Web statistics: how to collect them, can we trust them? *(figure 1-title)*

Presentation for Library Science Talks, march 27th and 28th 2006, Geneva and Bern
www.gaa.nl/support/statistics

Good afternoon ladies and gentlemen. I have been invited to answer two questions for you this afternoon. The first one is: How do we collect web statistics? and the second one: Are web statistics to be trusted? The first question is not so difficult. I am sure that you will understand the basics of the technology behind webstatistics when you leave this room in an hour or so. But I am not quite sure everybody will give the same answer to the second question. We will see.

But before starting, I would like to thank the three organising institutes for the invitation for this presentation. I realy feel greatly honoured to be invited to speak on a websubject in the city where in 1989 the WWW was born. (as a dutch archivist I feel greatly honored to speak for the National Library of Switzerland)

Ok, lets get started. I have to introduce you to some technical concepts, such as logfiles, robots, and referers. Don't worry, it sounds more frightening than it is. It does not take more than about twenty minutes, and then you know how web statistics can be collected.

Log files

First of all I am going to explain to you about log files, because the logfile is the fundament of webstatistics. Log files are files containing data about site visits. Thinks of the german word 'Logbuch' or 'Schiffstagebuch'.

This is how it works. *(figure 2 – diagram)*

If someone wants to view a page of a website, in other words if he wants to '*visit* a website', his computer sends a request to the web server the site is on, and this server then sends the requested page to the computer of the person who asked to view the site. The web server with the site on it, then appends a line with information about this request in what we call a log file. This information therefore relates to the visitor and that page. These are the primary data used to produce web statistics — data on visitors and requests for pages. Let me show you.

Here are the log files of the Amsterdam City Archives site. *(figure 3 - live log file)*

What you are seeing now is live. The movement of the lines means that someone somewhere is looking at a page on the site of the City Archives. In no time at all this grows to many gigabytes with complex and unwieldy information. That is why you need a statistics program that processes the data to produce clear and understandable graphics and tables, and takes care of filing the information.

This for instance is one of the graphics our statistics program produces out of these log files. It shows you the amount of pageviews per day of the first quarter of 2006. *(figure 4 – pageviews per day)*. These dips in the lines are weekends, people have always visited our site more during the week than in the weekend, though in recent years, with more people having fast internetconnections at home, Sundays have become more popular. These two peaks are days at which new items on the site were introduced.

To understand where the data of these graphics come from we have to go back to the logfiles. We will now take a closer look at a line from these log files. *(figure 5 - line from log file)*

The first thing you see is the IP-address of the computer from which the request for a page from our website was made. Then you see time and date of the request. The next thing is the url of the page which is requested. It is a page from a part of our website that is called Schatkamer - Treasure Trove, 300 schatten – 300 treasures, macht – power, geheimschrift – secret code.

And this is information about the webpage on which was clicked by the visitor to get to our website. In this case it was Google. The owner of the computer with this IP-address used Google to search for “Coenraad van Beuningen” (which is the name of a historical figure in Amsterdam). Google produced its well known list and at the moment this person who is looking for “Coenraad van Beuningen” actually clicked on the item in the Google-list that refers to our site he became a visitor of our site. And then this file of information was stored into our logfiles.

Let me show you how it works: *(figure 6 - gaa.nl/...coenraad_van_beuningen).* This is the page on the website of the Amsterdam City Archives we are talking about, a page about a letter written in a secret code from a burgomaster called Coenraad van Beuningen who was negotiating with the enemy in times of war.

Now we will switch to Google *(figure 7 – Google.nl).* This is the dutch version of Google. I enter ‘Coenraad van Beuningen’, I see the list with a total of about 14.000 hits and decide to look at 4th item on this list, the website of the Amsterdam City Archives.

This action that I have done just now, is registered in our logfiles in a way comparable to the registration we have just seen. I will show it to you again . *(figure 8 – line from log file)*

In these last few lines you see some technical information such as the visitor’s browser (Safari) and his operating system (Macintosh). This information is important too, but I won’t go into that now. In this meeting I want to concentrate on the information about visitors and pageviews, because these are the two basic elements of webstatistics.

