MEASURING THE WORLD-WIDE WEB

Denman Maroney, DMB&B Interactive

How the Web is measured

The World Wide Web is measured in two basic ways. Colloquially and not altogether accurately these ways are known as site centric and user centric. Site centric Web measurement is based on Web server access logs. User centric measurement is based on a sample of Web users. A hybrid method in development combines site and user centric methods.

Web measurement services

The major site centric measurement services are Nielsen/I-Pro, NetCount and Intersé. In addition to these there are dozens of companies at this end of the business. Interestingly, the largest share of Web analysis is done in-house by sites using systems of their own design. On the user centric side there are only two services — NPD Media Metrix (formerly PC Meter) and Relevant Knowledge — of which only one — NPD Media Metrix — is operating commercially. Nielsen/I-Pro is the only company working on developing a hybrid Web measurement system. I know what they're up to, but I'm not supposed to tell you about it.

What Web servers log

To understand site centric measurement all you have to know is what Web servers log, since what isn't logged can't be analyzed. A Web server log is a plain text file. Each line or record of that text is a series of fields, which represent the request source, information source, request time, status code, browser type and referring URL of every client/server interaction. In a perfect world this would mean that each record shows who asked for what, when, and by what means. Unfortunately, most of these items are ambiguous from a media research point of view. "Who" is not a person but a computer that may or may not and usually does not represent an individual human being. "What" is not a publication but a file representing an element of a page of a publication, whose exposure to a Web user may or may not result from a request logged by a server associated with a publisher trying to sell advertising. "When" is clear; computers are very dumb, but even they can tell time. "By what means" is what kind of browser, which is clear but rarely useful. The referring URL is nice to know, but often there isn't one, or it can't be known.

Request source

The request source represented in a server log is a unique IP or Internet Protocol address. An IP address is a series of four groups of three positive integers 0 through 9 separated by dots. To make this more intelligible for purposes of analysis, in some cases it is possible to convert it into what's called an Internet domain by looking it up in what's called a DNS look-up table. For example, looking up this address might reveal that it corresponds to the domain aol.com. Unfortunately that domain represents some 8 million individuals. Which one is not revealed by the log. In other cases a given IP address corresponds to no known domain.

The .com part of the domain aol.com is called the Top Level Domain name or TLD. Other TLD's include .edu, .org and .net. Respectively these TLD's reveal that a request source was educational, non-profit or Internet based. Sometimes this information is useful. Unfortunately the vast majority of domains are .com, so TLD's tend to be not very discriminant. The number of TLD's is about to be expanded. This will make them certainly more discriminant and hopefully more useful. The AOL part of aol.com is called the Second Level Domain or SLD. Some SLD's can be mapped to SIC (Standard Industrial Classification) codes. This information can be useful as well.

Information source

The information source is a file request. Web information is served up in pieces called pages. For example, here is the Web home page of Mindspring, my Internet Service Provider. With this page loaded on my browser, which happens to be Netscape, I can pull down the "View" menu and select "Document Info." This tells me that the Mindspring Home Page comprises eight files — one text and seven images. When I hit the "Home" button on my browser, the Mindspring Web server serves these eight files individually and logs eight file requests or "hits," and my browser assembles these eight files on my computer screen to form the Mindspring Home Page. If this were my first visit to this page, the Mindspring server would log eight hits. But this is not my first visit. Accordingly my browser retrieves the page not from Mindspring but from my computer, which knows it as local cache file MOPVJICM. In this case, the server logs zero hits. You see the difficulty.

Web measurement standards

Every company and every constituency in the Web measurement business has its own idea about what should be standard measurement procedure. In other words, there is no standard. Two US industry organizations have attacked this problem: CASIE and the IAB. CASIE is the Coalition for Advertising Supported Information & Entertainment, a joint venture of the American Association of Advertising Agencies and the Association of National Advertisers. The IAB is the Internet Advertising Bureau.

CASIE represents Web advertising buyers. The IAB represents Web advertising sellers. CASIE calls its standards "Principles," and the IAB calls its standards "Metrics."

Now I'm going to compare the CASIE Principles and IAB Metrics. Next I'll compare these to the measurement units used by NPD Media Metrix on the user centric side. Finally, I'll simulate a short Web session and see how each of the three systems would record my behavior.

CASIE Principles

The CASIE Principles define four units of Web measurement: Hit, Page, Visit and User.

Hit
CASIE defines a Hit as "a record in a Web server log." This unit is not advertiser relevant, but it begs definition because it is the basis of site centric measurement. As I just showed, I hope, it is understated by caching. Caching is an even bigger problem than I indicated, because not only individual browsers but also Internet service providers of all kinds engage in caching. This latter manifestation is called proxy caching. It saves time and money, but it also wreaks havoc on Web measurement.

Page
CASIE defines a Page as "a transfer request for a single Web document (or set of associated Web files)." In media research terminology, this corresponds roughly to a measure of gross page circulation. However, as I illustrated with my Mindspring Home Page example, it begs the question of how to distinguish documents from files. It also ignores frames, a technique of serving Web information in sections. Frames allow sections of Web pages to stay on screen through subsequent page requests. To appreciate the difficulty of this from a media research point of view, imagine what kinds of answers you would get from people if you asked them how many pages they read of a magazine whose every page had been cut into a different number of strips.

