

# *Performance Indicators For Web Sites*

Brian Kelly

UKOLN

University of Bath

Bath, BA2 7AY

**Email:** [B.Kelly@ukoln.ac.uk](mailto:B.Kelly@ukoln.ac.uk)

**URL:** <http://www.ukoln.ac.uk/>

## **Aims of Talk:**

- To explain why we need performance indicators for Web services
- To describe difficulties in using Web log statistics
- To describe tools for analysing Web logs
- To mention other possibilities for providing performance indicators

---

UKOLN is funded by Resource: The Council for Museums, Archives and Libraries, the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

# *Contents*

Introduction

Web Statistics

- Server log files
- The problems
- Tools and techniques

Other Indicators

- Links to your web site
- Search engine coverage
- User feedback

Conclusions

# *About This Talk*

This talk:

- Based on article on *Performance Indicators For Your Web Site* published in **Exploit Interactive** (see <http://www.exploit-lib.org/issue5/indicators/>)
- Article written to advise funding bodies and monitoring agencies and providers of Web services
- Focuses on the analysis of usage data for Web sites
- Gives a technical rather than a service provider perspective

*Caveat:*

- I am not a librarian! The audience needs to identify which aspects are relevant to Usage Data for E-Collections

# *Background*

*".. the development of the electronic journal is promising much better usage data than we have ever had with paper journals"*

Roger Brown in "*Exploitation and Usage Analysis*",  
The Serials Management Handbook, ed. Kidd & Rees-Jones

Is this true?

*"Web statistics are (worse than) meaningless"*

<URL: <http://www.cranfield.ac.uk/docs/stats/>> Is this true?

Besides web server statistics, what other criteria can be used to provide performance indicators?

# *Why Have Performance Indicators?*

Performance indicators for Web sites can be used for several purposes:

- Use in management reports showing service growth
- For Service Level Agreements with funding agencies
- As basis of negotiations with advertisers
- If closing alternative (paper-based) services
- To identify gaps in service provision
- To predict and plan for future load patterns
- To monitor performance levels
- To advise on deployment of new technologies
- To inform and motivate contributors

# Web Statistics

```
#Software: Microsoft Internet Information Server 4.0
#Version: 1.0 #Date: 1999-12-25 00:00:21
#Fields: date time c-ip cs-username cs-method cs-uri-stem cs-uri-query sc-status sc-bytes cs(User-Agent) cs(Cookie) cs(Referer)
1999-12-25 00:00:21 194.237.174.119 - GET /issue1/jobs/Default.asp - 200 20407 AltaVista-Intranet/V2.3A+(www.altavista.co.uk+jan.gelin@av.com) - -
1999-12-25 00:03:39 194.237.174.119 - GET /statistics/ExpIntHits1.asp - 200 10519 AltaVista-Intranet/V2.3A+(www.altavista.co.uk+jan.gelin@av.com) - -
1999-12-25 00:26:54 209.67.247.158 - GET /robots.txt - 200 303 FAST-WebCrawler/2.0.9+(crawler@fast.no;+http://www.fast.no/...) - -
1999-12-25 00:32:47 194.237.174.119 - GET /issue2/default.asp - 200 5332 AltaVista-Intranet/V2.3A+(www.altavista.co.uk+jan.gelin@av.com) - -
1999-12-25 01:49:54 206.186.25.7 - GET /resources/images/main/bg.gif - 200 300 Mozilla/2.0+(compatible;+MSIE+3.02;+AK;+Windows+NT)
    ASPSESSIONIDGQQQGAD=IIHCBIFDIECKPAPGICDEOJII;+SITESERVER=ID=22e0a17296b8c2ed1f77460cde75c27f http://www.exploit-lib.org/issue1/webtechs/
1999-12-25 01:49:54 206.186.25.7 - GET /issue1/webtechs/Default.asp - 200 24659 Mozilla/2.0+(compatible;+MSIE+3.02;+AK;+Windows+NT) -
    http://www.statslab.cam.ac.uk/%7Esret1/analog/webtechs.html
1999-12-25 01:49:54 206.186.25.7 - GET /resources/images/main/global_home_h.gif - 200 487 Mozilla/2.0+(compatible;+MSIE+3.02;+AK;+Windows+NT)
    ASPSESSIONIDGQQQGAD=IIHCBIFDIECKPAPGICDEOJII;+SITESERVER=ID=22e0a17296b8c2ed1f77460cde75c27f http://www.exploit-lib.org/issue1/webtechs/
1999-12-25 01:49:54 206.186.25.7 - GET /resources/images/main/global_search.gif - 200 534 Mozilla/2.0+(compatible;+MSIE+3.02;+AK;+Windows+NT)
    ASPSESSIONIDGQQQGAD=IIHCBIFDIECKPAPGICDEOJII;+SITESERVER=ID=22e0a17296b8c2ed1f77460cde75c27f http://www.exploit-lib.org/issue1/webtechs/
1999-12-25 01:49:56 206.186.25.7 - GET /resources/images/main/local_home01.gif - 200 663 Mozilla/2.0+(compatible;+MSIE+3.02;+AK;+Windows+NT)
    ASPSESSIONIDGQQQGAD=IIHCBIFDIECKPAPGICDEOJII;+SITESERVER=ID=22e0a17296b8c2ed1f77460cde75c27f http://www.exploit-lib.org/issue1/webtechs/
```

