RePAH:
Reseach Portals in the Arts and Humanities
A user analysis project

Appendix A5:
Work-Package 2: Web-Log Analysis Report

WP2 Report prepared by Mark Greengrass and Robert Ross

http://repah.dmu.ac.uk/report

Published by:

HriOnline

for

The RePAH Project

Knowledge Media Design
De Montfort University
Portland 2.3a
The Gateway
Leicester LE1 9BH

&

The Humanities Research Institute
University of Sheffield
34 Gell St
Sheffield South Yorkshire S10 2TN

ISBN: 0-9542608-8-0

Also Available at http://repah.dmu.ac.uk/report

September 2006

Contents

The RePAH Project

In July 2005, the RePAH Project was commissioned to carry out a survey of user-needs for information portals in the Arts and Humanities by the AHRC ICT in Arts and Humanities Programme. It sought to understand how the arts and humanities research community finds and exploits the internet resources it needs.

In order to do this the RePAH project:

- Examined the existing literature on user needs with regard to web gateways and portals,
- Analysed the web-logs from the Arts and Humanities Data Service (AHDS) subject centres and the Resource Discovery Network's (RDN) humanities and arts web hubs (prior to July 2006 these were known as Humbul and Artifact, but now have been harmonised into Intute-Arts and Humanities)
- Conducted focus groups, interviews and a Delphi exercise with members of the arts and humanities community
- Developed and tested a paper-based demonstrator for a managed research environment to explore possible ways forward with regard to web-based research resources.

The project was carried out in 7 work packages:

- WP1 RePAH Online Questionnaire--this report examines an online survey of the Arts and Humanities Community's use of web resources.
- WP2 Web-Log Analysis--this report analyses web-logs from several of the Arts and Humanities Data Service subject centres as well as Humbul and Artifact of the Resource Discovery Network (now Intute).
- WP3 First Focus Group--this report studies the responses from a series of five focus groups conducted at the University of Sheffield and three interviews from DeMontfort University. Respondents discussed their use of web resources in general and portals in particular.
- WP4 Delphi Exercise--this report considers the results of a Delphi exercise conducted around the feasibility of various web-based tools.
- WP5 Demonstrator of a Managed Research Environment--this report is an exploration of a paper-based demonstrator of a variety of features that might be applied as portlets and used by the Arts and Humanities research community.
- WP6 Phase II User Trials of Portal Demonstrator--this report brought the paper-based demonstrator to scholars in eight subjects within the Arts and Humanities community and asked them to evaluate the features and functionality of possible portlet tools.
- WP7 Intute in Light of this Report--this report explores Intute-Arts and Humanities with reference to the features and functionality explored in the paper-based managed research environment demonstrator.

Additional appendices within the RePAH Project report include an overview of the Arts and Humanities research community [Appendix A2], and a review of the literature relevant to user requirements for digital resources and web-based research facilities [Appendix A3].

This appendix reports on Work Package 7 which examines Intute-Arts and Humanities with reference to the features and functionality explored in the paper-based managed research environment demonstrator, as well as some the data harvesting of the AHDS by Intute.

To see the full report and the other appendices see **http://repah.dmu.ac.uk/report**.

A5.1 Introduction

This report is prepared on the basis of web-log data provided by the AHDS and RDN subject portals over the following periods:

AHDS Central Server Access Logs [February-September 2005]
AHDS History Server Logs [*data still under analysis*]
AHDS Visual Arts Logs [*data still under analysis*]
Humbul Server Logs [January-December 2005]
Artifact Server Logs [October-December 2005]

The data provided is difficult to analyse, and for a number of reasons. Because of the distributed nature of the AHDS service, many users access their online resources through the particular service rather than the central server. In the case of AHDS History, however, its server logs are rolled up with the Data Archive. They were able to strip out for us the traffic that was not relevant to our needs. In the case of AHDS Archaeology, where the traffic is heavily influenced by non-HEI needs, we did not feel that the analysis would be relevant. In all the AHDS data, there is internal traffic between the AHDS sites that we have not been able to strip out from our analysis. In the case of Artifact, the server logs were not archived prior to October 2005 and so we have only a fragment of the picture to go on. No web-log data was forthcoming from AHDS Language, Literature and Linguistics, or from AHDS Performing Arts. Although the JISC requires some statistics from web-log activity to be published from the services that it supports, they are not published in a coherent fashion. We would expect to be able to recover sample statistics of the following from the Annual Reports of the services as 'surface-measures' of user traffic:

- Site Visits
- Total Page/\Item Views per month
- Average No of Pages/Items consulted per day
- A statistical reflection of particular function-usage. In the case of AHDS this may be collection downloads. In the case of Intute, it may be registered users.

In reality this is not the case. The statistical analysis of these data-logs was undertaken for the project by Dr Paul Huntington of CIBER, UCL. Based on that evidence, this Report has been prepared by the Project, which is responsible for its conclusions.

A5.2 Web-Log Analysis Methodology

Web-server logs record simple traffic statistics and data such as number of page requests per month and originating addresses of page requests. Deep-log analysis (DLA) uses web logs from a server and, following a normal process of analysis, links the information with site-user profiles, or demographics, to produce a 'deeper, more meaningful data' picture of overall site usage. It is a four stage process:

1) Data definition, recording procedure and statistical significance are agreed.
2) A series of pre-defined metrics are used to ensure the data is analysed in

line with organisational goals and policies.

3) Enrichment of usage data with demographic data.
4) Identification of questions concerning information seeking behaviour that need to be asked by questionnaire, interview or observation.

The **working metric definitions** used in this report are:

User: A user is effectively a computer; sometimes that computer represents an individual, (i.e. a professor in his office), in other cases a number of people (i.e. students in the library). User identification can be based on a combination of "IP" number and browser details or by use of cookies.

Sessions: They are identified in the logs by a session identification number. Logs include a session beginning tag and a session ending tag, which enables time calculations as well. Items viewed/requests made. The key usage sub-metrics are: type of items viewed, number of items viewed in a session and return visits. These sub-metrics offer extremely good platforms for characterising and comparing the information seeking behaviour of sub-groups of users because generalisations based upon millions of users, while sounding impressive, can prove very misleading, camouflaging possibly big differences between individual user groups, like that between students and professors. A complete item might be all the pages, charts etc. from an article, and this is recorded as a single item and hence the digital library logs are quite different from traditional server log files that record pictures and text documents separately. The logs may also recorded views to the home page and a returned search screen.

Items viewed/requests made: This is defined as a 'complete' item returned by the server to the client in response to a user action. Typically this might be a menu page or a search screen. Logs do not record all items viewed by the client since, once seen, the item will be cached to the clients' machine. If a client returns to view the page that view will be made from the copy in the cache and not from the server. A page will remain in the cache for a variable length of time.

Robots: Web-logs often record 'robot' users to sites. These are mechanical agents, used mainly be search engines, to index web-pages. Robots should report to the site's 'robot.txt' document, which identifies the accesses by this IP number as a robot and informs the robot as to what pages to index. Several robots were identified in the course of the CIBER analysis which did not conform to that convention and these were also stripped out. Views to automatic feeds were also stripped out from the analysis.

Internet protocol (IP) numbers: These are identities that facilitate an access to view items on the internet. IP numbers also act like registration numbers and can be used to access additional information about the user in a process called reverse DNS (Domain Name Server) lookup. This process, when successful, reveals the user's organisation name, the type and the location of organisation. However many users mis-register. So, for example, a UK user may register for a US-style domain name or a net-provider will often register as a commercial organisation. Further, not all IP numbers can be identified by this process. Though these difficulties limit, they do not negate the usefulness of such data. Academic institutions, in particular, rarely mis-register either their location of organisational grouping.

Referrer Links: These are the identified site link from which the user accessed the site being investigated.

A more powerful way of examining the number of items viewed is to categorise search sessions by the **number of items viewed**. This is called 'site penetration'. Research on the subject has shown that many web users graze lightly, examining just a few items/pages before they leave with no substantial content consumed, although knowledge might have been gained [Nicholas et al, 2004c]. High levels of penetration can be assumed when there is evidence of:

a) 'natural movement' through the site
b) a massive choice of data on offer
c) the investigative nature of some information-seeking
d) the presence of an embedded search engine and other retrieval aids.

Returning to a site also constitutes evidence of conscious and direct use. However, research on that subject suggests that people view only a small proposition of a site's contents and, further, return to it very rarely [Nicholas et al, 2004c]. In theory, how frequently they return should depend on the nature of the site – a newspaper site, for instance, might be expected to obtain more return visits. But there is no natural frequency for any particular kind of site. But, in the case of academic information-seeking behaviour, one might expect a more developed repeat-behaviour (in order to satisfy reiterated information needs) than other internet information-seekers. In general, the ability to generate useful information via DLA relies on adding user demographic data (e.g. occupation, subject specialism), via data obtained from a subscriber database (preferable) or online questionnaires (less preferable, since user data cannot be mapped so closely onto usage data). Of course, logs and user databases enable us to map the digital environment more accurately but provide little by way of explanation, satisfaction and impacts.

A5.3 Humbul Web-Log Analysis

A5.3.1 Items Viewed:

Figure 1 looks at the daily number of items viewed over 2005. Usage generally fell within the range of about 2,000 to 4,000 daily views at weekends and between 6,000 to 8,000 item views on weekdays. The year started relatively low between 3,000 to 6,000 daily views and then rose to a peak in early to mid February of 8,000 (weekdays) before declining and reaching a relatively low figure of 6,000 (weekdays) in late march early April. Use recovered over the next month before entering a decline over May to early September. Over this period the range between weekday and weekend use declined. Use picked up fairly strongly from mid-September to December.

Figure 1: Daily number of items viewed 2005.

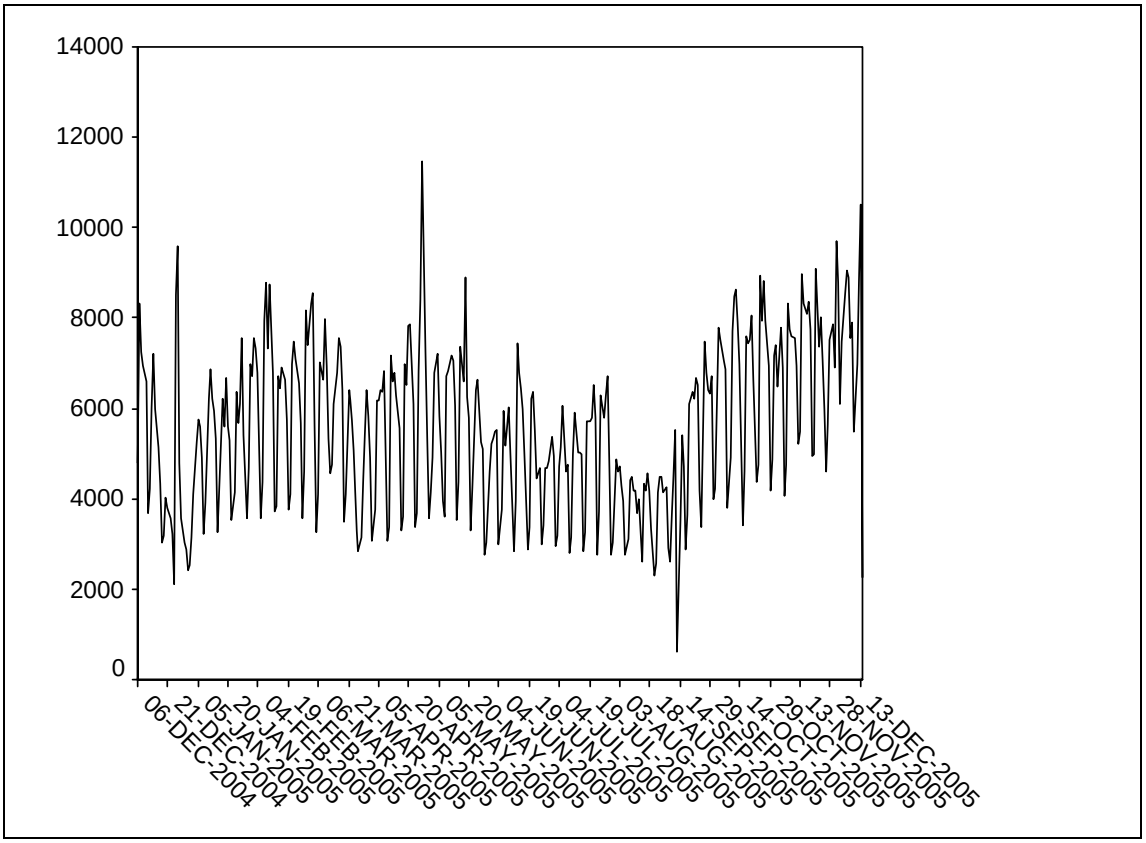


Figure 2 gives the percentage distribution of usage over day of week. Usage at the weekend was about two thirds of usage on weekdays.