Visit
CASIE defines a Visit as "a page request by a unique individual." This corresponds roughly to a net audience estimate. However, as I have shown, it also poses a difficulty, since neither the IP address logged by a Web server nor its corresponding domain name (if any) is likely to represent an individual Web user. Web analysis systems attack this problem by one of four means — inference, referrals, tokens or cookies. Inference says consecutive requests from the same source are from the same person. Referrals say consecutive requests for internal information are from the same person. Tokens and cookies are two ways of tagging request sources for subsequent identification. Of the four methods, none is very reliable, but inference is worst, and cookies are best. By whatever means, a visit count as defined by CASIE ignores logged audience characteristics like .com, .edu and so forth. More gravely it also lacks a time dimension. What that means will become clear when I get to talking about the IAB Metrics.

User
CASIE defines a user as "a Page request by a unique individual who has disclosed personal information." This corresponds to an estimate of net audience demographics. Users are self selected, as in magazine subscriber studies. Web users disclose personal information through registration.

IAB Metrics
The IAB defines six units of Web measurement: Page request, Audience Characteristics, Audience Behavior, Ad Page Request, Ad Click, and Ad Click Rate.

Page Request (Non-Framed)
A Non-Framed Page Request according to the IAB is "the opportunity for an HTML document to appear in a browser window as a direct result of a visitor's interaction with a Web site." This differs from the CASIE definition in answering the question of how to distinguish documents from files by defining documents as HTML files and in accounting for frames in a manner I am about to describe.

An HTML file is a text file. An early draft of the CASIE Principles defined a Page as "an HTML file request," which in all but semantics is exactly how the IAB defines it. This wording was abandoned by CASIE in response to criticism from Web publishers, who told us they serve many pages with no associated HTML files. Evidently the IAB succeeded where CASIE failed in persuading publishers to accept the reality of what is possible.

Page Request (Framed)
Here is how the IAB accounts for frames: "A single page request will be recorded when an HTML document is requested that will replace the entire window or a portion of the window that was present at the time of the request." By this means the section that stays on screen is not recounted.

Audience Characteristics
The IAB accounts for logged audience characteristics "as defined by browser type, platform, domain, referral link, and regionality." Such variables are as close as server log analysis can get to net audience "demographics," although clearly they have little to do with traditional demographics like age, sex, or income. You need registration for that.

Audience Behavior
The IAB puts several metrics under the heading of Audience Behavior, including "Visitor, Visit, Return Visits, Time and Average Time."

A Visitor is someone "identified by unique registration, unique cookies, unique URL tagging (tokens), or unique IP address with heuristic." Thus what the IAB calls a Visitor is a combination of what CASIE calls a Visit and a User. I'm not sure what is meant by heuristic, but it sounds like a euphemism for inference, the term used by CASIE.

A Visit is "a series of page requests by a visitor without 30 consecutive minutes of inactivity."This adds a necessary time dimension to the CASIE definition.

A Return Visit is "the average number of times a visitor returns to a site over a period of time." This corresponds to an estimate of average frequency. It has no CASIE analog.

Time is "the elapsed time from the first to the last page request that constitutes a visit, and adding the average time per page for such a visit." This definition infers that time elapsed is time exposed, even though the elapsed time could be time spent in the bathroom or playing computer games instead of being exposed to Web pages. Adding the average time per page is an attempt to account for time spent with the last page requested, since a Web server has no way to know when an exposure ends unless another page is requested. This has no CASIE analog either.

Average Time per Page Request is "the elapsed time from the first to the last page request that constitutes a visit divided by the number of page requests in that time, minus one." This too infers that time elapsed is time exposed. This too by necessity ignores time elapsed at the last page requested. This too has no CASIE analog.

Ad Click
The IAB defines an Ad Click as "the opportunity for a visitor to be transferred to a location by clicking on an advertisement, as recorded by the server." This is a measurement of advertising response. Note that it does not guarantee advertising exposure, anymore than a Page Request estimate guarantees page exposure. It has no CASIE analog.

Ad Click Rate
The IAB defines Ad Click Rate as "Ad Clicks divided by Ad Requests." This too is a measure of advertising response. It too has no CASIE analog.

User Centric Measurement
I now turn to user centric measurement.

NPD Media Metrix' PC Meter
Despite its name change NPD Media Metrix still uses a PC meter. This device is a software program installed on Windows PC's in a stratified quota sample of some 10,000 US homes. In this respect it is prone to every bias that every sample based audience measurement system is prone to. In this case, specifically, the sample is not representative of Web users, because the Web is used by people who do not use Windows, people outside the home and people outside the US. The meter runs at the operating system level, which is to say it works passively except that it asks users to identify themselves at the outset of each usage session. It records Web usage by user by page by time. This is a considerable improvement on site centric measurement, which records site not Web usage, identifies visitors not users by inference, referrals, tokens or cookies, infers page usage from HTML file requests, and infers time exposed from time elapsed.