This log file shows visits to the *Exploit Interactive* web site from 00:00:00 on 25 Dec 1999:

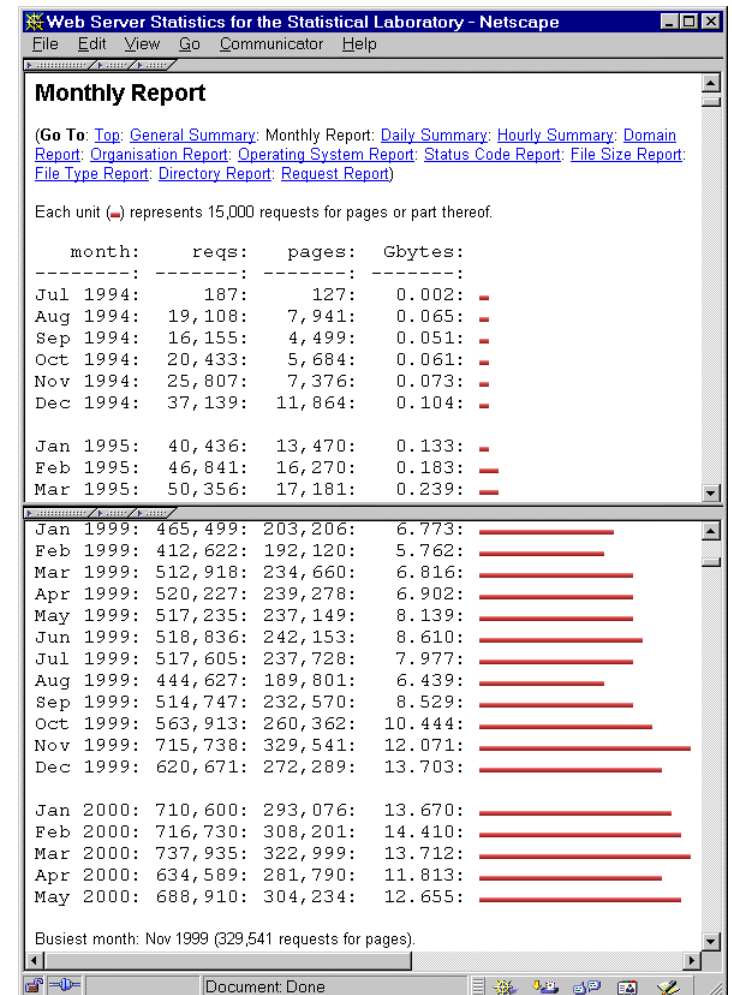
- A visit from an **AltaVista robot** in UK, downloading several text files
- A visit from a **FAST-Crawler** robot in Norway
- A visit from a PC (WinNT) user of an IE browser who followed a link at `<http://www.statslab.cam.ac.uk/%7Esret1/analog/webtechs.html>` and downloaded a **HTML page** and **several images**

# Viewing Web Statistics

The **Analog** program (Cambridge Univ) was one of the first packages to provide a graphical summary of web log file.