Figure 2 The percentage distribution of usage over day of week

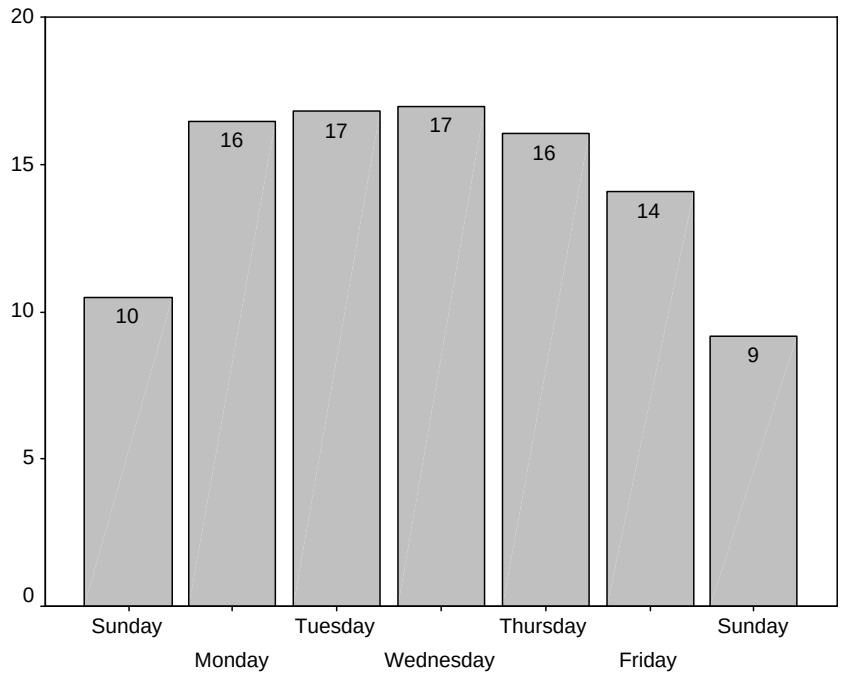
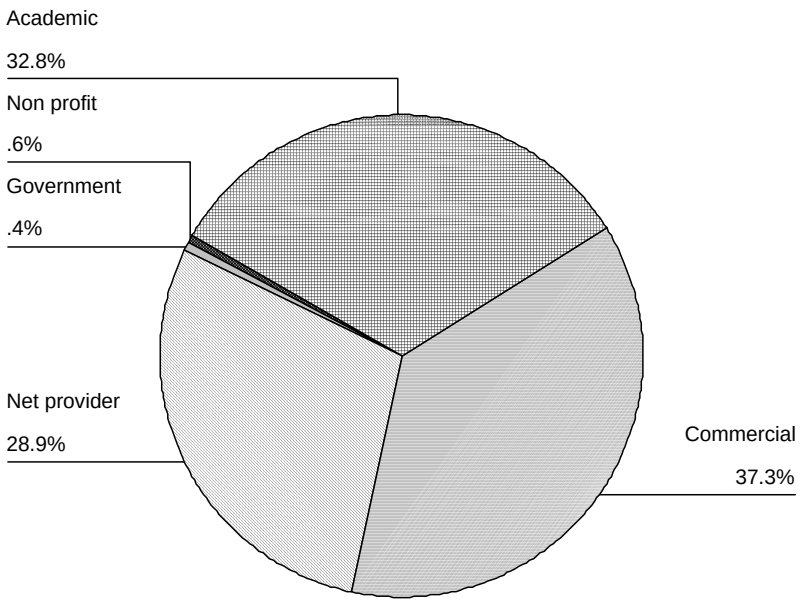


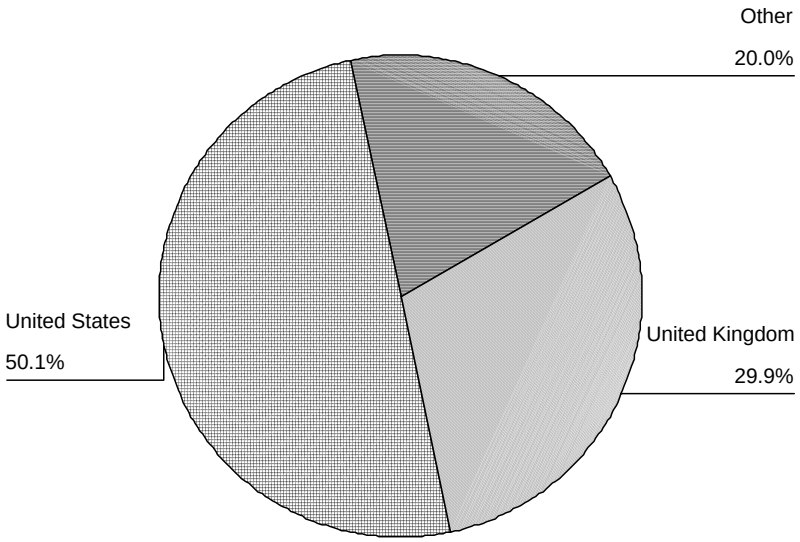
Figure 3 gives the share of usage broken down by organisation type, as retrieved from the reverse DNS lookup. About a third (32.8%) of usage is attributed to academic users, just over a third (37.3%) to commercial users and over a quarter (28.9%) of usage is attributed to net providers.

Figure 3 The share of usage broken down by organisation (DNS) type



The main two countries from which users were using the Humbul services by reverse DNS look up was the US (50.1%) and UK (29.9%). The group other includes all countries that each made up less than 2% of usage.

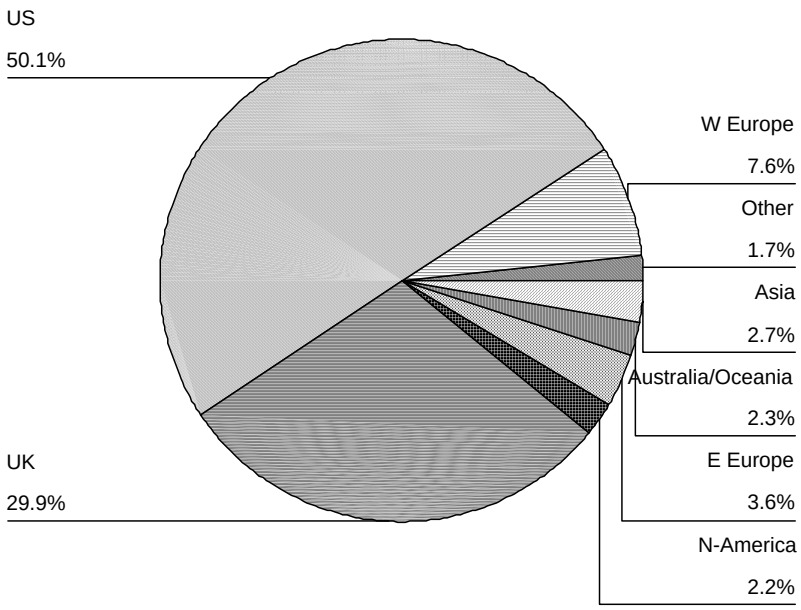
Figure 4 The share of usage broken down by user (DNS) country code



Other all countries accounting for less than 2% of use

Figure 5 gives the same information but breaks it down according to the country location by world regions. Western Europe (excluding UK) made up about 7 to 8% of usage while Eastern Europe made up between 3 to 4%.

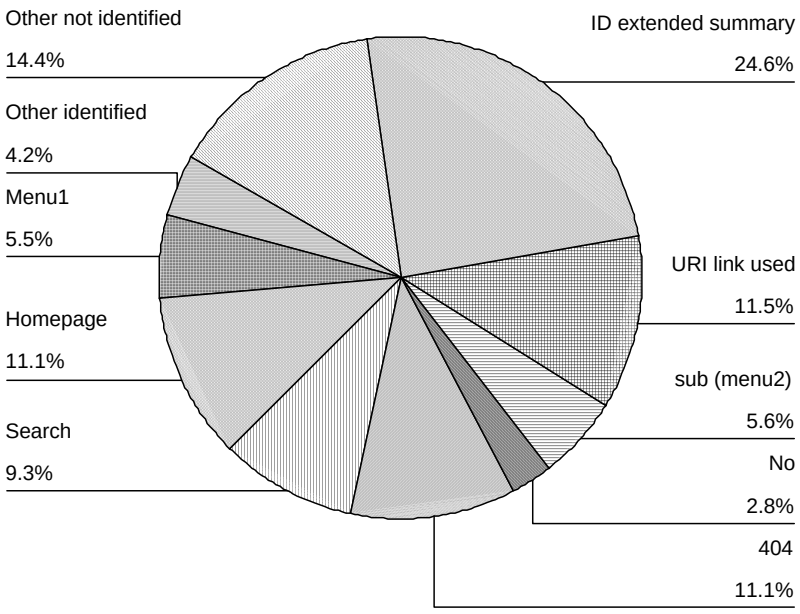
Figure 5 The share of usage broken down by user (DNS) country codes grouped into world regions.



Other: grouped regions that individually account for less than 1% of usage.

Figure 6 gives an idea of item-type viewed. The homepage, that accounted for about 11% of views, is the opening Humbul page that is viewed on opening the site at www.Humbul.ac.uk. This page includes a variety of subject links. It is defined here as the 'menu1' page. There were approximately 5.5% views to this page. This page offers links to sub-categories within the subject. Clicking on any of these links takes the user to a sub-menu (menu2) page – which made up 5.6% of views. The menu2 page offers users a list of resources to link to. Under each resource is a reduced summary, a link to the extended summary and a link to the resource. Should the user opt for the extended summary the user is taken to the ID (extended summary). This gives an extended summary of the resource and a link to the resource. About a quarter (24.6%) of items viewed were to the extended summary. About 11.5% of users activated the link to the external URI (universal resource indicator). Rather than use the menus, users may alternatively activate the on-site search facility. About 9.3% of usage related to items where the word-search appeared. Other identified items were to do with 'jobsearch' and other items that appear on the left hand menu of the Humbul homepage. Other unidentified pages made up 14% of usage. Most (about 75%) of the other unidentified group was accounted for by the following item names: describe (19%), user (14%), vts (9%), about (8%), help (8%), submit (8%), topics (7%), output (5%).

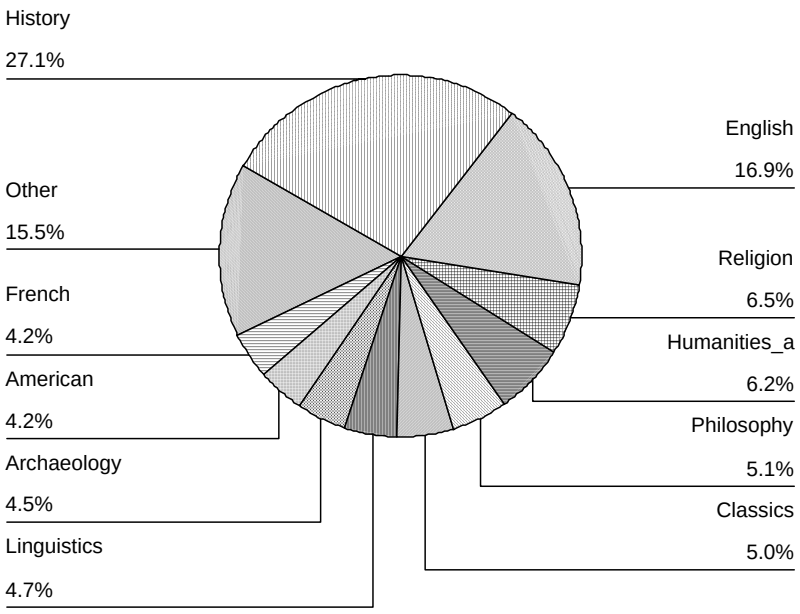
Figure 6: Distribution of item type viewed



Other identified groups items where each item accounted for less than 2% of usage.

In terms of menu1 usage the following gives an idea of subject-usage. History is the most popular subject and about a quarter (27.1%) of subject-use relates to this. Other popular subjects are English (16.9%), Religion (6.5%), Humanities_a (6.2%) and Philosophy (5.1%).

Figure 6: Distribution of subject item (Menu1) viewed



Other groups subjects that each account for less than 3% of use.

The Humbul logs also give the site address and directory of the linking resource. If the user decides to visit a resource, the logs record the site visited. About 11.5% of items viewed were users who then actively clicked through to the resource. Throughout the year, 7,463 separate resources were accessed via the Humbul site. The following Table lists the top 40 and the accompanying spreadsheet gives the full list.

