Web statistics of heritage institutions

Summary of a survey held under the authority of Digitaal Erfgoed Nederland
This English summary has been written to support the work on Europeana
* Henk Voorbij and Bauke Freiburg / Web statistics: backgrounds, feasibilities and pitfalls, 2009
* Henk Voorbij / Usage of web statistics at heritage institutions, 2009

Increasingly, cultural heritage institutions provide access to digitized resources on their websites. By using web statistics one can get an impression to what extent these resources are being consulted. To a lesser degree, physical numbers of visitors offer a good idea of the usage of an institution. In addition data about the usage of digital resources are increasingly required.

Under the authority of Digitaal Erfgoed Nederland a dual survey of web statistics was conducted. Firstly a literature survey was carried out to obtain insight in the backgrounds, feasibilities and limitations of web statistics. Secondly a statement of affairs was drawn up on 112 cultural heritage institutions to ascertain to what extent these institutions dispose of web statistics, which software packages are used for that purpose, and which statistics are considered useful for mutual comparison. An interview was held with several of the institutions. Finally annual reports of heritage institutions were studied to ascertain to what extent they make reference to web statistics.

Web statistics offer insight in the usage of a website. Among other things, the statistics show how often the website is visited, which pages are most viewed, via which pages people enter the website and via which pages people leave the website, how many different visitors there were, from which countries or domains they originated, how the visitors arrived at the website (via a search engine, via a hyperlink on another website, or directly) and during which hours or days the traffic is heaviest. More than a hundred software packages are available to generate these statistics.

An important distinction exists between log file analysis and page tagging. Log file analysis implies that all transactions which have been recorded in the server's log files are converted into ready-made statistics using software such as AWStats. Page tagging requires that a small bit of Javascript is inserted on each page of the website. Each visit to these pages is registered and converted into statistics. The best known program is Google Analytics. A restriction of logging and tagging is that the possibility of distinguishing visitors is inadequate. However, cookies do offer this facility. In practice particularly applications based on page tagging use cookies.

The most significant differences between logging and tagging are:
– Page tagging requires that pages are equipped with a script. This may entail an increased workload. Log file analysis does not require any extra workload in this respect: the necessary data is already available
– With page tagging the data is usually stored on an external server. This does not require storage space, but permanent availability is not guaranteed. Moreover, usage for other – less desirable – purposes is not excluded. With log file analysis the data is available on one's own server.

- The reliability of statistics based on log files depends among other things on the following factors: to what extent can the program filter out visits by robots, do visitors arrive via a proxy server, do visitors have dynamic IP-addresses or are previously viewed pages being called up from the cache. The main problems when using page tagging is blocking or deleting of cookies. In both cases the identification of individual visitors is not perfect: different users could be seen as one person, and conversely one user is sometimes seen as different persons.

Points of particular interest when interpreting web statistics are:
- The programs do not provide absolute values, but do offer insight in the development of the usage in the course of time.
- It cannot be simply declared that a higher score is preferable. A high number of page views per visit and a high average duration per visit could indicate that the visitors find the website interesting enough to view several pages, but it could also indicate that the navigation has shortcomings and that the required information is hard to find.
- The scores are heavily influenced by definitions. A visit to a website is usually deemed to have ended after thirty minutes of inactivity. There is no guarantee whatsoever that the visitor actually viewed the page for as long as the timekeeping indicates. Perhaps he left some time ago, but may have forgotten to close the window. The number of unique visitors also depends on the definition: a visitor is unique within a certain reporting period. It makes quite a difference whether this is a day, a month or a year. The number of unique visitors per day or month may under no account be cumulated to a yearly total.

To obtain insight in the usage of web statistics by heritage institutions in the Netherlands, a questionnaire was sent by email to 112 institutions. These institutions participated in an earlier survey held under the authority of DEN and are possibly more than averagely orientated towards digitization and information technology. The response was 47%. The most important conclusions are:
- Most institutions offer digital reproductions of their material collection on their website (84%)
- Most institutions collect web statistics (78%)
- The institutions use a huge variety of software packages to generate web statistics. Google Analytics is the most popular package, followed by AWStats.
- More than half of the institutions deems the following data of importance for mutual comparison: the number of sessions (visits), the number of unique visitors, the average duration of the visit, the number of pages viewed (page views), the number of new or returning visitors, and the number of visits by origin (via search engine, via URL, directly).

The following was obtained through interviews with thirteen of these institutions:
- The degree of insight into the backgrounds of web statistics differs considerably.
- There is a need for a liaison committee where heritage institutions can exchange experiences on web statistics.
- In nearly all cases, the data which is offered by the packages is adopted without adjustment. Some institutions deem adjustments useful to arrive at more reliable data, however one institution holds the opinion that even these efforts will not lead to reliable data and that one should have an eye for trends.
- Seven institutions analyse the web statistics for practical purposes, such as adapting the website, establishing priorities for (further) digitization, or attracting more visitors.

- Nine institutions consider using the website to be an important performance indicator, if only because physical visits are decreasing.

Finally we checked whether the 112 institutions indicate the web statistics in their online annual report and if so in what way. An annual report was found online of exactly half (56) of the institutions. Analysis of the content yielded the following results:

- 37 institutions (66%) incorporate web statistics in their annual report. This is certainly valid for archives and museums. (University) libraries offer comprehensive figures on the usage of external, digital files.
- The most employed indicators are the number of visits and the number of visitors. At times the obsolete, misleading notion of hits is still employed.
- It is awkward to interpret the data. Often an explanation of the definitions and measuring methods are lacking. The importance of the number of (unique) visitors is unclear when it is unknown whether these are being counted per day or per month, and how these numbers have been cumulated to a yearly total.
- Physical visits are usually measured in terms of visits, not unique visitors. That is perhaps the reason why the notion of visits is a more relevant indicator than the notion of unique visitors for a heritage institution.