What can we say about the web site from Jul 1994 - Mar 1995 (top) to Jan 1999-May 2000 (bottom)?

See <http://www.statslab.cam.ac.uk/~sret1/stats/stats.html>



# *Hits, Requests and Pages*

## **The HTTP Process:**

- A user clicks a link or enters a URL
- The remote web server downloads the HTML page
- The HTML page is interpreted and any inline objects are also downloaded:
  - Each image (occurrence of `<IMG SRC="foo">`)
  - Background image or sound
  - External JavaScript or stylesheet file

## **Summary**

Each individual user request for a **page** can produce multiple **requests** at the remote server and generate multiple **hits**.

# *Fluctuations in Hits & Requests*

## **Scenarios**

**1** In 1993 images are introduced across a web site (two images per text page)

**Result:** Nos. of hits trebles, while number of page requests remains constant

**2** In 1998 external JavaScript files are used to animate menus when they are selected

**Result:** Nos. of hits increases while number of page requests remain constant

**3** In 1999 internal style sheets are used to replace images of *University name*

**Result:** Nos. of hits decrease while number of page requests remain constant

# *Conclusions*

- The term **hit** is not very useful as the number of hits can be affected by developments to the web site architecture.
- Hits, however, are needed in order to monitor server performance levels.
- **Pages** (page requests) are a better indicator than hits
- But who is looking at the pages?

# *Users and Visits*

Registration not normally needed to access Web resources.

Can we track **users** easily? Can we profile users?

```
1999-12-25 01:49:54 206.186.25.7 - GET /issue1/webtechs/Default.asp - 200 24659
Mozilla/2.0+(compatible;+MSIE+3.02;+AK;+Windows+NT) -
http://www.statslab.cam.ac.uk/%7Esret1/analog/webtechs.html
```

The web log tells us:

- User on computer with IP address 206.186.25.7
- Using IE 3.02 on Windows NT Platform

DNS lookup enables 206.186.25.7 to be mapped to  
redpine.canadian.net

Can we use IP addresses to monitor growth in numbers of users visiting our web site?

Can we use domain names of visitors to monitor growth in accesses from countries?

# Caching

- ☺ Caching is important to speed up the Web
- ☺ JISC funds a national caching infrastructure for UK HE
- ☹ Caching makes it difficult to interpret web statistics:
  - User A requests file
  - Request goes to Institutional / National cache via local proxy
  - If not in cache, resource retrieved (hits generated) and kept in cache
  - User B request same file
  - Resource retrieved from cache (no hits generated)
  - Users C-Z all request same file. No hits generated

# Robots

You want robots to visit your Web site:

- AltaVista (and other indexing robots) to enable your resources to be found
- Auditing robots e.g. to validate links, to count size of Web
- Specialist robots used within research community
- Off-line browsers (are these robots?)

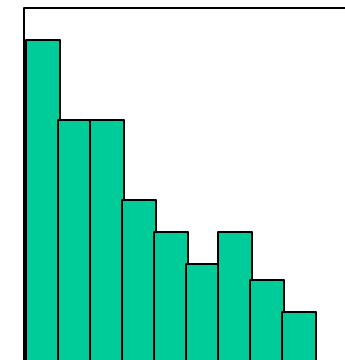
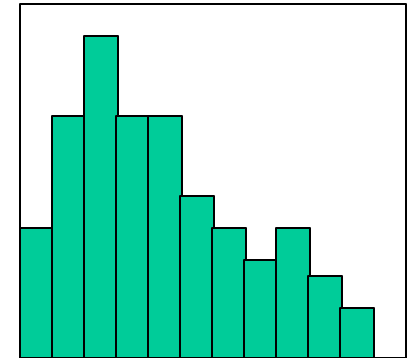
But:

- Robots generate hits
- Does a growth in the number of hits simply indicate a growth in the numbers of robots
- Some robots may revisit your website regularly

# One-Off Visitors

What do you think is the modal number of pages retrieved from a Web site in a visit?