Figure 7: Top 40 resource sites accessed via Humbul

URI Site	Number	Percentage
www.bbc.co.uk	4166	1.5
www.wsu.edu	2473	.9
www.geocities.com	1969	.7
www.nd.edu	1517	.6
ads.ahds.ac.uk	1216	.4
www.bl.uk	1047	.4
www.arts.ed.ac.uk	1042	.4
www.pbs.org	1031	.4
www.emule.com	936	.3
memory.loc.gov	836	.3
www.fordham.edu	813	.3
www.shef.ac.uk	811	.3
www.channel4.com	789	.3
www.newadvent.org	713	.3
www.llgc.org.uk	680	.3
www.spartacus.school	659	.2
www.luminarium.org	659	.2
etext.lib.virginia.e	649	.2
uk.cambridge.org	643	.2
www.ucl.ac.uk	636	.2
www.iwm.org.uk	624	.2
www.loc.gov	614	.2
ccat.sas.upenn.edu	606	.2
www.gre.ac.uk	599	.2
www.archives.gov.on.	575	.2
www3.oup.co.uk	573	.2
www.archives.gov	563	.2
www.accd.edu	560	.2
www.nationalarchives	559	.2
www.georgetown.edu	546	.2
www.hti.umich.edu	540	.2
www.sas.ac.uk	536	.2
www.kb.nl	520	.2
etext.virginia.edu	506	.2
www.bu.edu	504	.2
www.stoa.org	503	.2
history.hanover.edu	499	.2
raven.cc.ku.edu	490	.2
learningcurve.pro.go	485	.2
www.17thc.us	479	.2
		12.6%

In terms of referrer link, about 40% of use related to users coming in via Yahoo, 20% via Google. Other sites include Wanadoo (3.2), ox (4.9), RDN (4%), Altavista (1.8) and the BBC (1.4). There is a specific reason for the apparently disproportionate number of user coming to Humbul via Yahoo. We understand that Humbul exposed its metadata via OAI to Yahoo for them to index their aggregated collection of harvested metadata. As a result

Humbul's metadata records are high in Yahoos rankings. Yahoo is a commonly used as commercial search engine of choice, particularly among non-academics in North America. So far as we are aware this is the only example of OAI metadata being made available for harvesting by the commercial search engines from the service providers. Its significant impact upon usage patterns should be noted.

A5.3.2 Part 2 Session level analysis

The following relates to the number of sessions. The site attracted between 1,500 (weekend) to 2,500 (weekday) sessions a day. The pattern of session over the year followed the same pattern as for items viewed.

Figure 8: Daily number of sessions - 2005

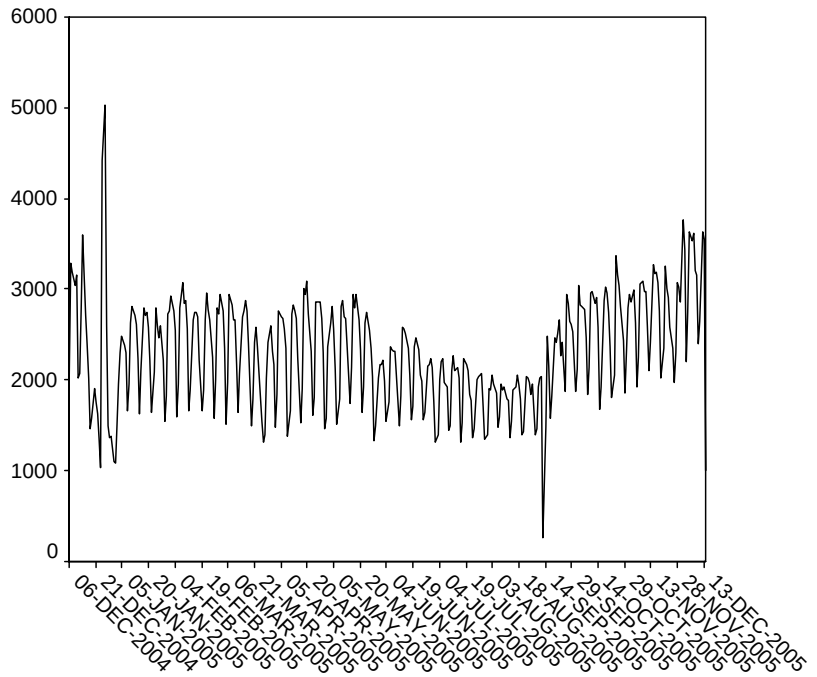


Figure 8 gives the same distribution but as a percentage.

Figure 8: Daily number of sessions - 2005

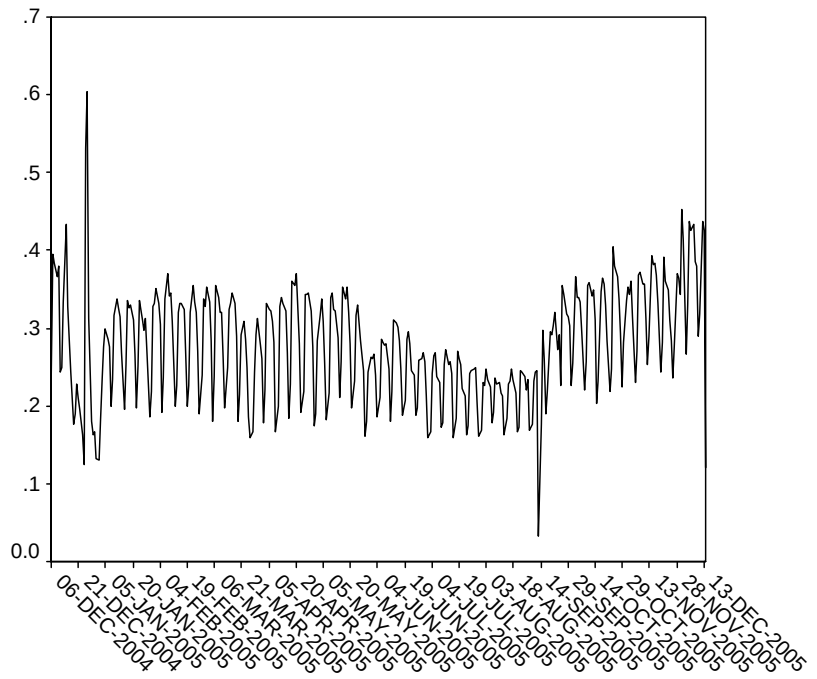


Figure 9 gives the number of sessions for each month for 2005.

Figure 9 The number of sessions for each month for 2005.

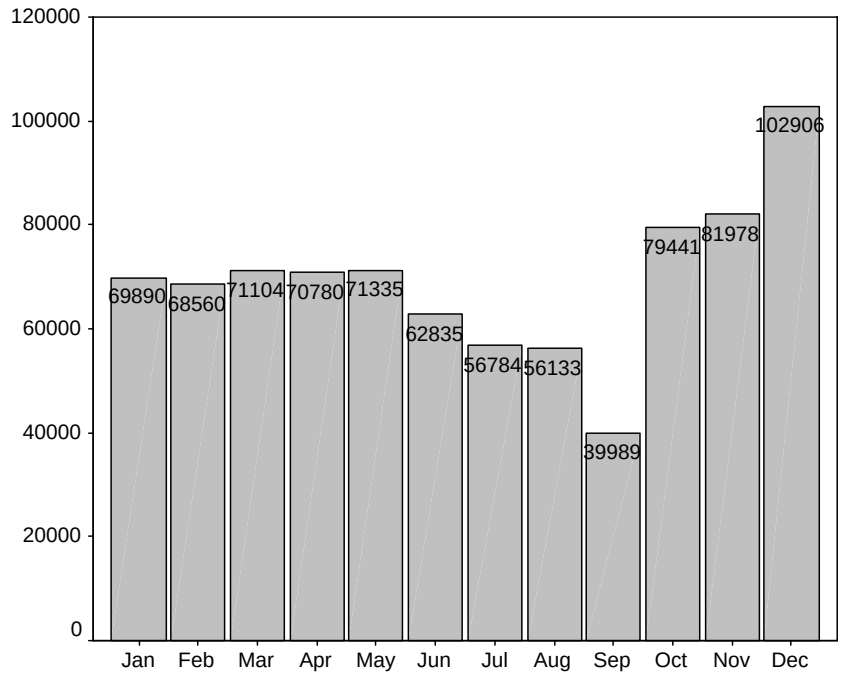


Figure 10: location of user as given by DNS registration details.

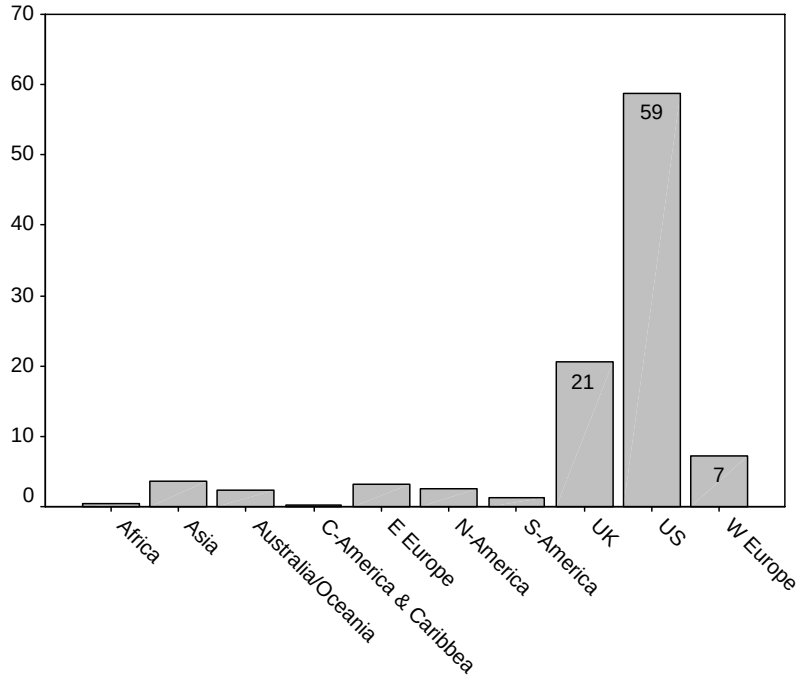


Figure 11 gives the organisation type of the user as given by the DNS registration details. Under a quarter (22%) of sessions were attributed to academic institutions. Most were attributed to either commercial (42%) or net provider (35%) organisations.

Figure 11: Organisation type of user as given by DNS registration details

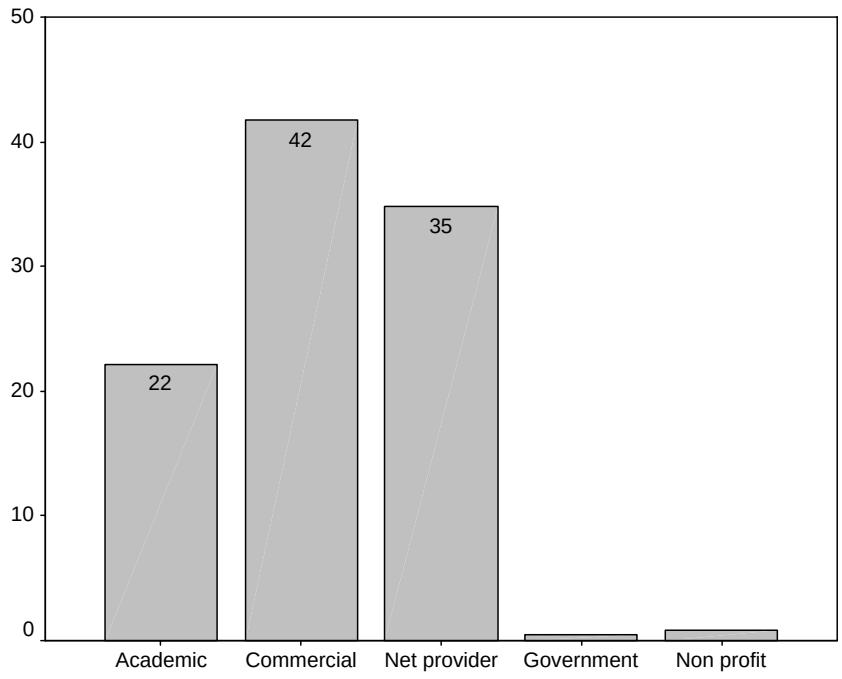


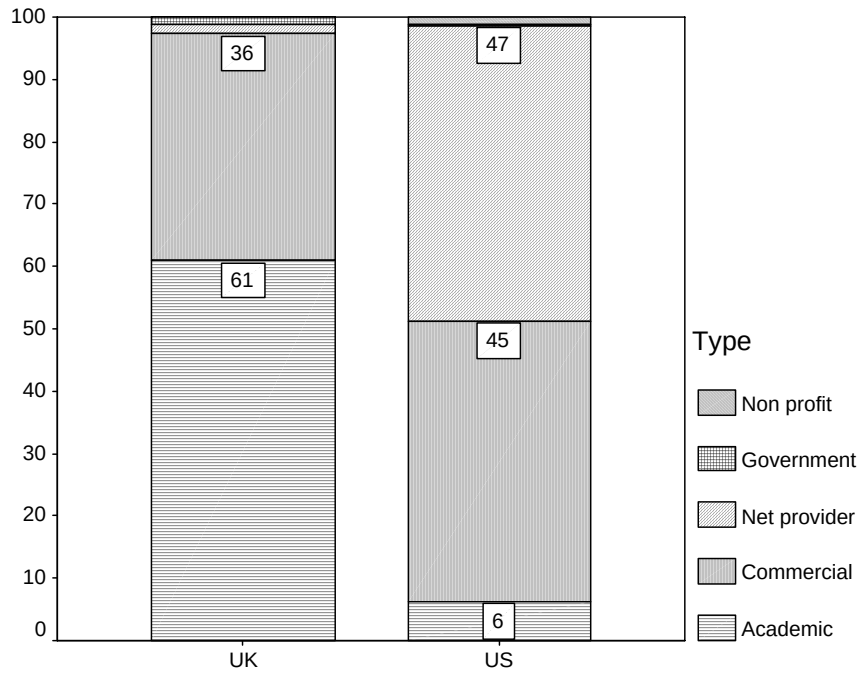
Figure 12 gives the list, top 30, of user academic organisation codes. The most important by some margin is Oxford (ox). About a quarter (24.9%) of sessions are attributed as coming from that source.