Media Metrix
Media Metrics reports eight units of Web measurement: Page, Reach, Usage Days per Person, Page Requests Per Usage Day, Viewings per Page Request, Seconds Per Viewing, Minutes Per Page Request, Minutes Per Usage Day, and Minutes of Usage per Person.

Page
A Page is "a unique URL." The URL corresponding to the current request is not logged by any Web server. If it were, measuring page requests would not be such a challenge for log analysis. In this respect the Media Metrix Page is an improvement on both the CASIE and IAB versions. In another respect, too, the Media Metrix Page is an improvement. The PC meter is immune to caching. Whether the respondent's browser retrieves a page from the respondent's computer, a proxy server or the server where the page originates is a matter of indifference to the PC meter. If the page is rendered by the browser, it is recorded by the meter.

Reach
Reach is the "percent of individuals that visited a specific site or site category divided by the total number of individuals who accessed the Web that month." This corresponds to a Web audience share estimate. It is unique to user centric measurement. No site centric system can make such an estimate, for no site centric system measures a universe of Web users, representative or not.

Usage Days Per Person
Usage Days Per Person is defined by Media Metrix as "the number of different days the individual used the Web in the current month." This unit, too, is unique to user centric measurement, for the same reason.

Page Requests Per Usage Day
Page Requests Per Usage Day is defined as "the number of unique pages the individual views in one day." This unit, too, is unique to user centric measurement. For the same reason.

Viewings Per Page Request
Viewings Per Page Request is "the number of viewings per unique page the individual views in one month." It, too, is unique. It too includes cached pages.

Seconds Per Viewing
Seconds Per Viewing is "the average amount of time (in seconds) the individual spends on a particular view." This is comparable to the IAB's definition of "Time," but includes cached pages.

Minutes Per Page Request
Minutes Per Page Request is "the average amount of time (in minutes) the individual spends on a unique page." This is comparable to the IAB Metric of "Average Time," but unlike site centric measurement, it includes cached pages and doesn't have to play games to estimate the time spent viewing the last page retrieved from a given site.

Minutes Per Usage Day
Minutes Per Usage Day is "the average amount of time (in minutes) the individual spends on the Web in one day." This is analogous to the IAB Metric of "Time" but unlike site centric measurement includes multiple sites.

Minutes of Usage Per Person
Minutes of Usage Per Person is "the average amount of time (in minutes) the individual spends on the Web in one month." This is unique to user centric measurement.

A Short Web Session
Now let's follow me through a short Web session and see how it would be recorded by following the CASIE Principles, IAB Metrics and NPD Media Metrix.

Before I begin I clear my cache.

At 12:05 I begin by going to the Mindspring Home page. Here it is; note the ad for Windows 95 at the top.

At 12:06 I follow an internal link to the Mindspring News Page.

At 12:07 I follow an external link to Wired News. I see an ad for something in the upper hand corner, but I can't see what it is.

At 12:08 I return to the Mindspring Home Page by hitting the "Home" button on my toolbar (not shown here). There's the Windows 95 ad again.

At 12:08 I quit.

In summary, I visit four pages, two of which are the same, namely the Mindspring Home Page, in the course of four minutes. These pages comprise 8, 4, 16 and 8 files respectively, for a total of 36.

Now suppose we analyze the Mindspring server log.

Following the CASIE Principles, we would count 28 Hits, missing 8 that my browser retrieved from my local cache when I returned to the Home Page; an unknown number of Pages, since CASIE is unclear on how to distinguish page from file requests; 2 Visits, since CASIE puts no time limit on its Visit definition, and 1 User, since I subscribe to Mindspring and therefore am identifiable to them.

Following the IAB Metrics, we would count 2 Page Requests, missing 1 due to caching; we would characterize me as coming from a commercial TLD and the Mindspring SLD, which from an SIC database we would identify as an ISP; we would count 1 Visitor; 1 Visit due to caching; 0 Return Visits due to caching; 2' Time equals 1' at the Home plus 1' at the News Page; 1' Average Time; 1 Ad Page Request — the Windows 95 ad on the Home Page; 0 Ad Clicks; and a 0%Ad Click Rate.

Now suppose I am a member of the NPD Media Metrix panel.

The PC meter on my computer would count 4 Pages, 1 Usage Day, 4 Page Requests Per Usage Day, 1.33 Viewings Per Page Request since I viewed 3 pages once and one twice, 60 Seconds Per Viewing, 1 Minute Per Page Request, 4 Minutes Per Usage Day and 4 Minutes of Usage. On the other hand, unless Mindspring is visited by other panelists besides me, its audience may be too small for NPD to report in such detail due to sample size limitations.

Summary

In summary, the CASIE Principles are good, the IAB Metrics are better because they specify what is possible with site centric measurement, the NPD Media Metrix are best but subject to sample bias, and what's possible now is unacceptable long term, because none of it describes as accurately as possible what's really happening on the Web. Therefore, some sort of hybrid approach holds promise for the future. I wish I could tell you about it.

Thank you.