

Paper 191-30

Drilling Down on Web Usage: Custom Applications of SAS® Web Analytics

Carol Kassel, Columbia University, New York, NY

ABSTRACT

Columbia University was one of the first large institutions to begin using WebHound™ (now SAS Web Analytics), a web usage statistics package that runs on top of base SAS. We licensed it in order to answer some questions we had about usage of some of our websites. Since that time, as our understanding of the software increased, we discovered that we could customize it to answer more involved questions.

Some of our custom applications of this solution include: understanding where users went from the Columbia home page and changes in this pattern after a redesign; combining this information with an analysis of searches from the home page to help users find areas of interest quickly; providing reports to subscribers of Columbia databases to inform them of their constituents' usage, and making these reports conform to specific standards; and providing site-specific custom reports to requesting groups on campus on an as-needed basis.

Overall, our impression has been positive, with a few caveats about setup and hardware requirements. Our goal has been, and continues to be, to maximize the benefit of this SAS solution by using it as widely as possible.

No previous knowledge of SAS Web Analytics or WebHound is required to read this paper; previous knowledge of SAS is recommended but not required.

INTRODUCTION

Columbia University already had a license for base SAS and several experienced users when we decided to license WebHound, a new solution, in February 2002. We selected WebHound from among other web usage statistics packages in part because of our existing experience with SAS, but mainly because of WebHound's ability to "drill down" on a variety of metrics. That is, it can generate reports in which you travel from general information to more specific information as you click (e.g., start with a list of referring domains, click on a specific domain to find out the top entry points from that domain). Most web usage statistics packages allow for single-variable metrics, but WebHound allows for multi-variable metrics out of the box. We wanted to make sure our audiences were finding what they needed, and we wanted to examine web usage to see if these goals were being met.

Licensing, setup, and usage of WebHound was a joint project between the Academic Information Systems (AcIS) department, which provides campus-wide hardware, software, networking, and support, and Digital Knowledge Ventures (DKV), which helps faculty and administrators use digital media to communicate with audiences across and beyond campus. We found that setting up WebHound was an involved process and learning how to use it required a significant commitment. However, once it was running and we learned how to use it, we began extending its use for a variety of purposes with relative ease. We also worked closely with representatives from SAS, especially during our first year of use.

In 2004, SAS upgraded WebHound, selecting a new name, SAS Web Analytics, to reflect a solution that was transformed from its previous iteration. SAS Web Analytics uses SAS 9 and incorporates many of the suggestions made by WebHound customers (including us!).

In this paper, I will discuss some of the ways we used WebHound, to give an idea of the ways in which this solution can be customized to obtain specified information. These examples are geared toward a general audience: those who already use this solution, those who are evaluating it for their organization, and those who are just curious.

Note: The goal of this paper is to provide a "real-world" example of usage of this SAS solution. I will be discussing the questions we asked and the way we found the answers. There are capabilities of SAS Web Analytics, such as customer segmentation, executive dashboard functionality, and the like, that do not apply as readily to a university setting. If you work for a commercial entity, you may want to investigate these features further on your own.

NOTE ON TERMINOLOGY: WEBHOUND VS. WEB ANALYTICS

In this paper, I will be referring mostly to WebHound because we have been using WebHound at Columbia. We do plan to migrate to SAS Web Analytics as soon as possible, but we have not done so by the writing of this paper. However, while SAS Web Analytics has capabilities that WebHound does not, many of the same concepts apply. I

have seen a preview of SAS Web Analytics so I have a sense of what its capabilities are. If there is a feature specific to SAS Web Analytics, I will mention it; otherwise, I will say “WebHound” when describing this software.