Figure 12 The top 30 user academic organisation DNS codes

Academic code	Number of sessions	Percentage of academic sessions
ox	27706	24.9
cam	2436	2.2
nottingham	1887	1.7
uea	1883	1.7
unimelb	1729	1.6
ucl	1702	1.5
dundee	1289	1.2
leeds	1202	1.1
le	1173	1.1
bham	1131	1.0
soton	1011	.9
mmu	996	.9
gla	957	.9
bris	955	.9
bton	905	.8
uu	896	.8
york	847	.8
kcl	845	.8
man	745	.7
hw	697	.6
shef	686	.6
ex	666	.6
ed	615	.6
dur	598	.5
uni-leipzig	573	.5
virginia	572	.5
open	571	.5
glam	558	.5
shu	554	.5
		50.8

In looking at the distribution of type of user over location (US and UK) it can be seen that while over half (61%) of UK user sessions were academic only 6% of US user-sessions were from academic institutions. US commercial users, who made up 45% of sessions here, were btcentralplus (19.4%), AOL (18.8%), btopenworld (4.7%) and many users of these organisations will have been used by UK users to access the Internet. In terms of UK-classified organisations internet access facilities were provided by Cable and Wireless (54%), blueyonder (23.3%), demon (4.9%).

Figure 13: Distribution of type of user over location (US and UK)



Referrer links were only recorded for 49.9% of sessions. For a third of all sessions (but 69% of sessions where a referrer was identified) the user accessed the Humbul site via a search engine. In terms of search engines used Yahoo made up about two thirds (64.3%) followed by Google (28.5%), Altavista (2.6), BBC (2.1%) and MSN (1.5%).

The following tables the distribution if a search engine was used by type of user (as recorded in their DNS). Where a DNS lookup or a referrer link was not found, the evidence has been excluded. Academics were least likely to use a search engine; but just over half did. Users coming in via a net-provider were most likely to have done so (76%). About two thirds of the ‘commercial users’ accessed the site via a search engine.

Figure 14: The percentage share distribution if a search engine was used by type of user by DNS registration.

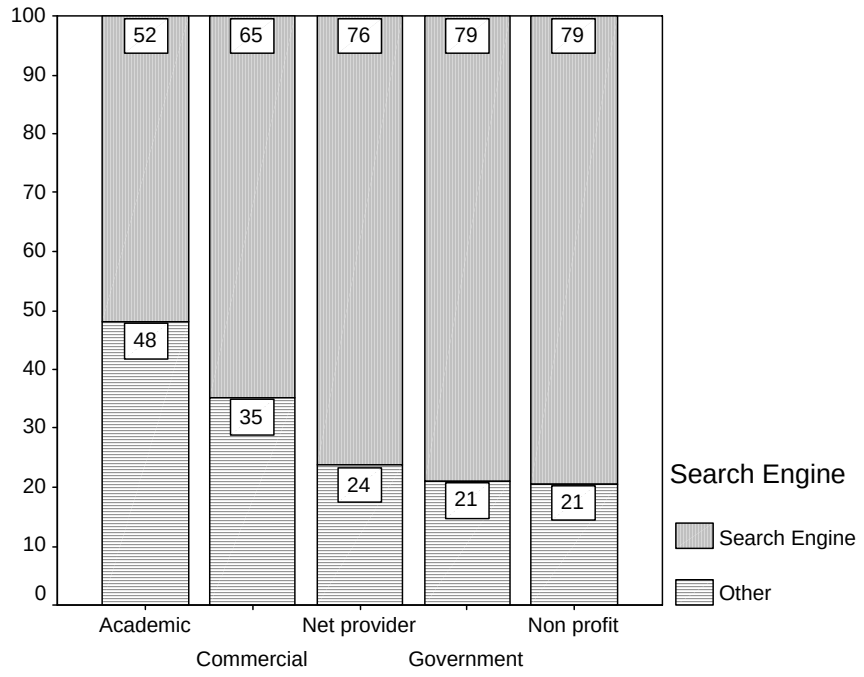
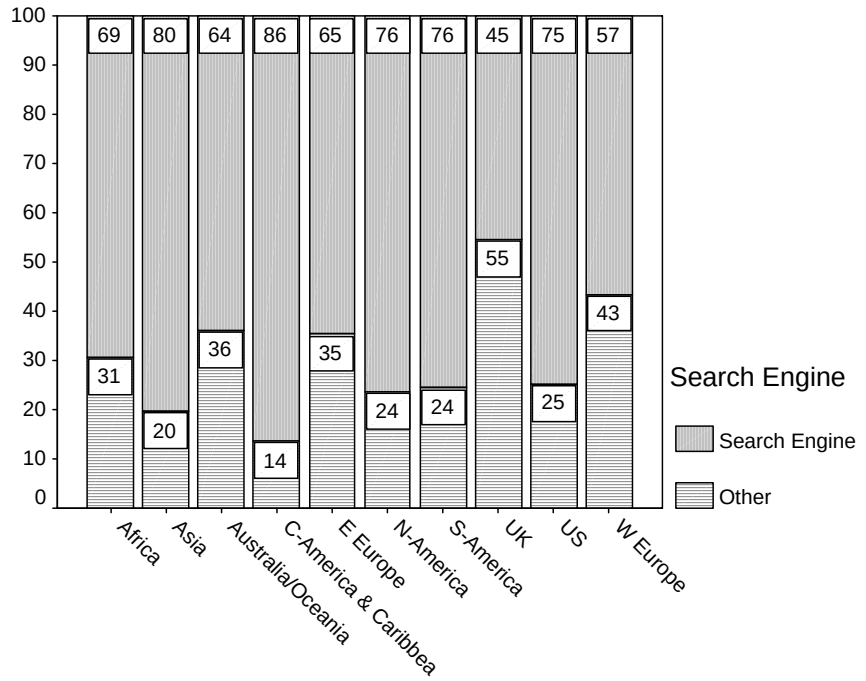


Figure 15 gives the same information but over a regional location. The UK had the lowest use of search engines, with just 45% of UK users accessing the site using a search engine. This is precisely what one would expect, however, since the UK also had the highest usage by academic users.

Figure 15: The percentage share distribution of if a search engine was used by country of user by DNS registration.



In terms of the number of pages viewed in a session, about 38% of sessions viewed more than one page; an estimated 25% viewed 2 to 3 pages; 10% 4 to 10; and 3% 11 or more pages in a session.

Figure 16 gives the percentage distribution of the number of pages (grouped) viewed in a session.

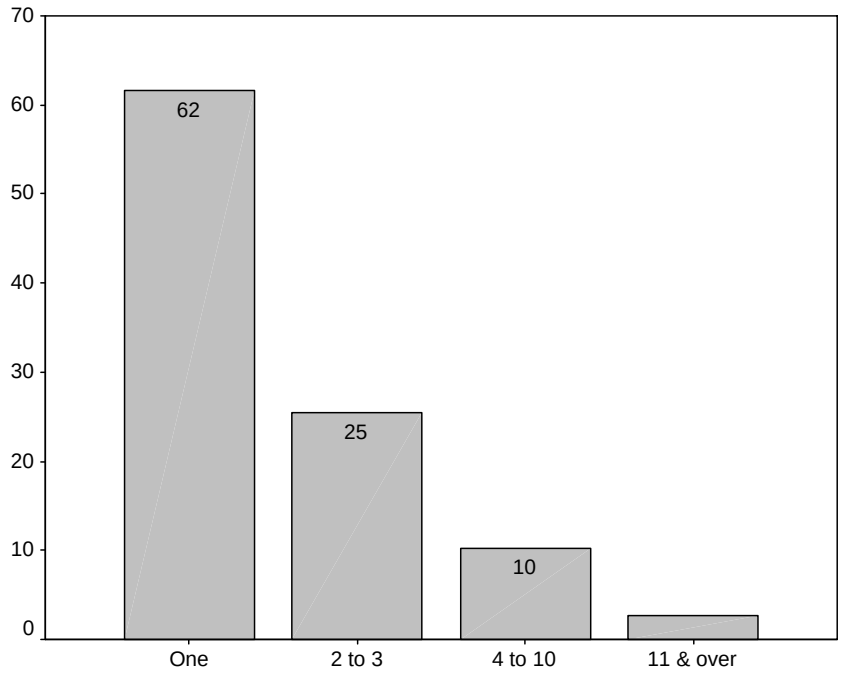


Figure 17 gives the distribution of views in a session by DNS organisation type of user. Net (63%) and commercial users (60%) were most likely to view just one page while academic users (54%) were least likely to do so. The web-logs confirm, in other words, what one would have expected: academic users tended to be the more serious users.

Figure 17 the percentage distribution of views in a session (grouped) by DNS organisation type of user

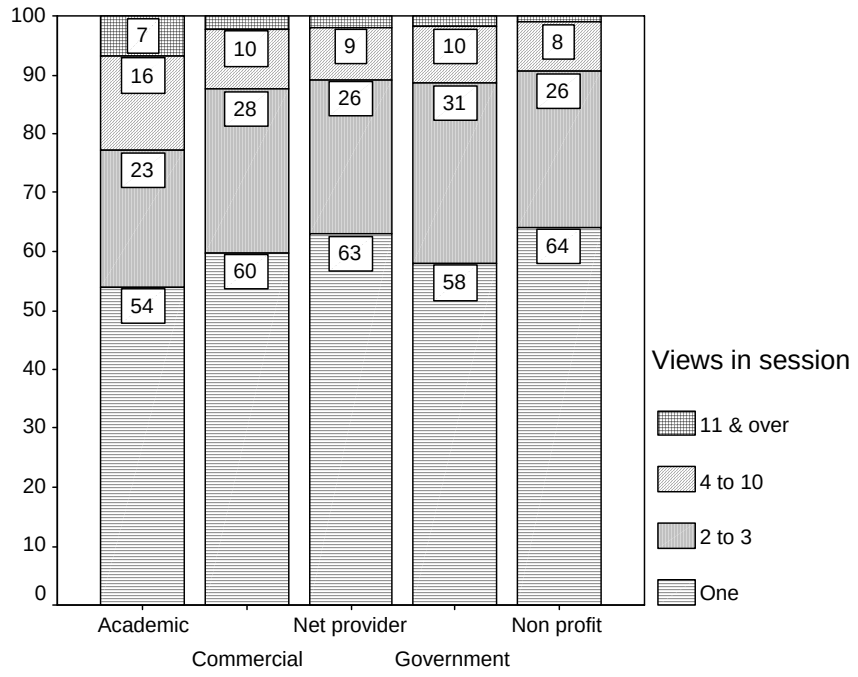


Figure 18 gives the distribution of views in a session, measured in terms of users who had accessed the site via a search engine during the session. Those viewing just one page (73%), or just 2 to 3 pages (74%), were more likely to have come in via a search engine. Sessions where more than 4 pages were less likely to have accessed the site using a search engine. It is possible, of course, that these sessions consisted of viewing more menu pages rather than penetrating to the resources in the collection. However, it is also clear that some users were coming in via a search engine and just browsing one or two pages, and then leaving again.

Figure 18 The distribution of views in a session by if the user had used or accessed the site via a search engine during the session

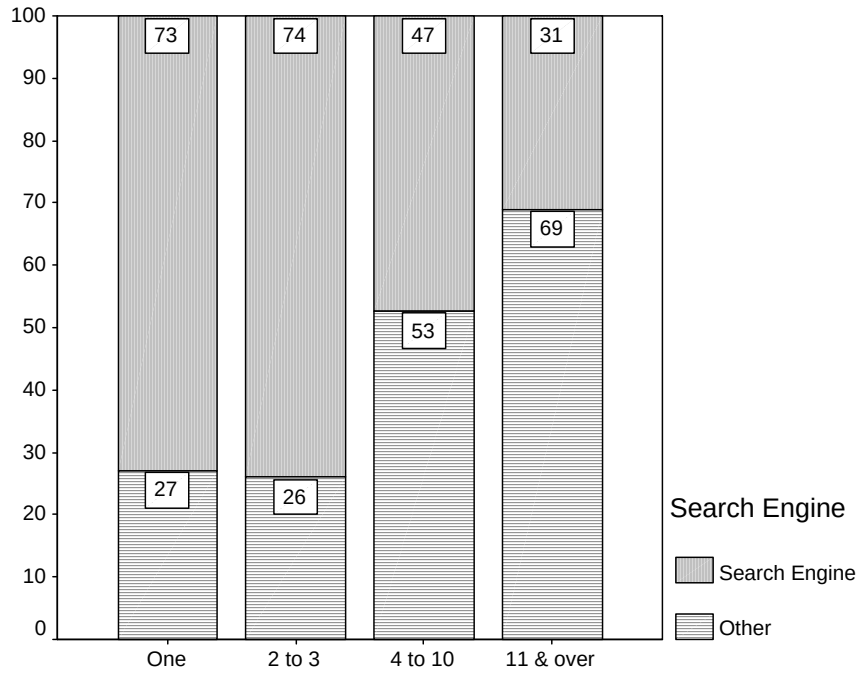


Figure 19 gives the distribution of views in a session by DNS country type of user. US users (63%) are most likely to view just one page; UK users (52%) were least likely to do so.