We will now take a closer look at the concept of a Visitor.

Visitors — IP address, cookie

The first piece of information from our log file relates to the visitor, or, as you know now, to the IP address of the computer from which a page was requested.

There are two ways to count visitors. The first method is the one based on IP addresses. An IP address is a numerical code and every computer connected to the internet uses an IP address. If you count the number of different IP addresses from which a page on your website is requested, you will apparently know how many different visitors you have.

However, there is a problem here. On the one hand, a computer network with a huge number of visitors can be using one IP address. For instance, all visitors to our website of the network of the university of Amsterdam are counted as one visitor only. And of course there are many different people, different 'visitors', visiting our website from the university every day. But they would be counted as only one 'visitor' in the statistics IF only the IP-address would be used to count them.

On the other hand, in the case of dial-up connections a different IP address can be allocated to a computer for each session. So each time one person, one computer, dials in to a service provider, this provider may assign a different ip-adress for that particular session. And this one person would be counted as a 'visitor' as many times as he is connected, IF the IP-address would be used to count.

Therefore, we need a more accurate method for collecting information about visitors. This method uses cookies.

This is a sample cookie from our site. *(figure 9 – cookie)* It is a relatively simple cookie, consisting of 9 digits. But it is enough for statistical purposes.

Every page on our website contains an instruction to the browser to store a totally random number on the visitor's computer, unless one is present already. On each subsequent request the visitor makes to our website, the browser will return that same number to our webserver, which then recognizes this visitor as a unique, "returning" visitor.

Virtually all large sites use cookies, and a commercial site in no time puts a whole jarful of cookies on your computer. But there is no need to worry because your computer can handle huge numbers of cookies, since they are very small.

However, some visitors instruct their computers to refuse cookies because they consider them to be an infringement of their privacy. And, I must admit, they can be used for that goal. But the Amsterdam City Archives doesn´t. We have made clear on our site that we conform to the P3P protocol for privacy developed by the W3C. W3C stands for the World Wide Web Consortium. P3P stands for Platform for Privacy Preferences.

Furthermore we have made our webstatistics accessible on the site to everyone and we provide extensive comments on how they function. *(figure 10- visitors per day per quarter, show navigation and introduction)*

To count visitors the web statistics of the Amsterdam City Archives are based primarily on the analysis of cookies and secondarily on the analysis of IP addresses. Some 60 to 70 % of our visitors allow the use of cookies. The other users are identified by the IP address. Most professional statistical programs only count cookies. Simpler programs such as webalizer use IP addresses.

Robots

There is one other important factor which can really distort the number of visitors in statistics, or perhaps you could call it 'sex up' the amount of visitors for fans of big numbers, because the counts of visitors and page views could easily be increased by 20 to 40% for our site. For some other sites the percentage of the increase could amount to hundreds. That depends on how big the site is, and how many visits are paid to the site.

I am talking about search engine robots. They are programs, operated from search-engines, Internet archives, or other forms of electronic intelligence gathering. They are called non-human agents, or robots, or spiders. They automatically request pages on the entire web at high speed. Should your statistical program not recognize these as robots, and should search engine robots be included in the count of visitors, it makes a big difference.

This for example shows the difference of one single day *(figure 11 – visitors and robots on March 7th)*

On the day I wrote this presentation for exemple, on March 7th, 2,820 visitors came to our website viewing a total of 43.234 pages. This count EXCLUDES robots. On that same day 30 robots paid us a visit, and jointly looked at 5.787 pages. So especially the number of pageviews would therefore have been significantly higher had we included the robots.

This is a list of robots that frequently visit our site, I show the list just to make you acquinted with them, so that next time you see one of them mentioned in an article on the web, you recognize it. *(figure 12 – list of robots)*

So, be sure whether your count of visitors includes or excludes robots. By the way, this does not mean your website should not be welcoming to robots. On the contrary, a well built website is optimized to be visited and indexed by robots. Otherwise you will miss a large potential audience. Just be sure they are not counted as ‘ normal visitors’ in your statistics, at least if you do not want to fool yourself.