A MYRIAD OF WEBSITES, A VARIETY OF QUESTIONS

At Columbia, we have a variety of websites, from Columbia’s home page to departmental sites to aggregations of scholarly publishing, developed and maintained by different groups on campus. At first, our questions about usage were fairly basic, such as the number of visitors to each site, the number of page views, etc. We were able to obtain such information using the “out-of-the-box” processing and reporting features. Here is an example of a standard report, with the navigational frames that make moving from report to report simple (Figure 1):

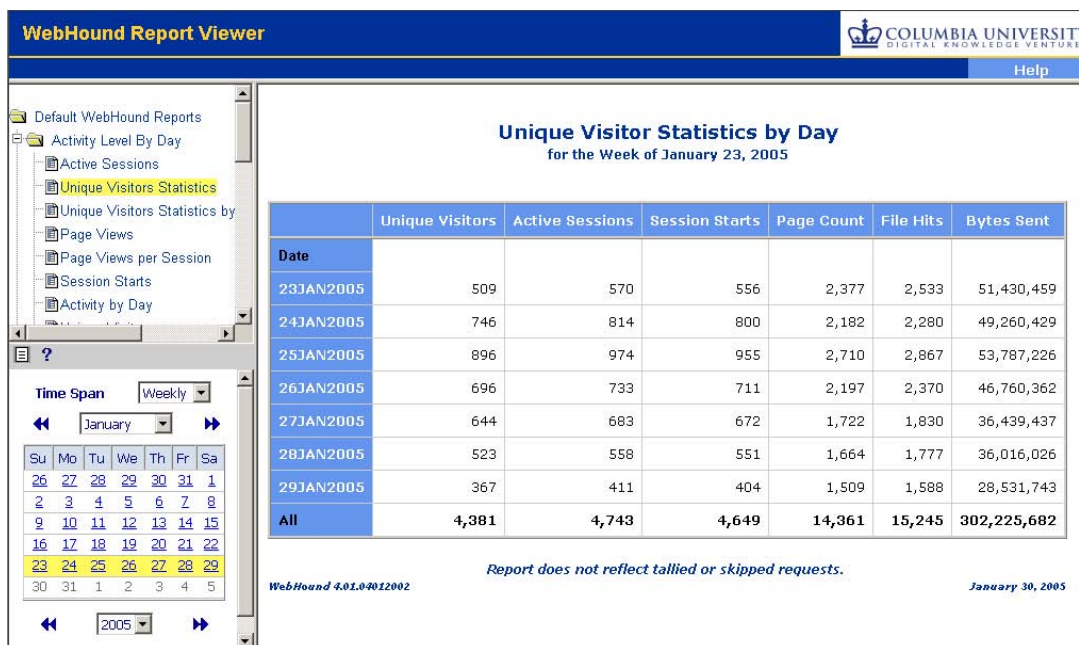


FIGURE 1. STANDARD REPORT

Many of the standard reports were quite impressive, especially the drillable reports, and also reports based on pathing data. That is, WebHound organizes all of the requests into paths and determines the most popular clicks from one page to the next. One report provides a visualization of common paths through the site (Figure 2):

Clickstream Visualization: Entry Points to Exit Points

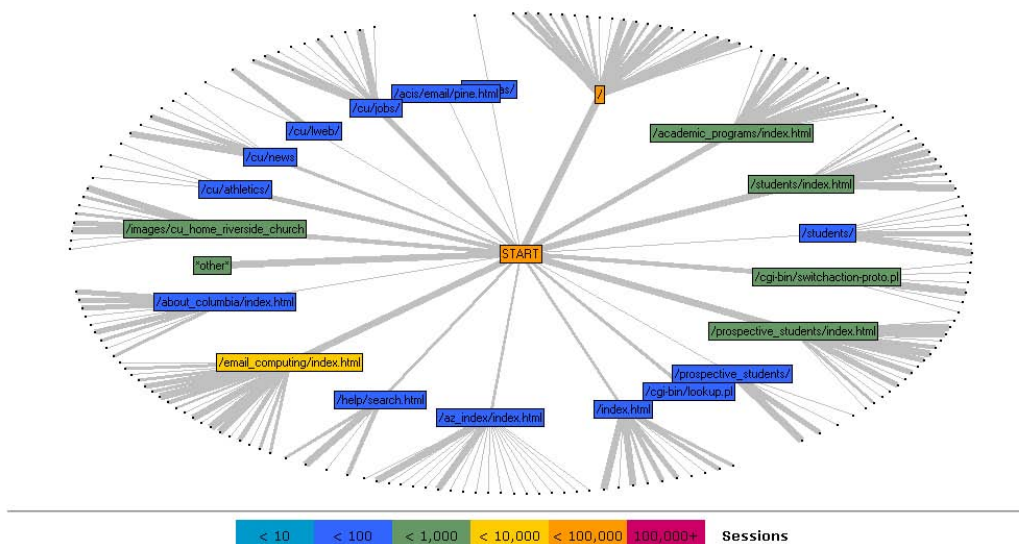


FIGURE 2. VISUALIZATION REPORT

We were able to learn a lot from the standard reports. But once we got a handle on the software and what it could do, we also began to address more complex questions, such as:

