promotion International*, 28(1): 36–42.
14. Plaza, B. 201). Google Analytics for measuring website performance. Tourism *Management*, 32(3): 477–481.
15. Turner, S.J. 2010. Website statistics 2.0: Using Google Analytics to measure library website effectiveness. *Technical Services Quarterly,* 27(3): 261–278.
16. Plaza, B. 2009. Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series. *Aslib Proceedings*, 61(5): 474–482.
17. Dragos, S.M. Why Google Analytics cannot be used for educational web content. In *Next Generation Web Services Practices (NWeSP), 2011 7th International Conference*. IEEE, 2011, pp. 113–118).
18. Hess, K. 2012. Discovering digital library user behavior with Google Analytics, Code 4 Lib Journal, issue 17, http://journal.code4lib.org/articles/6942
19. http://analytics.blogspot.co.uk/2013/11/full-customer-journey-three-lenses-of.html

David Clark and David Nicholas
CIBER Research Ltd
1 Westwood Farmhouse, Greenham
Newbury, RG14 7RU, UK
Email: david.clark@ciber-research.eu,
Dave.Nicholas@ciber-research.eu

Hamid R. Jamali
49 Mofatteh Avenue
Department of Library and Information Studies
Kharazmi University, Tehran, Iran
Email:h.jamali@gmail.com