Figure 19 the percentage distribution of views in a session (grouped) by DNS country type of user

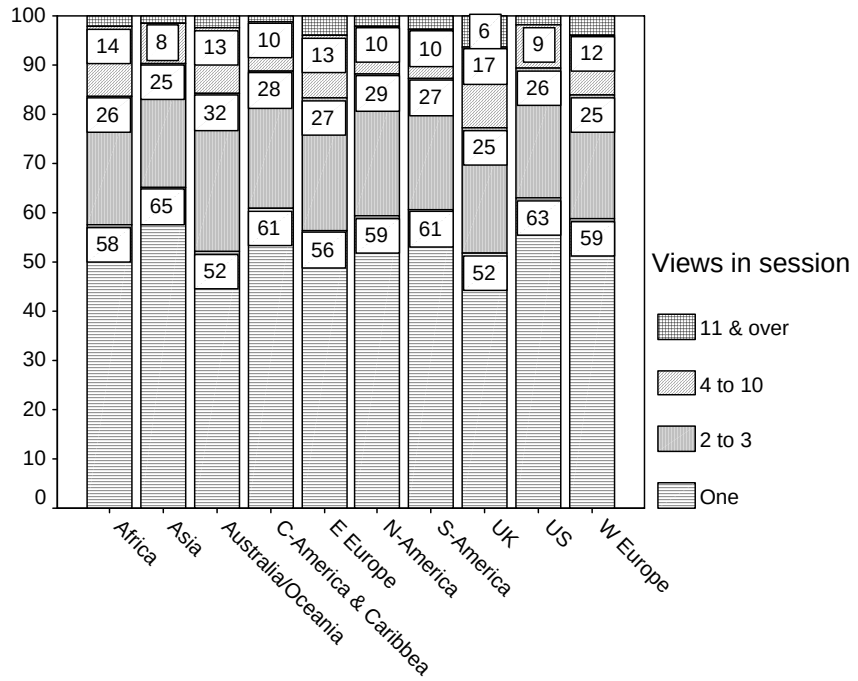


Figure 20 gives the percentage distribution of session time, grouped by DNS organisation type of user. Academics were recorded as having longer sessions. 31% had sessions lasting over 3 minutes. The comparable figure for commercial session users is 21%, with 19% for net-organisation based sessions.

Figure 20 the percentage distribution of session time (grouped) by DNS organisation type of user.

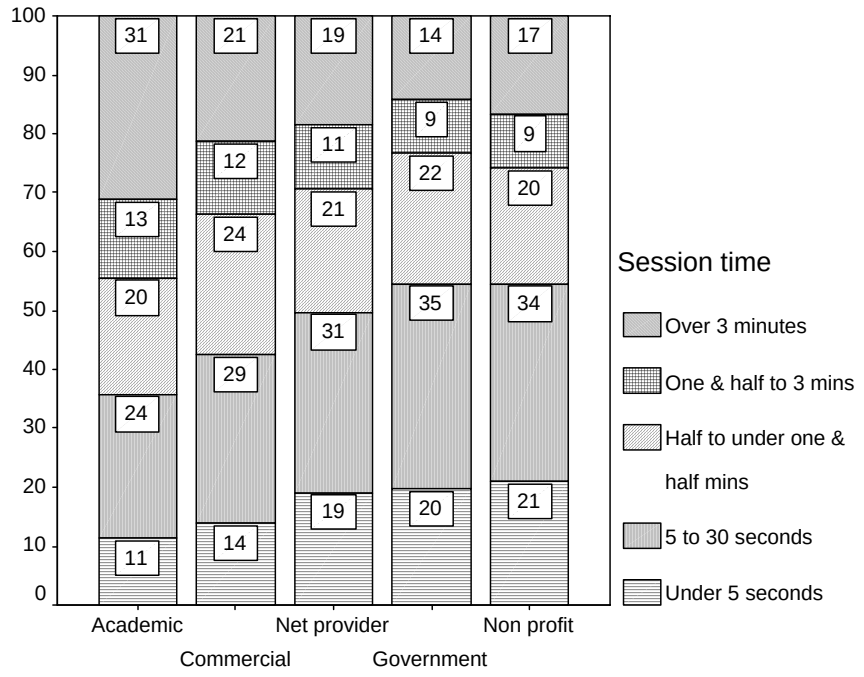
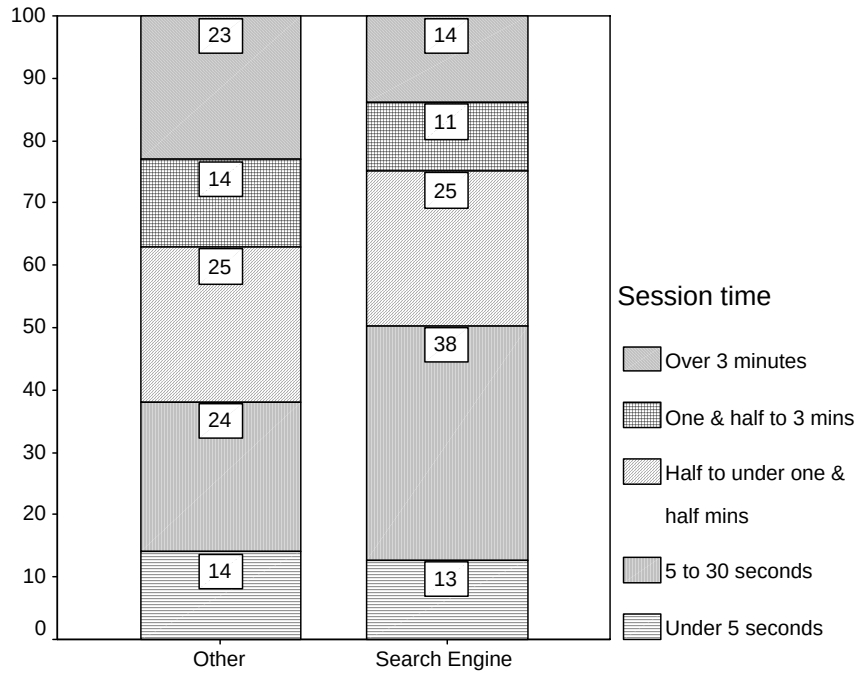


Figure 21 gives the percentage distribution of session time, grouped by whether the user had entered the site from a search engine. Those users not using a search engine were far more likely to have longer sessions. Twenty-three percent had sessions that lasted 3 or more minutes compared to fourteen percent of users who had entered the site via a search engine.

Figure 21 the percentage distribution of session time (grouped) by DNS organisation type of user.



The following (Figure 22) classifies and compares how users navigated their way around Humbul. The key outcome variable here was if the user had accessed an ID (extended summary page), or if the user had clicked on a URI (resource link). The user could find this information using one of three methods: a) a search engine (such as Yahoo or Google); b) the on-site search facility; or c) the site menus - or a combination of these three. A 'menu-user' is defined here as a user who had viewed a sub (menu2) level menu at least once. This grouping of navigation accounted for about half of the sessions. However, about 12% of sessions had just visited the homepage and did not go on to view any subject menu or outcome views. A further 19% of all sessions were 'other' referrer users who also did not view any subject or outcome views.

By examining users' navigational path, we discovered that about two thirds (64%) navigated the site via a search engine; about 12% of sessions used the on-site search facility; 12% used menus; and 11% used some combination of the three method at least once in their session.

Figure 22: Distribution of navigation method

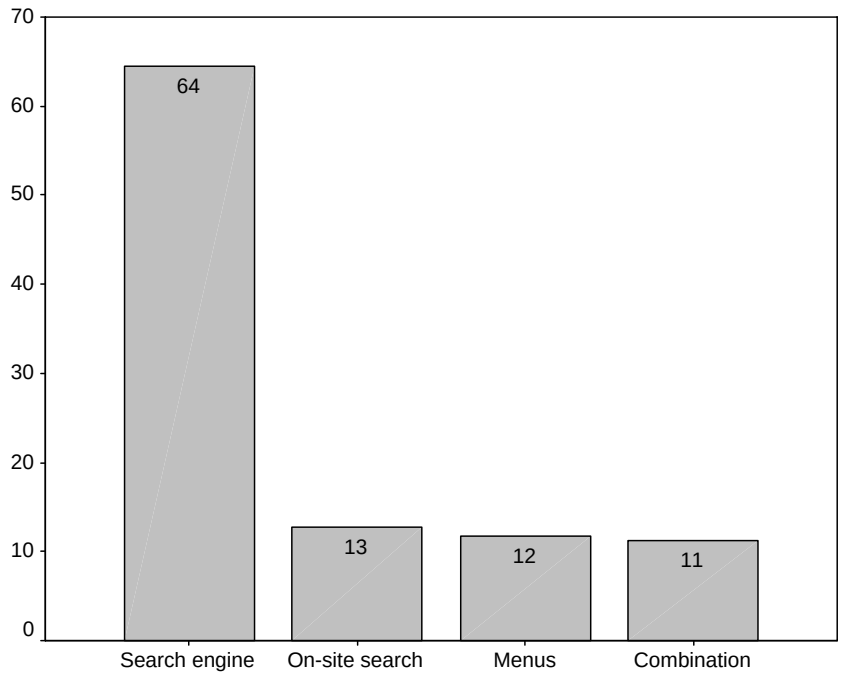


Figure 23 gives the distribution of navigation access by month. There appears to be a greater use of on-site searching and use of a combination of methods between September to December compared to other months.

Figure 23 The distribution of navigation access by month

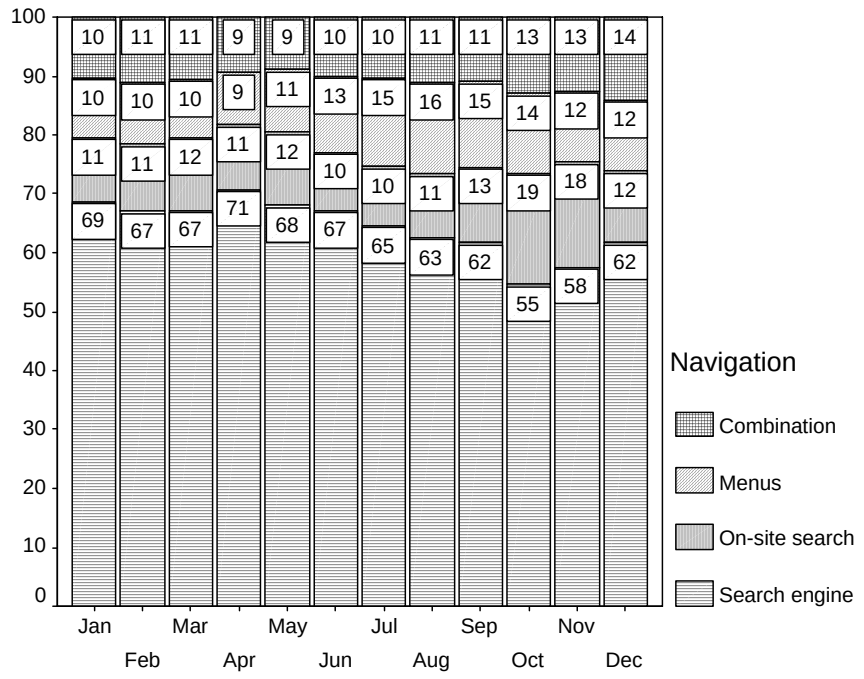


Figure 24 gives the distribution of navigation access by organisation type of user session. Academic users (c.38%) were least likely to navigate using a search engine as compared with 63% of commercial user sessions and just under three quarters (73%) of net type sessions. Academic users were much more likely to use the on-site search facility (29%), menus (18%) or a combination of methods (15%).

Figure 24 Percentage distribution of navigation access by organisation type of user session.