- Where were users going from the Columbia home page? A redesign was planned, but how could we figure out what to keep and what to change, and how would we know if the redesign was successful?
- What were users searching for, and from what page within our system were they searching? How did this vary with on-campus vs. off-campus users? Which were users doing most from the Columbia home page: website searches or “people searches” (searching the Columbia directory)? Finally, did the search terms tell us anything about links we should be featuring more prominently?
- For our online specialty websites, available by subscription, how could we provide statistics to each subscribing institution on their constituents’ usage?
- Which videos on our servers were being viewed the most? How much of each video were users actually watching? Overall, how many minutes or hours of video were being watched on a monthly basis?
- Could we use the software to provide custom reports for other campus groups as well as ourselves?

First, we needed to figure out what information would answer these questions. Then we needed to write the appropriate SAS code to obtain the answers. The basic rule of thumb, we discovered, was that if we could quantify a question, we could customize WebHound to find the answer. The reason? WebHound, and SAS Web Analytics as well, uses the already-powerful SAS software. It provides a lot of shortcuts to produce useful reports without extra effort, but it also allows those with knowledge of SAS to make adjustments using methods they already know.

UNDER THE HOOD

How does WebHound work, and what is its relationship to base SAS? To put it in simplest terms, WebHound consists of a series of highly sophisticated SAS macros and procedures, compiled for your convenience so that you do not have to do a mountain of work manually. It takes the contents of a web log file, pulls the data into a series of SAS tables, and generates reports based on your specification.

Web log files vary in format depending on your web server as well as any customizations you may have made for your organization. WebHound reads a variety of file formats by default; you can also add custom code to handle additional information that might be in your logs.

Log files are processed and the information is brought into SAS tables during the ETL phase. “ETL,” for those unfamiliar with the acronym, stands for “extract, transform, and load,” and it is the process in which raw data is integrated, managed, and converted into SAS data sets. In WebHound’s ETL process, there are two types of SAS tables created: detail tables, in which each log record is a separate entry, all fields are populated, and each table represents all of the data from a given ETL; and summary tables, in which log records are combined, small subsets of the fields are populated for each table, and tables are divided by time period. Reports generally use the summary tables, but the detail tables are useful for you to check the ETL and make sure all your fields were populated correctly. Apparently, SAS Web Analytics’ ETL is similar.

The report process generates a combination of static and dynamic reports. Each report has a preparation source file and a presentation source file, which are essentially SAS commands to obtain the appropriate data from the appropriate summary table, massage the data, and then produce the report using either PROC tabulate or PROC report. For the dynamic tables (drillable reports), WebHound creates MDDBs for each report. MDDB, or “multidimensional database,” is a storage facility that pulls in data from a data warehouse, summarizes it, and holds it in a matrix-like format. When you click on a drillable report, it accesses the data in the MDDB and returns the appropriate view. I gather that SAS Web Analytics reports are almost all dynamic, generated “on the fly” upon user request. Rather than MDDBs, they use OLAP, or “online analytical processing,” which is another storage mechanism that also pulls data from a data warehouse and holds it in a matrix, or “cube,” but it is apparently more efficient than MDDBs. In this way, the user can customize the output of every report, not just the drillable reports—re-sorting at will, exporting to another program such as Excel, and so forth.

Both WebHound and SAS Web Analytics have the notion of “webmarts,” which are collections of data, rules, and reports. You can run reports on multiple websites using one webmart as long as your rules are the same for all. You create additional webmarts when you encounter a site or group of sites with a distinct set of rules. At Columbia, we currently have over 20 webmarts, some of which are used to generate reports for multiple sites.