- Research suggests that users use search engines to find resources, examine a Web site and then leave if its not of interest
- Does a growth in the number of visitors imply indicate a growth in the number of users of the Internet?



# *Tools*

- Can we conclude that Web statistics are meaningless?
- Would we say that TV viewing figures are meaningless?
- Web statistics need to be treated with caution.
- Web log analysis packages with data-mining capabilities can:
  - Indicate trends
  - Interrogate the data (e.g. strip out hits from robots)

# *Log Analysis Tools*

Many tools available:

- **Analog**: free, easily automated. However little data-mining capabilities and management graphs limited.
- **WebTrends**: Popular desktop package. Several versions. May be expensive for reporting on multiple Web sites.
- **Webaliser, aWebVisit, HitList**, etc. (see CD-ROM on many Internet magazines)
- Lists available at `<ipw.internet.com>` and `<www.yahoo.co.uk>`

# Externally-Hosted Services

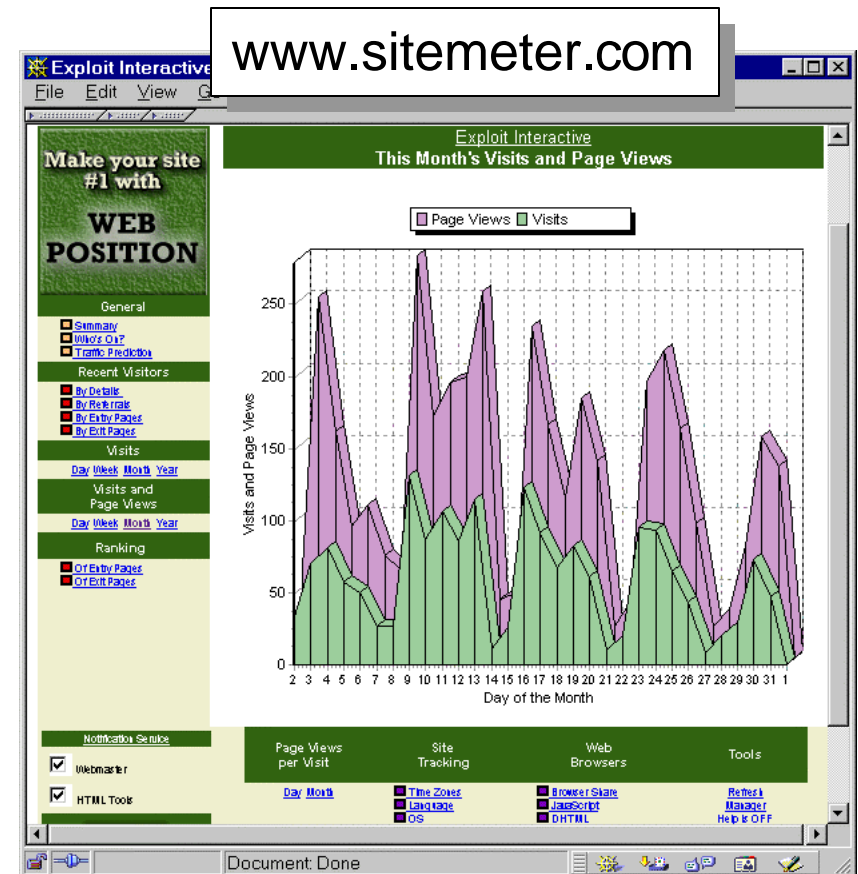
*Exploit Interactive* has been evaluating two externally-hosted statistical services: **SiteMeter** and **NedStat**.

## Advantages:

- No software to buy, install, configure and run or powerful PC to run software on
- No log files to manage
- Uses "*cache-busting*" images
- Can monitor extra features

## Disadvantages:

- Limited data-mining
- Ownership of data
- Dependency on external service
- Fails to monitor text browsers



The services can monitor client-side features, such as browser plugins, screen resolution, etc.