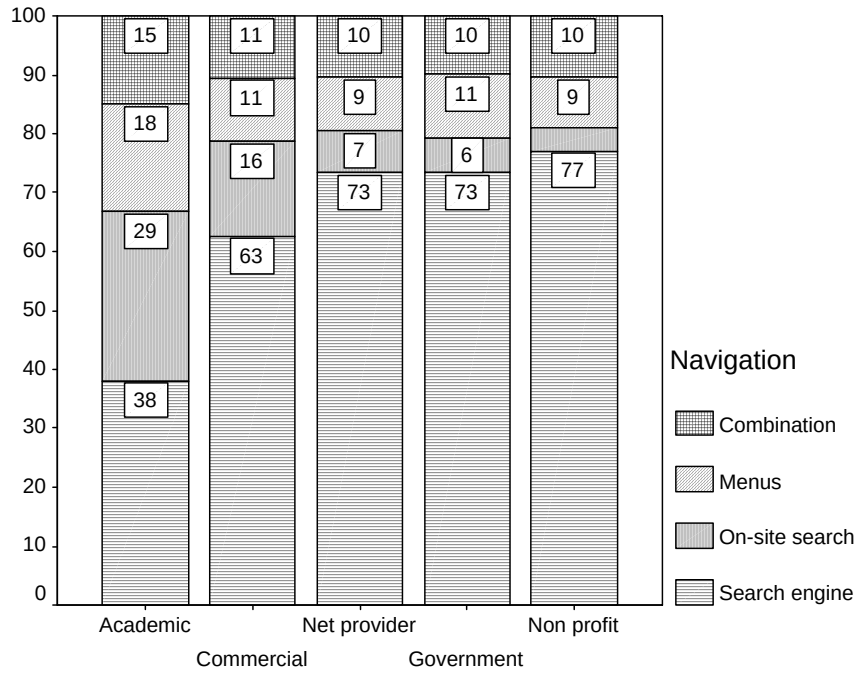
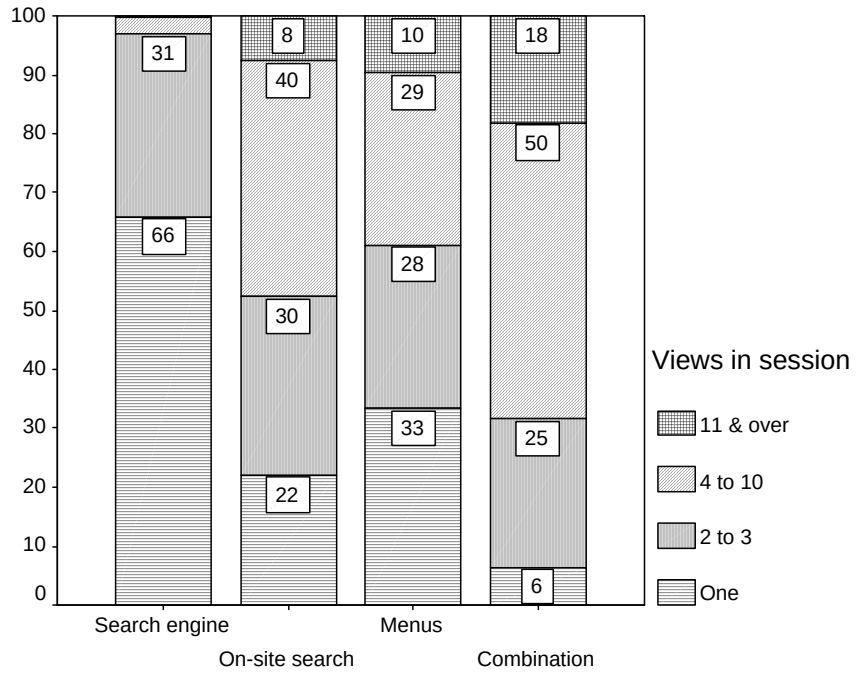


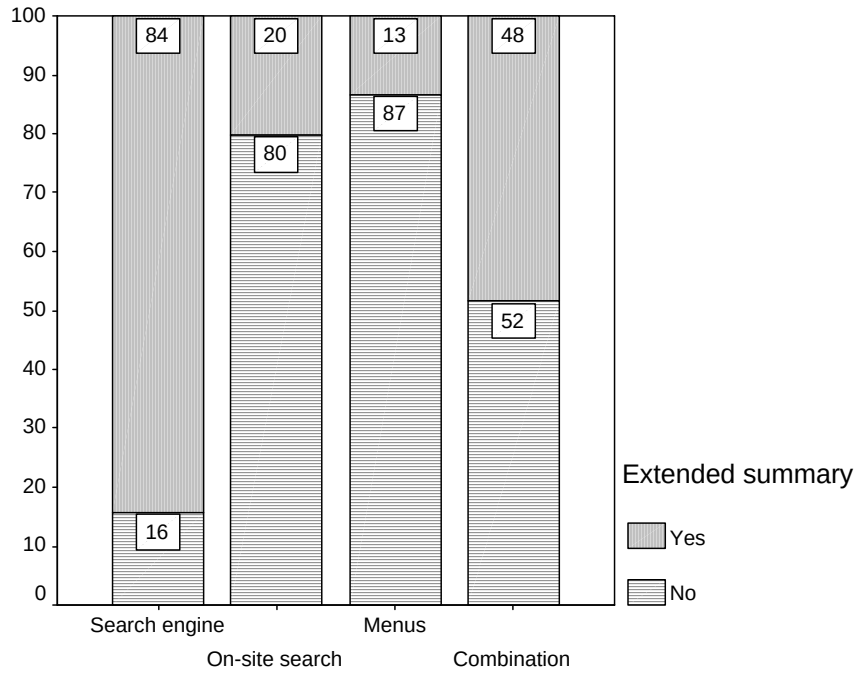
Figure 25 gives the percentage distribution of number of pages viewed in a session across navigation access. Those users coming into the site by a search engine were far more likely just to view one page and leave 66% did so compared to 22% of sessions where the on-site search facility was used or 33% of menu users.

Figure 25 The percentage distribution of number of pages viewed in a session across navigation access.



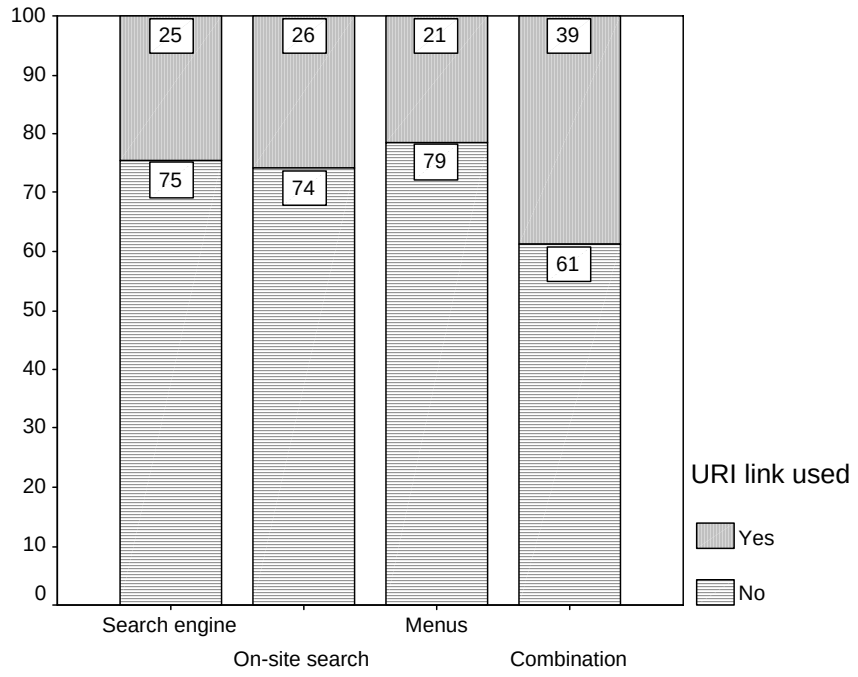
In terms of accessing the item ID extended summary, 84% of those who used a search engine to access the site, also accessed an extended summary. About 20% of those using the on-site search facility accessed an extended summary, 13% of menu-users did so and 48% of those using a combination of access methods accessed such items.

Figure 26 The percentage distribution of ID extended summary items viewed and cross navigation access (sessions)



In terms of accessing a URI link, about a quarter (25%) of those using a search engine went on to link to an external resource, a quarter (26%) of those using the on-site search facility did so. About 1 in 5 of users navigating via menus went on to link to a resource and about 39% of those using a combination of methods did so.

Figure 27 The percentage distribution of URI link resources used by navigation access. (sessions)



The following gives the distribution of subjects (first subject viewed) viewed (menu users only) in a session. History (25.6%) attracted the most use, followed by English (16.8%), Religion (6.6%), Humanities_a (6.3%) and Philosophy (5.6%).

Figure 28: Distribution of first subject viewed (sessions)

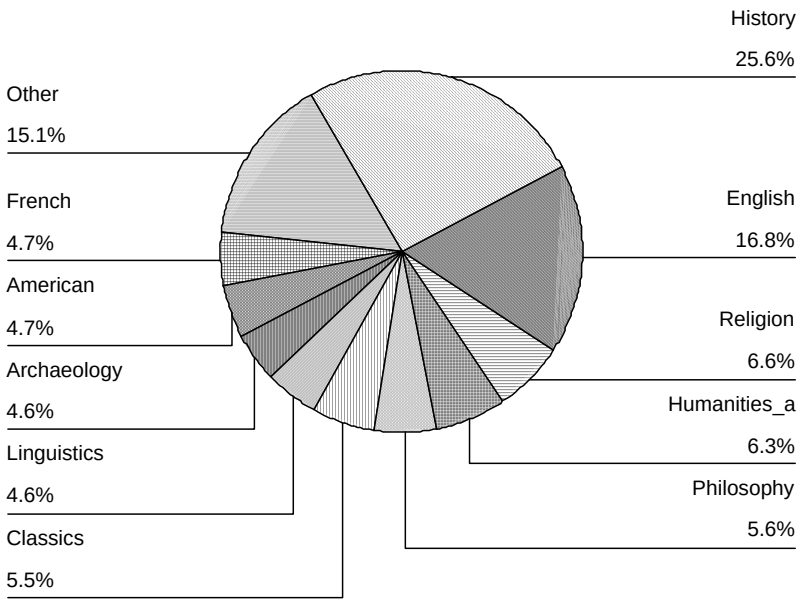


Figure 29 gives the number of items viewed in a session across subject. Those viewing Humanities_a tend to view more items in a session compared to other subjects.

Figure 29 The number of items viewed in a session across subject.

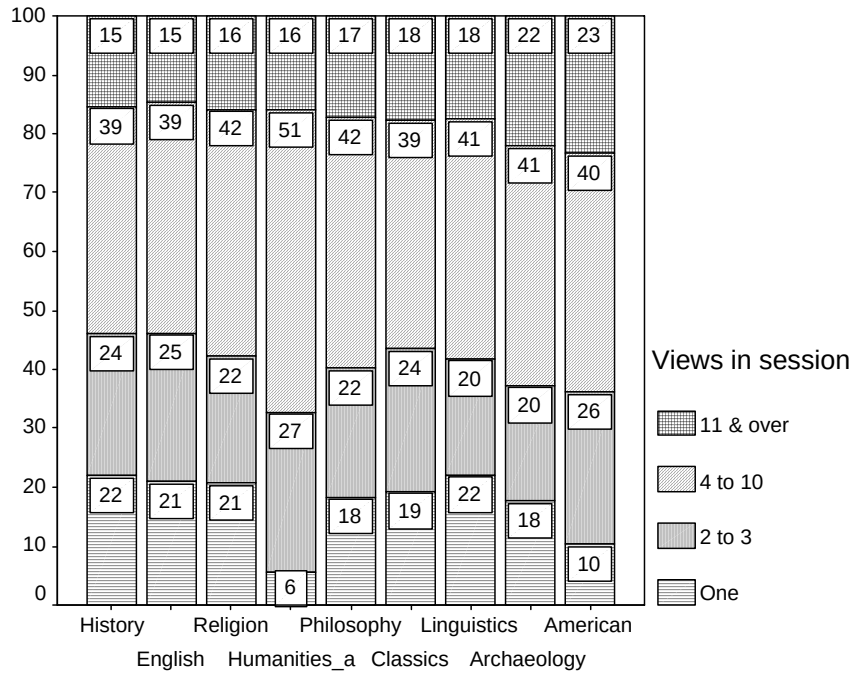
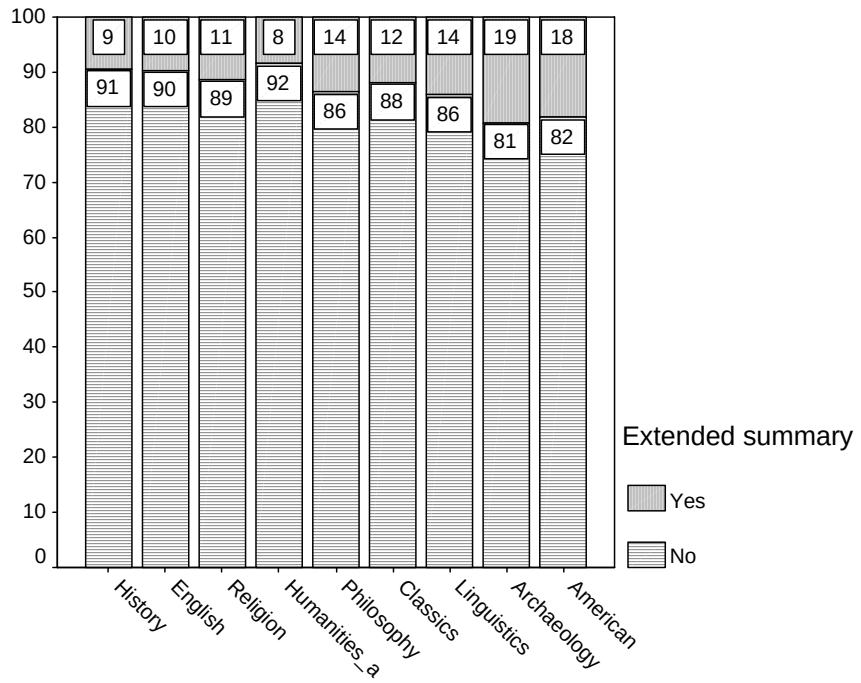


Figure 30 gives the number of extended items viewed across subject. Archaeology (19%) and American (18%) studies attract a greater percentage views to ID extended summary pages and Humanities_a (8%) the least.