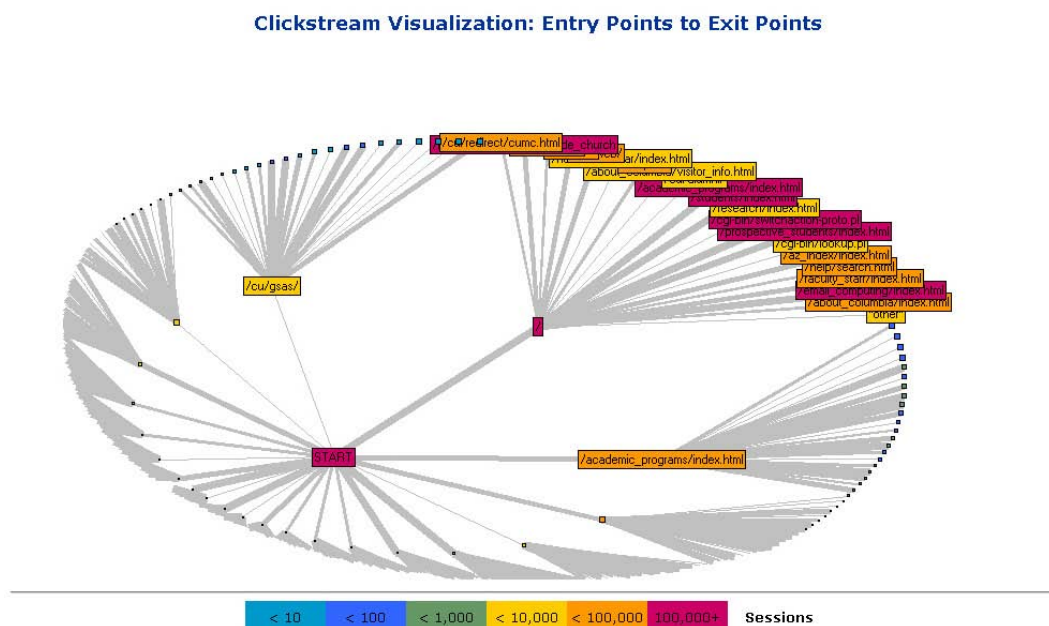
COLUMBIA HOME PAGE

In the second half of 2003, DKV was asked to provide new designs for the Columbia home page. We formed part of a university-wide committee to determine the requirements for the new design. As part of our gathering of information, we wanted to look at current usage. One of our main objectives was to figure out where users were going from the Columbia home page. This information would help us to decide what to keep and what to discard when redesigning it. We also wanted to track overall usage after the redesign.

We collected the relevant log records and loaded them into WebHound. To determine where users were going from the Columbia home page, we used a standard report that provides a visualization of common paths through the site. It uses SAS's ability to link together a series of events and analyze them statistically. In this case, the events were views of web pages. Isolating the home page, we could see the pages that users visited next and find out which pages were the most popular. For instance, we determined that “Computing and Email,” “Schools and Departments,” and “Admissions” were important links that should be kept somehow in the new design. These links were renamed but their essential function remained, whereas other, less-used links were folded into other categories or moved to the next level according to topic.

During our analysis, we discovered that there were some links that we could not track—including searches from the home page and links to sites on other servers (such as Athletics or Medical Center)—because the information was unavailable. As we implemented the redesign, we also added pages, essentially redirects, that would allow clicks to those sites to be logged so that we could count them.

Post-redesign, the paths looked like this (Figure 3):



The slightly renamed “Email and Computing” was still the top destination, followed by “Academic Programs” (formerly “Schools and Departments”), “Search,” “Students,” and “Libraries.” These destinations were consistently the top five. We were also pleased to note that all of the links on the new home page were followed at least some of the time!

DOUBLE-CHECKING USAGE: SEARCHES FROM THE HOME PAGE

When we began analyzing usage of the home page, we realized that we needed to look at searches from the home page. After all, if users were searching for certain terms often enough, that might indicate something that users found important that they were not finding on the home page. It might warrant adding those terms as links.

In addition, we have two types of search: “websites” and “people.” The website search returns web pages on our servers that match the given term. The people search looks at the Columbia directory and returns matches from there. We guessed that we should default to websites – but did the data support the hypothesis?

First, we put together a “top search terms” report for all searches regardless of their source. We created a new variable called “search_term” and populated it based on SAS’s “scan” command. Then it was a simple matter to set up a report that grouped all of the search terms and sorted them by popularity. This report now comes standard as part of SAS Web Analytics.