# *Other Indicators*

What other indicators may be of interest:

## **Links *To Your Site***

Indicators that people are interested in your service (and can deliver traffic)

## **Coverage By Search Engines**

Indicators that users can find resources on your Web site

## **User Feedback**

Comments, voting, etc.

## **Technical Indicators**

Browser support, server-uptime, etc

# Links To Your Site



- Search engines can be used to report on the numbers of links to a Web site
- **LinkPopularity.com** provides an interface to 3 search engines
- Monthly reports can be obtained
- Links are an indication of **potential** use of your Web site

A survey of the number of links to University web sites is available at <http://www.ariadne.ac.uk/issue23/web-watch/>.

EEVL used this approach to obtain sponsorship (nos. of links to EEVL was much larger than links to the sponsoring company).  
Would regular monitoring of links to your Web site be useful to you?

# Coverage By Search Engines



- Have you promoted your Web site?
- Can your Web site be accessed by search engines?
- Are you near the top of the search results?
- Search engines can report on their coverage of your Web site
- Coverage is an indication of **potential** use of your Web site

For information on how to ensure that your web site has been indexed see <http://www.exploit-lib.org/issue4/promotion/>

# Links As Performance Indicator

What are links used for:

- Internal navigation
- References

How many:

- Links on your Web site (internal and external)?
- How many broken links?

	Jan 2000	Apr 2000
No. of pages	906	1,135
No. of internal links	6,842	6,861
No. of external links	3,117	3,172
No. of broken links	0	0
No. of images	234	378

<http://www.exploit-lib.org/issue5/exploit-audit/>

## References

1. Ernst & Young Press Release  
URL: <<http://www.ey.com/news/releases/090998.asp>>
2. The British Library's Online Public Access Catalogue  
URL: <<http://opac97.bl.uk/>>
3. Tacoma Public Library's OPAC, Topcat  
URL: <<http://www.tpl.lib.wa.us/topcat/>>
4. Amazon Associates Programme  
URL: <<http://www.amazon.co.uk/associates/>>
5. Amazon.co.uk  
URL: <<http://www.amazon.co.uk/>>

- Can links provide a performance indicator?
- Should broken links to external resource in Web journal be fixed, flagged or ignored?

# User Feedback

- It is now much easier to obtain and analyse user feedback
- Feedback and voting systems can be installed free-of-charge