Figure 30 The distribution of extended items viewed across subject. .



The following table rank lists the first word of Yahoo search expressions used. Common terms such as “the” were excluded. History seems a popular search word to include and 5% of search expressions included history as the first term in a search expression.

Figure 31: Yahoo search words – first search word used only

Term	count	Percentage of all words
History	5803	4.4
world	1037	.8
History	994	.8
English	775	.6
ancient	724	.5
philosophy	622	.5
roman	558	.4
victorian	557	.4
women	544	.4
find	539	.4
john	530	.4
British	525	.4
journal	515	.4
online	505	.4
Russian	461	.3
spark	461	.3
medieval	455	.3
pictures	446	.3
American	427	.3
language	416	.3
		12.8

A5.3.3 Examples of User-Behaviour

It would require a much more extensive analysis than was warranted by the limited user-information at our disposal to undertake an examination of individual user-behaviour. Here, we simply look at the three transactions by way of example.
The first (user-visit 1) is of an IP number that was recorded as visiting the site twice, once on 7 March 2005; then again on 10 October 2005. In March, this user viewed 5 pages. Via current procedures, we were not able to identify this user's domain name server (DNS) details. The user accessed the Humbul site via an organisation called www.netaddress.com and accessed an ID page. One minute and nine seconds later, the user looked at the Slavonic 'sub-menu' page. This is a menu-page listing resources in a reduced summary, providing a link to an extended summary, and then linking to the external resource. Sixteen seconds later, the user completed an internal search using the words 'poet poem poetry'. Nine seconds later, the user returned to the 'sub menu' Slavonic page. We infer that the user was looking for material relating to poetry in Slavonic languages, or relating to Slavonic subjects. We must also infer that the search did not initially provide anything of interest. Twenty-nine seconds later, the user revisited the 'id' page, once more visiting it via www.netaddress.com suggesting that the user had left the Humbul site and then revisited it once more. In all, this user-session lasted about two minutes. The user returned to the site about 7 months later on 10 November 2006. This time, the user just viewed one ID extended summary-item which they had found on the site via Yahoo and used the search terms 'women's work out' to find the document. The user just viewed one document and left.

Figure (DNS

5.32: user-visit (1)	unknown)			
07-Mar-05	14:49:52	id	6365	www.netaddress.com/
07-Mar-05	14:51:01	sub	Slavonic Poet poem	www.Humbul.ac.uk/
07-Mar-05	14:51:17	search	poetry	www.Humbul.ac.uk/
07-Mar-05	14:51:26	sub	Slavonic	www.Humbul.ac.uk/
07-Mar-05	14:51:55	id	6365	www.netaddress.com/
10-Oct-05	21:56:26	id	10759	mx.search.yahoo.com/

The example graphically illustrates the frustrating inconsequentiality of web-log analysis without supporting user information. We might tentatively infer that this user did not find anything significant from their first search; but that the site had achieved some 'recognition' for them to revisit it during a later search. In neither case, can the resource-discovery 'experience' be described as very 'rich'.

The second example (user-visit 2) is of a user who accessed the site on 2 May 2006. The user landed at an ID extended summary page, having found the site using Google. They had used the search terms 'allison pompeian households'. A second afterwards, the server delivered a page labelled '404.html'. A 404 code-page is normally one that informs the client that the page or item had not been found. Traditionally, web-item counting software identifies the 404 coded items in the status field of the logs and deletes these from the count. In our analysis, however, since the 404 item is delivered as an html coded page the analysis will count two items as viewed, even though (in terms of resource discovery), the site visit had yielded nothing by way of information.

Figure 33: user-visit (2)				
02-May-05	17:24:22	id	13518	www.google.com/
02-May-05	17:24:23			

In a final example, a user visited three times, each time using Yahoo to do so. On 17 March 2005, the user found an extended summary document via the search expression 'voices from gaps women writers of color'. The user did not view any other pages. The user returned to the site on 22 March 2006, using the search expression 'nikki giovanni biography timeline'. The user left the site, but then returned to the site 45 seconds later using the same search expression in Yahoo. Five seconds later, the user then clicked on the external resource link and left the Humbul site and visited 'nikki-giovanni.com'. The user then returned about three weeks later and visited the site twice, on both occasions using Yahoo. The first time, s/he used the search expression 'voices from the gaps' and on the second time, 49 seconds later, using the search terms 'voices from the gaps women writers of a color'. On both occasions the user was delivered an ID extended summary page. This user seemed to prefer to use Yahoo rather than the on-site search facility or menus. Perhaps the idea of getting to grips with a menu structure was more daunting than using Yahoo. At all events, we should probably classify this as a user that had 'found' a resource through the subject-portal on this occasion:

Figure 34
user-visit (3)

17-Mar-05	20:32:36	id	4405	search.yahoo.com/
22-Mar-05	18:00:04	id	9676	search.yahoo.com/
22-Mar-05	18:00:49	id	9676	search.yahoo.com/
22-Mar-05	18:00:54	URI	nikki-giovanni.com	www.Humbul.ac.uk/
15-Apr-05	14:57:45	id	4405	search.yahoo.com/
15-Apr-05	14:58:36	id	4405	search.yahoo.com/

A5.3.4 Humbul Web-Log Analysis: Conclusion

Site Usage.

The Humbul site saw, on average, about 6-8,000 items/pages viewed per weekday in an average of 2,500 sessions. About half the users were from the USA. Under a quarter of the sessions could be directly attributed to academic institutions. We may therefore presume that a minimum of c.550 sessions per day were from academics, with a further cohort of UK academics accessing the site via commercial servers – perhaps doubling that number of sessions. Of course, many of these academic visitors may have been for teaching purposes, or undergraduate visitors in search of materials for projects and dissertations. We should not disaggregate teaching and research too clinically in our resource-discovery analysis. Since site revisits appear to be at a low level (though CIBER was able to furnish us with no statistic upon revisits), we may cautiously infer that only a small proportion of the research cohort in the UK, identified as 50-60,000 in **A2** (above), used the Humbul service in 2005 – perhaps in the region of 1-10% with the likelihood that it is in the lower quartile of that range.

Discipline Distribution.

The dominance of History users of the site is even more pronounced than the statistics suggest. If we compare the subject distribution [figure 6] with the subject distribution by RAE2001 returns [**A2**], this is as striking as is the under-representation of Modern Languages and Linguistics, and (to a lesser extent) Philosophy, Law and Religious Studies (Law not being a subject represented in the RDN resource discovery networks).

Site Penetration

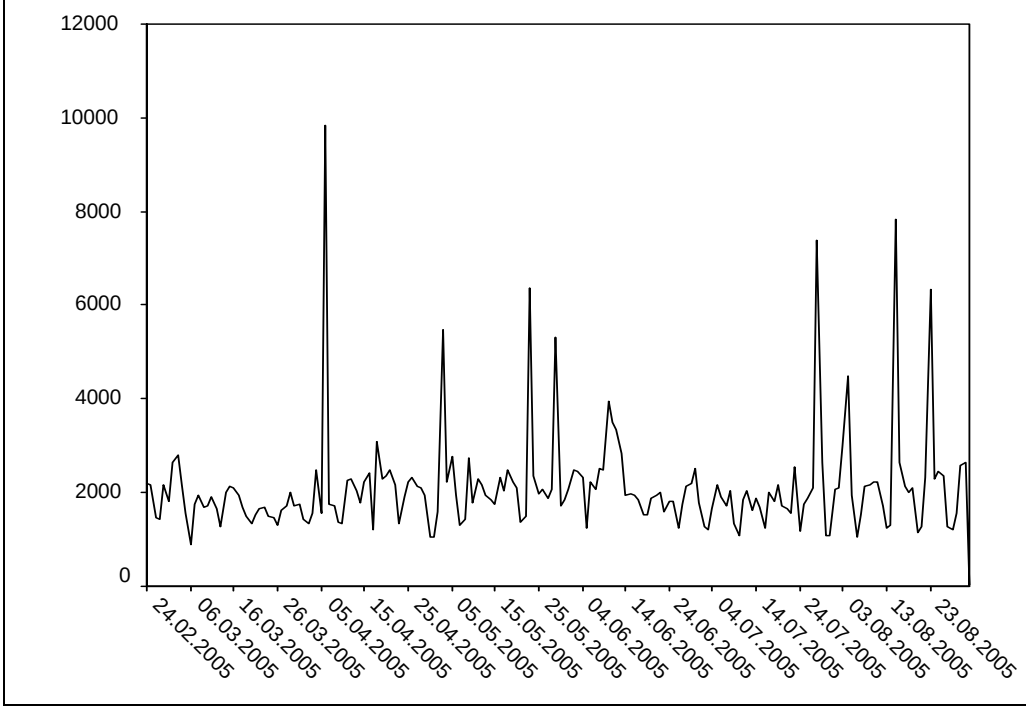
Academic users tended to make more serious use of the site when they visited it. They were least likely to 'bounce' out of the site having visited it. An encouraging statistic from the analysis is that 31% of the academic visitors spent over 3 minutes when they visited the site. Only a small proportion, however, used the on-site menus and search engines. The numbers of academic users who accessed an extended summery of a resource is also encouraging. But one of the most resonant conclusions of the analysis is that only a minority of these users went on then to link to an external resource.

A5.4 AHDS Web-Log Analysis

A5.4.1. Overall Site Usage

An overall view of site-usage is provided by the number of 'hits' or views per day. AHDS had approximately 1000 to 3000 views per day. Use is punctuated by occasional high volume day usage 'spikes' that can reach as high as 10,000 views.

Figure 35: Daily usage AHDS February to September 2005



Interestingly, a large percentage of 'hits' seem to have occurred in August – in contrast to the Humbul evidence above. We suspect that this intensity of usage is the result of the long vacation research traffic demands, coupled with MA and MPhil dissertation work. If this is the case we are struck by the fact that the comparable evidence fro Humbul is a trough at the equivalent period (Figure 8).

The way the AHDS site is constructed means many file-names and directories share the same name, irrespective of subject. If a record of subject-usage was to be found then this could only be done at the directory level. But the problem with using a directory-name approach is that there are a number of pages associated with a subject-directory. So, for example, 'History' has a number of menu-type pages and so a purely directory approach gives a biased view of activity, but perhaps provides an overview of subject popularity.