But finding out the top search terms site-wide did not tell us specifically about searches from the home page. In addition, we had added search boxes to the tops of other high-level pages (such as Email and Computing or Students) and we wanted to track searches from those pages also. Therefore, we needed to drill down on search requests to determine where they were coming from, and then look at the search terms therein.

In the resulting report, you click on a desired time period and then a referring page. Here is the next page if you click on "http://www.columbia.edu/" (i.e., the home page; Figure 4):

Referring Pages > Top 100 Search Terms

Drill Path: <http://www.columbia.edu/>

Search Term	↑	Page Views	
		Sum ↑ ↓	Percent of Sum ↑ ↓
~ALL OTHERS NOT IN THE TOP 100		116,916	74
cubmail		4,442	3
bookstore		1,981	1
courseworks		1,802	1
cce		1,366	1
map		1,191	1
acis		1,101	1
jobs		1,077	1
registrar		970	1
email		944	1
campus map		851	1
housing		772	0
study abroad		735	0
academic calendar		691	0

FIGURE 4. DRILLABLE SEARCH TERM REPORT

Now we can see specifically what users are searching for from the home page. For instance, many users are looking for the term "cubmail," which is the name of Columbia's web-based email program. Perhaps that suggests that a "cubmail" link should be added to the home page. This idea is currently under discussion.

SEARCHES: WEBSITES VS PEOPLE

In addition to knowing what users were searching for, we also wanted to know which type of search was most popular: websites or people. We offered a quick search box on the top of the home page and top-level pages, with radio buttons to select websites or people. We hypothesized that we should offer websites as our default. But did the data support this assertion?

It was a simple matter to create a new variable, Search_Type, that would have the value "Websites" or "People" depending on the URL of the requested file. Then we made some modifications to the standard "unique visitors statistics" report source files so that they would provide numbers by search type. We also generated another report that gave information only on searches from the home page by type (Figure 5):

Searches from the Home Page

Date	People				Websites			
	Unique Visitors	Active Sessions	Session Starts	Page Views	Unique Visitors	Active Sessions	Session Starts	Page Views
	813	865	734	990	2,056	2,209	2,133	2,625
	1,255	1,362	1,153	1,622	2,553	2,738	2,610	3,254
	2,206	2,671	2,244	3,273	3,864	4,223	3,967	5,023
	2,067	2,410	2,064	2,927	3,520	3,862	3,643	4,641
	1,359	1,447	1,215	1,644	3,021	3,238	3,081	3,796
	927	989	849	1,176	1,990	2,117	2,025	2,514
	691	720	618	850	1,813	1,908	1,833	2,237
Total	9,318	10,464	8,877	12,482	18,817	20,295	19,292	24,090

FIGURE 5. PEOPLE VS WEBSITE SEARCHES

We also wanted to check usage by different audiences. One of the ways we look at our users is to divide them into internal and external users (i.e., on-campus users vs. off-campus users). We wanted to see if these two audiences use the search function differently. To divide the log records accordingly, we created a new variable (“cuhost”) and populated it with a 0 or 1 according to the IP address or domain of the host. Then we applied a format (“CU Campus Host” or “Off-Campus Host”) and added the variable to the output of the relevant reports (Figure 6).

Searches from the Homepage by Search Type and Host Location

	CU Campus Host				Off-Campus Host			
	People		Websites		People		Websites	
	Unique Visitors	Page Views	Unique Visitors	Page Views	Unique Visitors	Page Views	Unique Visitors	Page Views
Date								
	107	144	139	183	706	846	1,917	2,442
	416	615	307	421	839	1,007	2,246	2,833
	919	1,707	727	1,074	1,287	1,566	3,137	3,949
	828	1,384	641	952	1,239	1,543	2,879	3,689
	321	433	318	432	1,038	1,211	2,703	3,364
	178	238	183	245	749	938	1,807	2,269
	135	189	159	216	556	661	1,654	2,021
Total	2,904	4,710	2,474	3,523	6,414	7,772	16,343	20,567

FIGURE 6. PEOPLE VS WEBSITE SEARCHES FROM THE HOME PAGE

We found that overall, more people were searching for websites than people. But on campus, there were more searches for people than websites. We considered changing the default for on-campus hosts, but decided this would confuse people. Based on our discussions and analysis of the data, we decided to keep the search default the way it was.