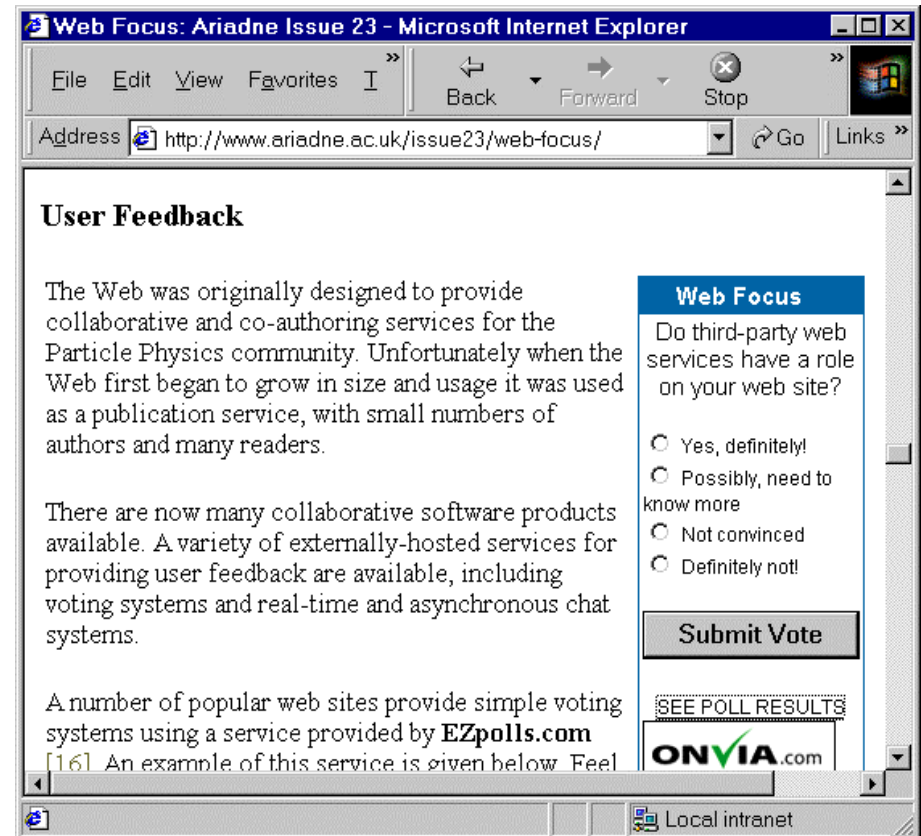
## Exploit Interactive

Do you:

- o Read all columns
- o Read 1-2 columns

Do you

- o Print out articles
- o Only read online



Feedback forms can be useful in quickly answering questions that can't be answered by Web log analysis e.g. do users print articles?

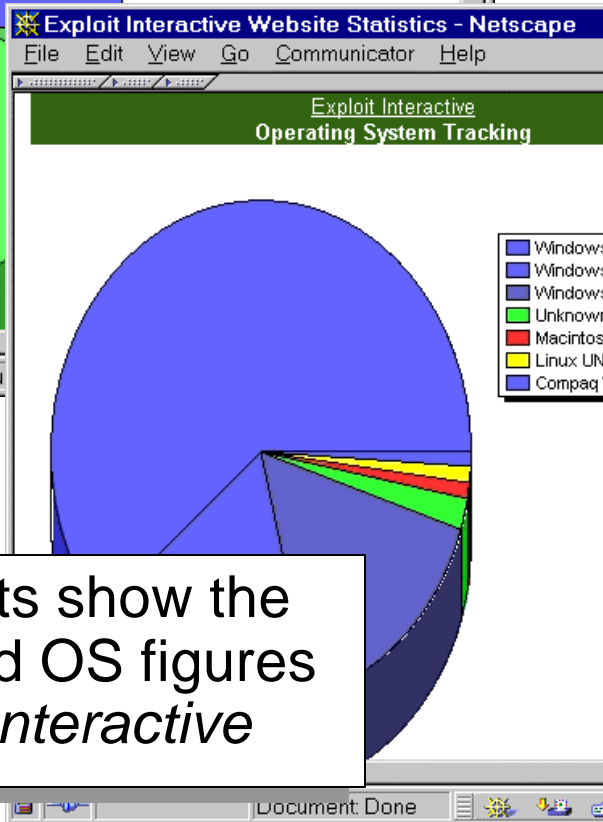
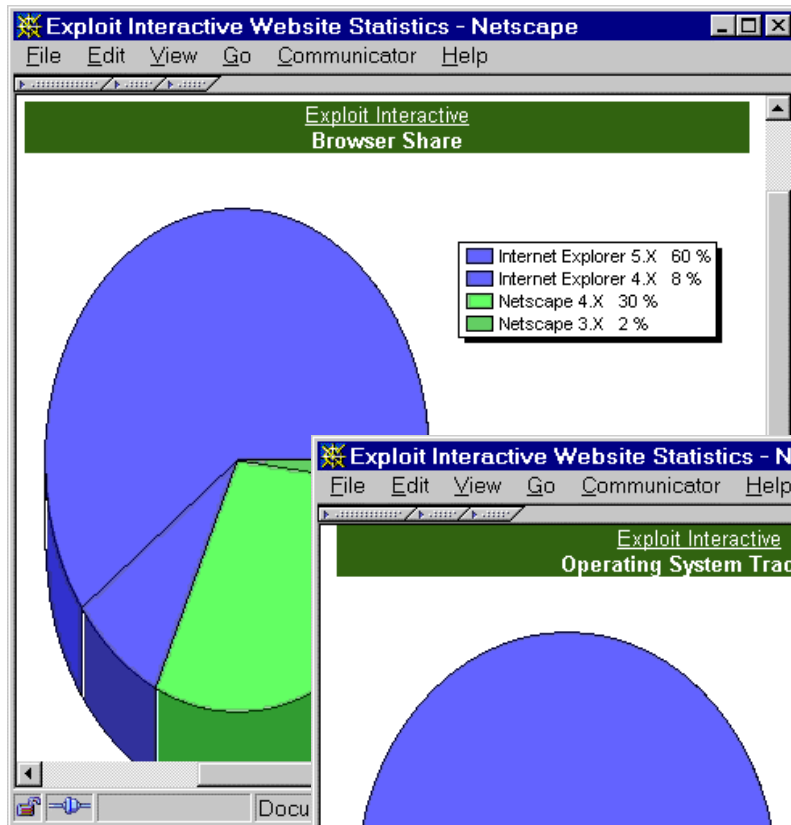
# *Technical Issues*