Page views

Ok. That is enough about visitors for the moment. We will now move on to the concept of pageviews. At the beginning of this presentation I said that there were two core concepts associated with statistics — visitors and pages.

Counting pageviews we encounter the same sort of problems as we have to deal with when counting visitors. There are just as many ways to count pages as there are to count visitors. So you have to be carefull as well with the amount of pageviews statistical programs offer you.

Even the concept of a ‘page’ is interpreted differently by the different statistical programs. So, although the meaning of the word ‘pageview’ seems to be obvious, it is necessary to ask the question ‘What is a pageview?’

To begin with there is a distinction between ‘hits’ and ‘page views’. A page as you see it on a website can consist of many elements: text, images, animations, javascript code. Each one of these elements produces a “hit” in the logfile. Let me give you an example. Here you see the homepage of our website. *(figure 13 – homepage gaa.nl)* This homepage is just one page but it is composed out of 20 elements. So it produces 20 “hits” in the logfiles. It is obvious that one gets a significant difference counting hits or counting pageviews. The number of hits gives you more information on how a site is built, than on how a site is used. In short, for statistical analyses, hits DO NOT MATTER. When a huge amount of hits is being mentioned in a newspaper, don’t trust the article. Fat chance the journalist who wrote the article does not now much about webstatistics and is more interested in big numbers than in the actual use of websites.

There are a lot more factors that influence the counting of pages. How, for example, do you count the use of the browser’s back button? Is that a new page or is it not? What if you click on the refreh button? What to do with pages that uses anchor links or others clickable options. To give you an example. In our website an inventory that can consist of a few thousand items is considered as only one pageview. And how do you count if the information changes within a single page? This happens when the web designer uses JavaScript, Flash or video technology, for instance. If you spend ten minutes playing a game, in your perception you have visited a lot of pages, but for statistics they may only count as one page.

So the figures of pageviews ALSO need to be used with a degree of caution. Do not attach too much value to them. As most customers like huge numbers, statistical programs tend to count in what you could call a ‘rich’ way. The question ‘What is counted as a pageview?’ is seldom asked or answered.

By the way, it is not true that the more pageviews a site gets, the more popular the site is. Did you know that a large number of page views is actually no compliment at all for a site? If a site is well designed and well built, you get the information you want quickly and without unnecessary clicking. It is not so difficult to build sites that are guaranteed to deliver many pageviews. All you have to do is make sure that the visitor to the most popular parts of your site has to click a lot in order to get his information and you are sure to end up with huge quantities of page views.

You might even give your visitor RSI. So far I have not heard of an internet user, not even in America, who has sued the owner of a badly built website because he has developed RSI, but I would not be surprised if it happens one of these days.

On the other hand it is similarly not good for the usability of a site (usability is web jargon for user-friendliness) to build hugely long pages. When building our image bank for instance (*figure 14 - imagebank, IJbeeld*) we had to decide how many pictures per page to provide when presenting search results. Searches in the image bank can lead to hundreds of results and you cannot put them on one page. We opted to present nine pictures per page. But if we had chosen to present say six or twelve pictures per page for example, this would have had a significant impact on the number of page views each day.

Referrers

Before rounding off this first part of my presentation I would like to explain to you one more concept of web statistics. That is the concept of the referrer. The referrer to a website is the url*)* of the website from which a visitor came to your website. If the visitor typed the url of your website into the location bar of his browser, of if he used his own collection of favourite websites, then the referrer is empty, but if he clicked on a link in another website or if he used a search engine, than the other website or the search engine is the referrer. *(if asked: url=uniform resource locator, first part indicates what protocol to use, second part the specifies ip address or domain name where resource is located*

In my opinion the list of referrers is one of the most interesting overviews your webstatistics can provide. It gives you information on the origin of your visitors. And what is more, the referrer enables you to see not only how many visitors came from search engines, you can even see which search query on a search engine brought the visitor to your site. Do you remember the line of our logfiles I used as an example? *(figure 15- line from log file).* The referer in this case was Google, and the query which brought the visitor to our site was ‘Coenraad van Beuningen’.