USAGE STATISTICS FOR SUBSCRIBING INSTITUTIONS

The work we did with Columbia’s home page usage statistics and with the search statistics shows the ways in which WebHound allows for the creation of new variables and the organization of data by any metric you choose. But we also faced a different challenge: dividing the data into hundreds of groups and running reports by group.

The Electronic Publishing Initiative at Columbia (EPIC) has three websites available by subscription to universities, libraries, and other institutions worldwide. These websites—Columbia International Affairs Online (CIAO), Columbia Earthscape, and Gutenberg<e>—are aggregations of scholarly content on a particular topic. When an institution subscribes, its members gain access to the content of the site either by IP address or domain recognition or by username and password. The IP address, domain, or username appear in the log file and can be used later to identify a given record as that of a given institution.

When a library purchases a book, it can determine the book’s usefulness to its members by tracking the number of times the book is checked out. Similarly, it can determine the usefulness of a web resource to which it subscribes by tracking usage by its members. The difference is that the library does not have the data on web usage; it must rely on the website provider for reports. Furthermore, in order to compare usage across different websites, the library must trust that all of the website providers are adhering to the same standards. To address this need, a consortium of library professionals developed a set of standards, called Project Counter (www.projectcounter.org), to regulate usage statistics across online publications.

Librarians and administrators at the hundreds of institutions that subscribe to EPIC’s websites had been asking for statistics for a long time. EPIC had a home-grown system in place for CIAO and Earthscape, but it was time- and resource-intensive—and they offered nothing for Gutenberg<e>. Furthermore, this system did not conform to Project Counter’s guidelines. Finally, there were some consortia (groups of institutions) who wanted to see not just reports for member institutions, but also aggregate reports for the entire consortium. EPIC approached DKV to see if we could use WebHound to develop the reports they needed.

Implementing this project involved several steps:

1. Gathering the relevant web log records for a given website, such as CIAO.
2. Exporting the IP address, domain, and username data for each institution for use in identifying each log record.

3. Comparing each log record to the institution information to determine its institution and recording that information.
4. Importing the log records into WebHound and writing the institution information to a field for later retrieval.
5. For a given log record, determining whether its institution is part of a consortium and recording that information for later retrieval.
6. Recording institution name and consortium name (if applicable) for each record.
7. Running the WebHound report process for each institution.
8. Running the WebHound report process for each consortium.

WebHound (and SAS Web Analytics as well) has anticipated the need to subdivide log records into an arbitrary number of groups. It provides an empty variable, called `Content_Group`, that appears in every summary table out of the box. SAS provided us with code to run the report process by content group and write the resulting reports for each group to a separate directory. We populated the `Content_Group` field with the ID number of each institution.

We decided to use a Perl script to accomplish steps 1, 2, and 3. The script wrote the resulting institution ID number directly to each record in the file we would import into WebHound. We customized the ETL process to read the ID number on input and populate the `Content_Group` field with it (step 4). We also created a SAS table that mapped institution IDs to consortium IDs and checked each record against that table, populating a consortium field (a new field called `Consortium_ID`) appropriately, during the ETL (step 5). We also had SAS tables that mapped institution ID to institution name and consortium ID to consortium name, and we each name in a separate field (step 6).

We then ran the report process by content group (step 7). Here is an example of a report (Figure 7):