- *My software developers want to use Dynamic HTML to improve the user interface.*
- *I'd like to deliver articles in PDF format with a Shockwave interface but I don't know if users will have the plugins.*

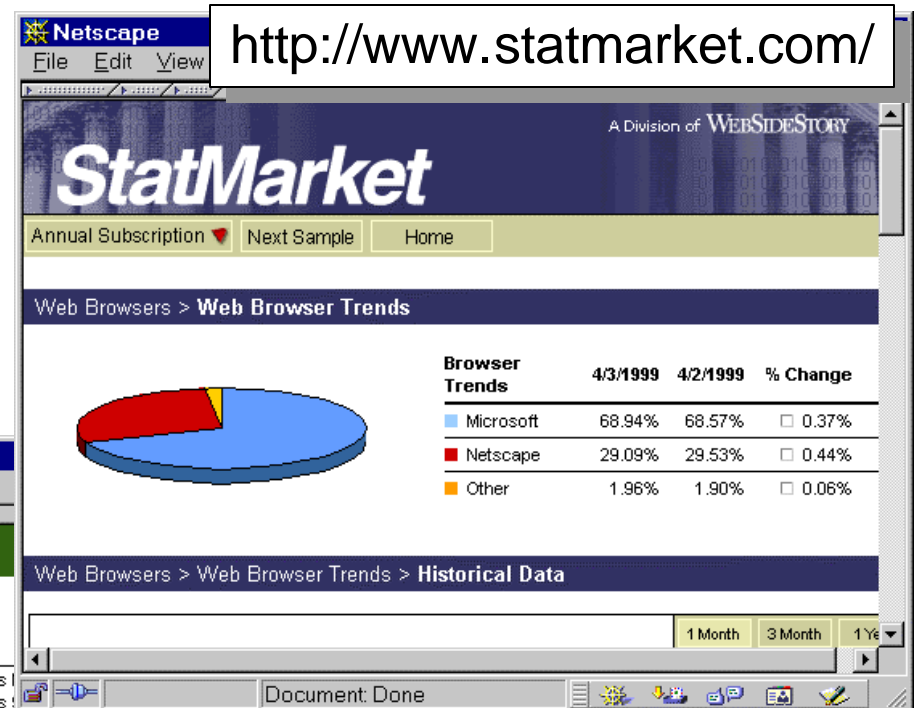
Nowadays developers face difficult choices when wishing to exploit new technologies.

- Information on browser profiles can be obtained from Web logs.
- Information on client capabilities and browser plugins can be obtained using, e.g., externally hosted services

# Technical Issues



These charts show the browser and OS figures for *Exploit Interactive*



**Statmarket** gives more comprehensive figures based on large nos. of visitors (40m) and Web sites and (100,000+)

# Conclusions

Roger Brown admitted that:

*"There are technical issues that may cause problems [caching, dynamic IP addresses, confidentiality]"*

This talk has reviewed some technical issues

- Web statistics can be difficult to interpret
- Analysis of Web statistics **is** needed
- Think about the tools you will need (and the resource implications in using them)
- Besides analysis of log files there are other performance indicators which may be of use
- Analyses will also help with in monitoring the performance of your Web site and planning future developments