Approximately 75% of the visitors to our website arrive via another site, in other words through a search engine or through a link from another page. Let me explain to you. *(figure 16 – amount of referrers on March 7th)*

On March 7th, a total of 3,085 times a link pointing to the GAA website on 290 different external webpages on 234 different domains have been clicked by a total of 1,998 visitors. These visitors represent 70.9 percent of all visitors of de GAA website on March 7th.

This is our referrer list of yesterday. *(figure 17 - referrers).* Let’s scroll through it. As you see, google is on top of the list. In this list if the referrer is a search engine it is green. These are the search queries people entered, and how many times the query was entered . As you see, a lot of people use google as startingpage. They don’t bother with favourites anymore. Just enter a name of an institution or site you remember and google brings you its website. Blue in our referrerlist means that the referrer is another website with a link to our site.

A referrer list gives you very valuable information on the background of your visitors. You can keep an eye on the queries in search engines that bring people to your site and you can see whether or not you reach your desired target groups. I think the referrer list is the most valuable item of your webstatistics, as this list gives you not only information on numbers, but it gives you information on matters of content.

I think it is now time to round off the first part of the presentation. I am sure you understand now how the basic principles of how information on visitors and pageviews is collected. The rough data are gathered in logfiles and your statistics program transforms these data for you into overviews, graphics and tables. Most used data are the numbers of visitors and pageviews, most valuable data (according to me) are the data on referers and search queries.

Now for the second question:

Are webstatistics to be trusted?

The answer to this question is twofold. The first answer is NO, don’t trust webstatistics. You have just learned that it is nearly impossible that two statistics programs count visitors and pageviews in the same way. There are too many different definitions of the concepts of pageviews and visitors. So especially don’t trust a comparison between different websites on account of their numbers of visitors or pageviews.

It is dangerous to compare websites on account of numbers, but people love comparisons, and if you really want to compare websites, use a tool that measures internet trafic using exactly the same software. Such software is currently available in the form of toolbars. There are many of them. I recommend that you take a look at Alexa. Alexa is the oldest and probably the most accurate among them.

Alexa.com (*figure 18 – alexa.com*)

Alexa was founded in 1996 and it was subsequently taken over by the well-known on-line retailer Amazon. Alexa's goal is to measure internet traffic worldwide on the basis of voluntary participation by as many internet users as possible.

Let's have a look at the information on worldwide internet traffic Alexa generates (*figure 19*). This is the ultimate 'benchmark': the global internet top 500. It starts with Yahoo, then MSN (the popular hotmail is part of this), then comes China, Google, and so on.

To see Alexa's rating for any other websites not included in these lists, you can click on "traffic rankings" and enter the address of a website. *(figure 20 alexa.com - enter cern.ch)* I have done this and now we will have a look at Alexa's statistics for the website of Cern. In the toolbar you see Alexa's ranking: at this moment the traffic rank for the cern website is 10.483, which is very high. Yesterday there were 9543 visitors. *(enter snl.ch)* The traffic rank of the Swiss National Library is 474.320 with an average of some 800 visitors a day. If you click on 'Learn more about traffic' you get an explanation on how Alexa calculates pageviews and how they define visitor reach.

I suggest you have a closer look at this later, as I want to come back once again to our question.

Are webstatistics to be trusted?

Knowing in how many different ways countings can be made, the conclusion seems obvious, don't not trust webstatistics. But I must say, my arguments can just as easily be used the other way round. Ofcourse you can trust your own webstatistics. Why not? You know how it works now. And of course you can compare the data from the same program form one day with the data from another day, from one year with another year. And, what is more important, your webstatistics give you much more information, and more *valuable* information, than just the countings of visitors and pageviews. So a more interesting question perhaps is:

What can be learned from webstatistics?