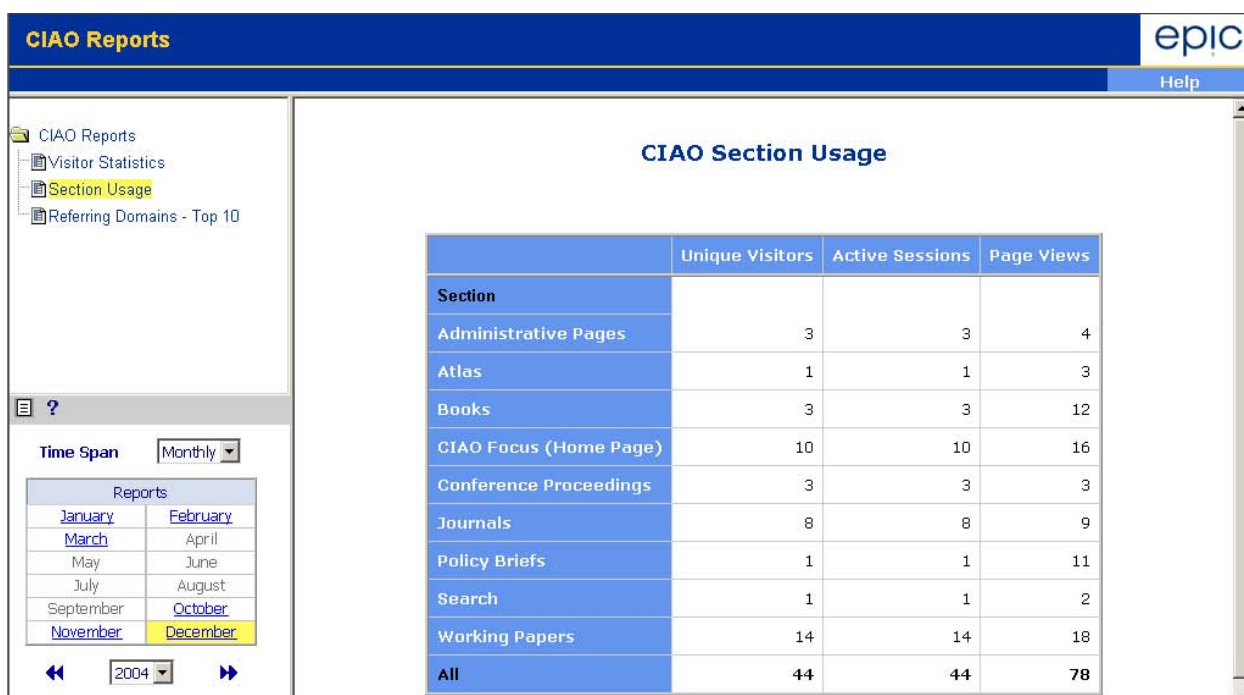


FIGURE 7. SUBSCRIBER REPORT

Finally, we ran the report process by consortium ID by modifying the code to run reports by content group and substituting `Consortium_ID` for `Content_Group` (step 8).

To conform to the standards of Project Counter, we did change some of WebHound's default settings. For instance, we discovered that accessing a PDF file could lead to several log records, just for one view! Project Counter – not to mention common sense – explicitly states that one view must be counted only once. We figured out how to determine which records were legitimate requests and which were partial requests, and we changed the SAS code that determines page views to reflect this discovery. Overall, though, we found that WebHound already conformed to the most important points in Project Counter's mandate, such as ways to determine a session.

VIDEO STATISTICS

We were focusing a lot on usage by website, analyzing requests to web pages. But there is another type of usage on the web: viewing streaming video files. In our system, streaming video in RealMedia format is served on a particular

server, configured for that purpose. Logs for that server are stored in a particular format in a distinct file. We wanted to obtain some information specific to videos, such as the number of hours per month that were being viewed, etc.

WebHound reads standard log formats out of the box, but it can also be easily configured to read in additional log information. We used this functionality in the subscriber report project when we wrote the institution ID to each record and read it into WebHound. But now, we wanted to report on the activity of our RealMedia server, whose log files were nonstandard. In particular, they contained multiple extra fields, such as file time, sent time, resends, etc. Meanwhile, they did not have referrer information in them. How could we accommodate the additions and the subtractions, and then use the additional data in our reports?

We actually made the main part of the log records conform to the standard log format of our other logs. We found that it was easier to do this on the pre-processing end (i.e., before importing the logs into WebHound). Meanwhile, bringing in multiple extra fields was as easy as bringing in one extra field and used the same logic. We used this additional information in some of our reports. Here is an example (Figure 8):