I would like to tell you about the most important lesson the Amsterdam City Archives have learned from their webstatistics a few years ago. You remember the referrerlist I showed you? Here it is again, this time for the first quarter of 2006. *(figure 21 – referrers per day, per quarter)*

In this list we can see that most of the visitors of the Amsterdam City Archives come to the site by using a search engine. Visitors came from a total of 142 different search engines, but most of them come from Google. When I add up the amount of visitors from google.nl, google.com and google.belgium this comes to more than 100.000 visitors from the total of about 138.000 visitors this quarter. That means that no less than about 70 % of our visitors came to our site through Google.

It has not always been like this. In 2001, which is the first year for which we have statistics for our website, more people came to our site via portals than via search engines. In 2001 not more than 34 visitors were counted coming via google.nl. This changed dramatically in 2002, when about half of our visitors came from Google.

These figures prompted the Amsterdam City Archives in 2003 to take the decision that every new part that we build in the website must be completely searchable by Google. Most webdesigners will tell you that this is not possible for the databases on your website, but that is not true. We know this because the databased search engine of Survey of Archives and Collections and all our Inventories, built after this decision was taken, are indexed by Google.

Let me show you *(figure 22 – google.nl)*

Let's say I am looking for information on one of Amsterdam's mayors. His name is Wim Polak. Google scores 67.500 results, but we all know that people browse through not more than some dozens of results. Then you have found what you are looking for, or you give up. So you want your website to be on the first pages of Google. With Wim Polak we are at the top of the list, and that is because google has located his name 19 times in the inventory of his archive. *(choose 2nd item Google)*

The inventories are only available in Dutch, but I am sure it will give you no problem to find the name of Wim Polak. We have highlighted the search term you started with in Google.

You could argue that perhaps every serious investigator should have found his way to the City Archives when he is looking for the files of an Amsterdam burgomaster. That is true. But we also have collections that no one would think of. For instance the archive of an international organisation for the protection of nature that has its head quarters in Amsterdam. And on one of the first days these inventories were online, we got a request from Dutch television for documents from this archive, because just then the remains of a sort of apeman were found in Indonesia and they discovered that there was a document in the archives of this organisation from 1932 in which the same finds were discussed.

And this document was found thanks to the fact that our website, including its databases that are added after 2004, is indexed by Google. By the way, do you know how to check if Google has indexed your site? Let me show you:

(figure 23 –Google intypen: site: gemeentearchief.amsterdam.nl - from our site 60.190 pages are indexed by Google – you want to know for Cern of SNL? – Cern: 10.300.000 it's a huge conglomerate of sites, over 500 I saw at netcraft (news.netcraft.com) all well built to receive robots, SNL: 14.800, could be better perhaps. National Library of the Netherlands counts more than 2 milion pages in google.)

So, always ask your webdesigner to garantee you that what he builds can be found by search engines. And don't accept no for an answer when you want this to be done for your databases as well. Ofcourse technically it is not at all easy to realise this, but even I can explain to you the principle of how this is done, so every self-respecting webdesigner should be able to realise it.

(do you want me to explain this to you now, or do you prefer to go on with another lesson to be learned from webstatistics?)

The key to the problem is that the website holds not one but two different versions of the data. One of them is structured in a way that is optimized for robots of search engines, and one is optimized for presentation to the public. The version for search engines is not suitable for the human eye. Let me show you a page from an inventory optimized for search engines *(figure 24)* This is what search engines love: a text with nearly no lay-out al all. It does not matter at all how long a page like this is. Computers have no problem with enormous quantities, but they do have problems with changing interpretation. Do you remember what I told you at the beginning? A website is hosted on a server and when your computer asks for the inventory of the archives of Wim Polak our website sends to your computer the inventory you saw at first.. When a robot of search engine asks for the same inventory, the server recognizes that the computer that made the request is not an ordinary computer but a robot and the server sends this version of the inventory. When you enter the words Wim Polak in Google, you get the famous google-list with all the urls of websites and when you choose our website then Google sends your request to our server. Our webserver recognizes that your computer is not a robot, so he gives you the text in 'human lay-out' but he also recognises your search entry in Google and then he highlights for you the words you originaly entered as search in Google. And all this is done in less than a second.