The Most Popular Pages

Top Pages and Percent of Total					
Rank	Requested File	Page Views	Percent of Total	File Time	Avg. Time Viewed
1	ciao/cbr/clip1.rm	6,589	13%	0:14:43	0:02:35
2	ciao/cbr/clip6.rm	2,325	4.7%	0:23:17	0:03:57
3	video/news/04/321_plight_of_theRoma/hillaryRodhamClinton.rm	2,056	4.2%	0:20:44	0:05:01
4	ciao/cbr/clip2.rm	1,649	3.4%	0:14:16	0:04:00
5	ciao/cbr/clip5.rm	1,380	2.8%	0:24:17	0:04:28
6	ciao/cbr/clip3.rm	1,209	2.5%	0:14:00	0:04:09
7	ciao/cbr/clip4.rm	1,108	2.3%	0:09:09	0:05:10
8	ciao/cbr/clip1low.rm	1,064	2.2%	0:14:47	0:02:50
9	law/CGLPSS/Sands041028.rm	655	1.3%	0:56:12	0:07:19
10	video/news/04/296_challenges_sustainable_development/jeffSachs.rm	509	1.0%	1:09:04	0:04:34
11	itcmedia/law/library/070d.rm	457	.93%	0:32:04	0:01:05
12	itcmedia/law/library/072_morris.rm	401	.82%	0:42:11	0:03:02

FIGURE 8. VIDEO USAGE REPORT

EFFICIENT CREATION OF NEW REPORTS

DKV's goal was to produce usage statistics for ourselves and other departments quickly and efficiently. We were pleased to find that, once we had a handle on WebHound, we could create new webmarts or generate new sets of reports with relative ease.

One site for which DKV had specific questions was a site we co-developed with other groups: Fathom, a distance learning site. Fathom had officially closed its doors, but it continued to offer an archive of free features. We wanted to know how many people were continuing to visit the site. We also wanted to see if there were trends in the popularity of the features to determine future development and marketing efforts. For instance, we might direct traffic to one of our other sites for more information on a particular topic, or we might develop courses on a given subject if we found that there was sufficient interest to justify it.

Each feature page on Fathom has a feature code in its URL. We created SAS tables with feature codes and titles, and set up formats based on the code in the requested file. We then generated a report similar to the standard "Top 25 Pages" report for features (Figure 9).

The 100 Most Popular Features

Top 100 Features and Percent of Total			
Rank	Feature Number	Page Views	Percent of Total
1	Tsunami: Where they Happen and Why	389	5.3%
2	Bollywood: Youth Culture and the Global Perspective	147	2.0%
3	Lost Egypt: Photography and the Early Documentation of Egyptian Monuments	81	1.1%
4	The Genius of Joseph Wolf, Father of Modern Wildlife Art	65	.89%
5	Human-Animal Relations in the Era of Postdomesticity	63	.86%
6	The Rolls-Royce Merlin	51	.70%
7	The Evolution of Modern Humans	50	.69%
8	The Boxer Rebellion, 1900: A Selection of Books, Prints and Photographs	48	.66%
9	Art Nouveau and the Erotic	46	.63%
10	Natural Leaders in St. Louis Gangs	46	.63%
11	Snakes: Versatile Predators of the Natural World	46	.63%
12	The Genetic Evidence for Human Evolution	46	.63%

FIGURE 9. FATHOM REPORT

We are still analyzing the traffic patterns but we have some information to help us determine future efforts.

Being able to create new webmarts quickly has proven invaluable, as has an understanding of the type of information that can be made available. Many of the groups who approach us simply want to know “something about usage” but they do not always know what is possible. We are able to work with them and make suggestions about what they might track. Knowing how to customize the results, we are able to react to suggestions efficiently.

CONCLUSION

From analyzing usage of the Columbia home page to reporting web usage to subscribing institutions and beyond, we have been using WebHound widely. We look forward to the migration to SAS Web Analytics so that we can take advantage of its additional features. Our general comments to those considering this solution are to customize it to answer the specific questions you may have about usage of your website. For us, it was the flexibility of the solution to provide answers for a myriad of questions that made the difference in our decision to continue licensing it.

ACKNOWLEDGMENTS

When I began using WebHound, I did not know a single line of SAS. I learned a tremendous amount from SAS expert Lynn Rohrs, Lead Consultant, Academic Information Systems. We worked on many of the customizations of WebHound together. I also owe a debt of gratitude to Rimma Ashkinadze, Systems Programmer, Academic Information Systems, who set up the hardware and software for WebHound and handled all upgrades and maintenance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Carol Kassel
 Columbia University
 514 West 113th Street
 New York, NY 10025
 Work Phone: 212-854-7838
 Fax: 212-854-7555
 Email: carol@dkv.columbia.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.