And it was the really fast and dramatic growth of the importance of search engines that we saw in our own webstatistics before we read about it in the newspapers, that convinced us of the need for the website to be absolutely friendly to search engines. I think this is the most important thing in the past we learned from our webstatistics.

The second lesson I want to tell you about has not been learned yet, but I hope we will in the course of this year. This year the website of the Amsterdam City Archives will undergo a complete redesign, and one of the most important changes will be a change in navigation and the homepage ofcourse is what you could call the 'front door', the startingpoint of navigation for a lot of (first) visitors. And when you stand in this doorway, it should be absolutely clear which way you have to go. But there are many different visitors, different target groups, and the homepage is to be a front door for all these groups. They all prefer to find the information they are looking for very near this frontdoor, if possible directly IN the doorway. But then it can get crowded there, with the risk that nobody finds his way easily.

Let me show you the homepage of the National Archives of England *(figure 25 – homepage National Archives - London)* By the way this is a site I admire a lot, when I think of something I would like to change in our own website, most of the time I find out that it has already been done in the website of the NA of England. But their homepage....., hmmm...., navigation bar, site search, experienced reseachers, new visitors, most visitors, newsitems in two different ways, three changing windows, all of it clickable, ... hmmm,..... I don't know....perhaps a bit crowded. But my opinion is not important, what do visitors think about it? And more important what links on this homepage are really used by visitors. Webstatistics can help you find an answer. We will experiment with changing navigation-options on our own homepage and monitor the use in the hope it will help us decide when designing a new homepage.

I think it is getting time to come to a conclusion now. I could go on telling you about the benefits of websatistics, but I am afraid then you won't be able to remember what this presentation was about in the first place. So I will summarize. At the beginning I have explained how the data can be collected on which webstatistics are build. I have introduced you to some of the basic concept of webstatistics. And I hope now you know enough about webstatistics to be able to decide for yourself in which context webstatistics are to be trusted or not. Most of the time a simple bit of common sense is enough when you know how things function.

Let's take a little test to see if you feel sure of yourself now, judging whether a site is to be trusted or not. (figure 26 – site olympic.org) Take a look at the first two paragraphs of this article about an olympic record: What information do you think is to be trusted and what is not? – exactly: the information on hits in the second paragraph is absolutely superfluaous.

New Daily Record for www.olympic.org

© IOC

19 February 2006

On 13 February 2006, the Olympic Movement's web site - www.olympic.org - reached a new daily record-high of 458,065 visitors. This figure surpasses the record set during the Athens 2004 Summer Games when 290,000 people visited the web site on its peak day. This is an increase of 57 per cent. The record number of visitors also resulted in a new overall record for page views: the 458,065 visitors consulted 2,875,137 pages.

Record for Organising Committee Web Site

The web site of the Turin Organising Committee for the Olympic Games - www.torino2006.org - also registered a new all-time daily record of 50,300,630 hits on 16 February 2006. This figure can be compared to the daily record high during the Athens 2004 Games (slightly over 50 million pages viewed) and the figures from the last Winter Games in Salt Lake City (around 13 million pages viewed in one day).

News and More

During the Turin Games, several news stories are posted every day on the homepage of www.olympic.org. Visitors are invited to enter a virtual Olympic village in Turin with 3D images to represent the main landmarks. The site also offers photos of the action and behind the scenes in Turin, including a Kodak-sponsored "photo of the day". As a reference web site for the Olympic Games, the site provides information, results, photos and videos on all the Olympic Games since 1896 and the outstanding performances of over 400 athletes. A special section is dedicated to the Olympic Museum in Lausanne.

Win a Video Game

Visitors to www.olympic.org can also test their sporting knowledge and take part in the competition to win one of 250 TORINO2006 video games. One of the questions to be answered is: Which athlete won the most gold medals during the XIX Olympic Winter Games in Salt Lake City in 2002? Was it Kjetil Andre Aamodt, Ole Einar Bjoerndale or Janica Kostelic?

http://www.olympic.org/uk/news/olympic_news/newsletter_full_story_uk.asp?id=1699