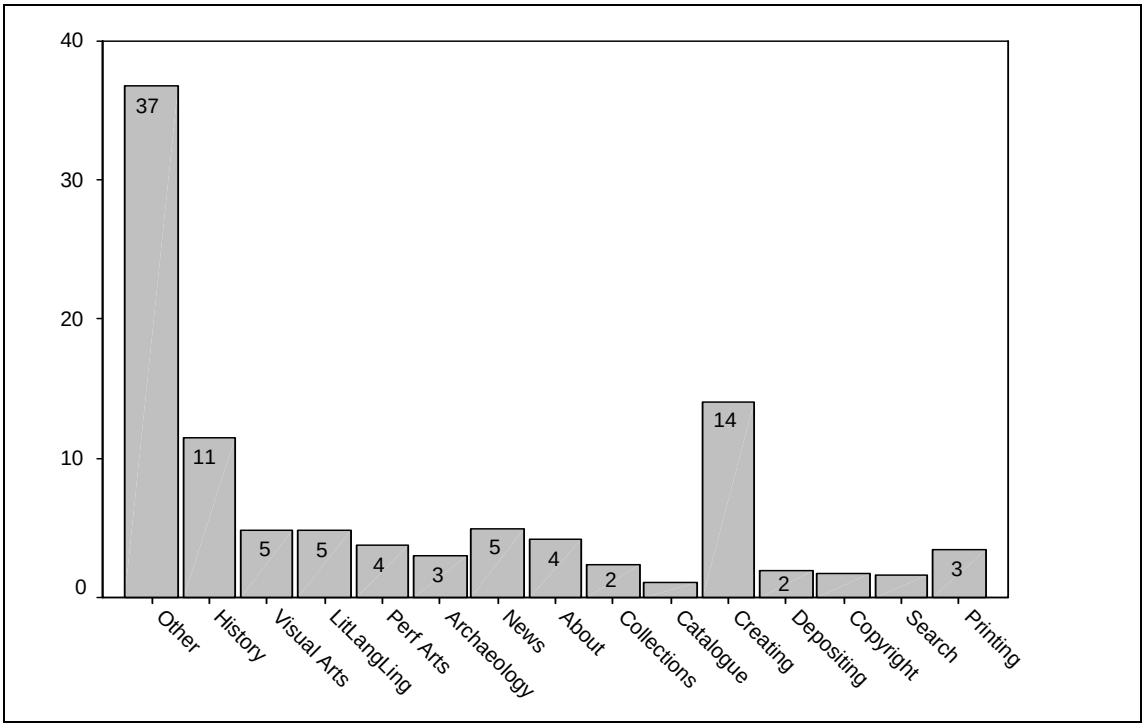
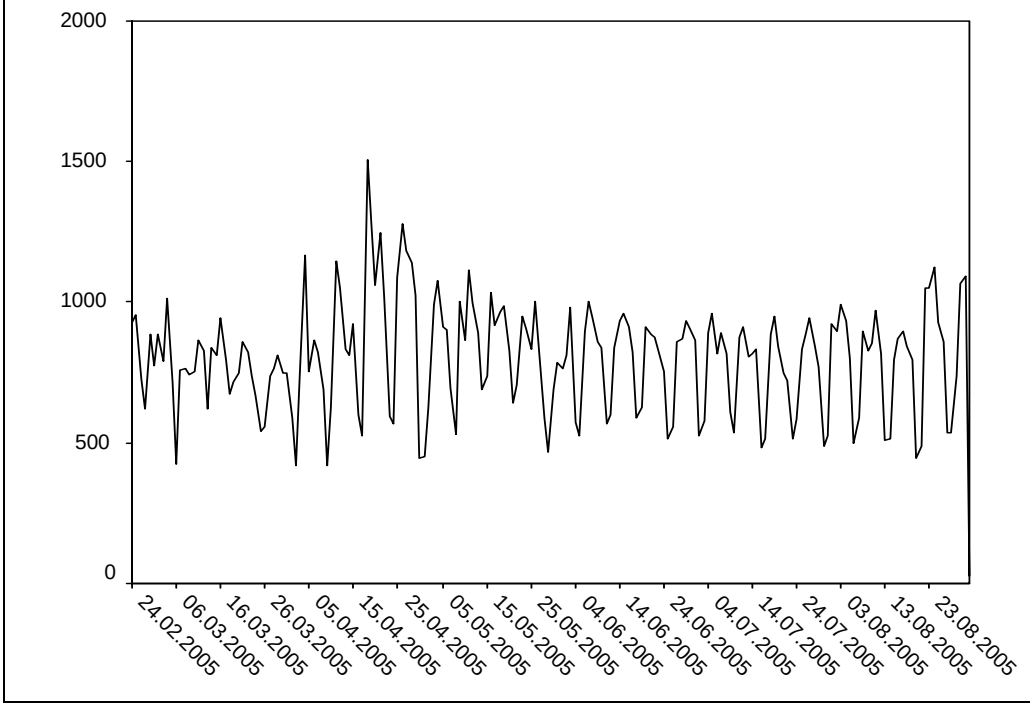


Figure 36: Top level directory usage over March to August for the five subjects

A5.4.2 User Session Analysis

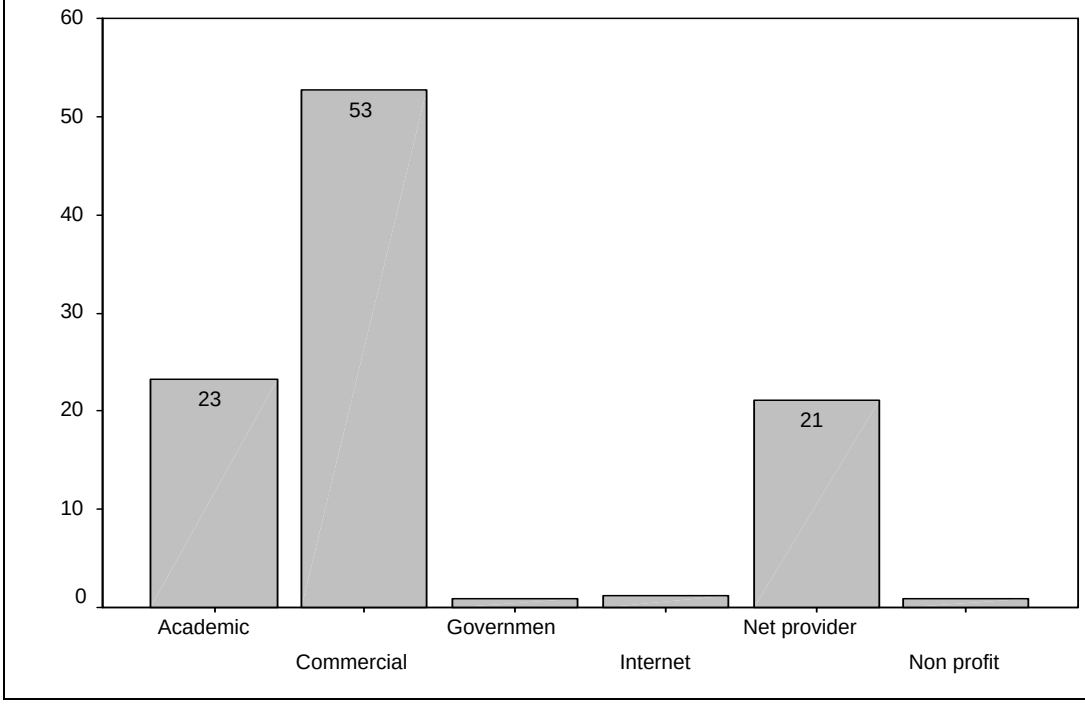
The number of user sessions was 151,998 over February to August - 600 to 900 user sessions a day. This provides the numbers (not 'hits') using the site.

Figure 37: Daily number of user sessions



The DNS Analysis of the AHDS site is as follows, with access from 'commercial' DNS dwarfing all other organizational entities:

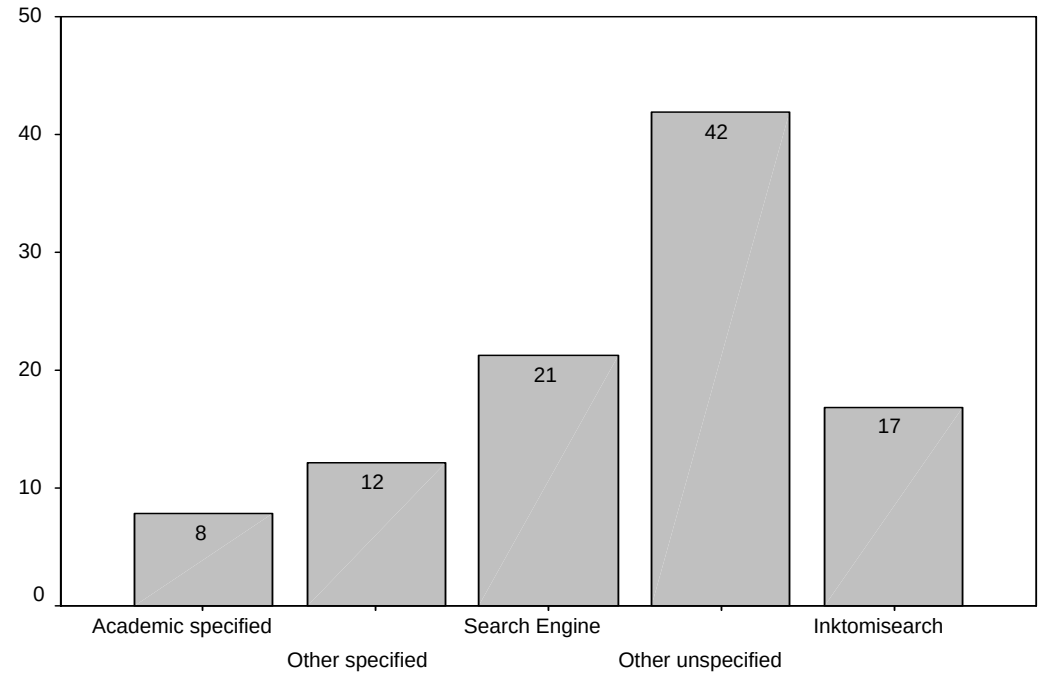
Figure 38: Distribution of sessions by DNS organisational usage



In the equivalent referrer-analysis to that conducted on the Humbul evidence, 21% of the users came to the AHDS via a search engine. There is, however, a high 'unknown user' element to this evidence and whether this should be accounted also as users who also arrived

via a search-engine is unclear. It is possible that the true percentage of search-engine derived visitors is 40% or more.

Figure 39: Distribution of sessions over referrer link.



The following table lists the top 15 referrer sites in the group ‘other unspecified’. The top 15 accounted for 51% of other unspecified sessions.

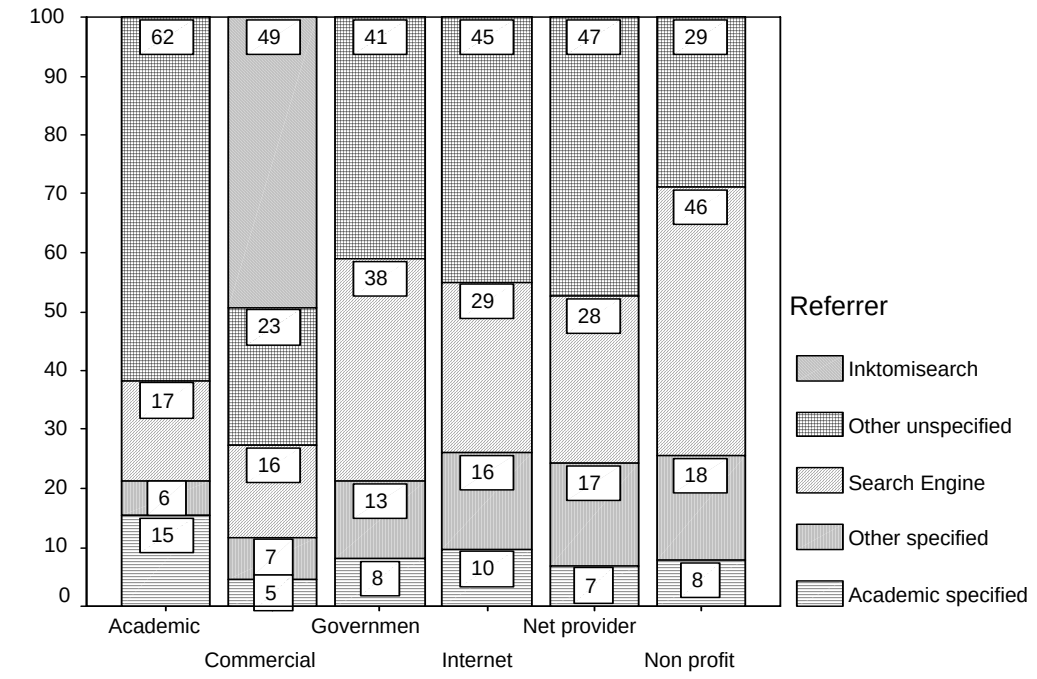
Figure 40: Top 15 referrers in "Other unspecified"
www.stumbleupon.com
www.onebird.com
www.ivritype.com
dublincore.org
www.uky.edu
aolsearch.aol.co.uk
www.cmswatch.com
www.tei-c.org
www.library.cornell.edu
www.ifla.org
www.minervaeurope.org
www.dlib.org
www.drh.org.uk
aolsearch.aol.com
www.nla.gov.au

The following table lists the top 15 referrer links in the group 'Academic specified'. The top 15 account for 60% of sessions. We have singled out in green the referrals that may have been significantly affected by internal traffic within the AHDS site. We have also singled out in red the referrals from Humbul and Artifact.

Figure 41: Top 15 referrers in "academic specified"
hds.essex.ac.uk
www.hw.ac.uk
www.data-archive.ac.uk
www.kcl.ac.uk
appserver.pads.arts.gla.ac.uk
census.data-archive.ac.uk
edina.ac.uk
www.Humbul.ac.uk
www.Artifact.ac.uk
www.lib.cam.ac.uk
www.jisc.ac.uk
www.esds.ac.uk
www.bodley.ox.ac.uk
www.ukoln.ac.uk
www.abdn.ac.uk

The next figure gives the distribution of referrer link by DNS organisation of user. In terms of academics, 15% entered the site via academic-specified links, 17% via a search engine and 62% via other unspecified.

Figure 42: The percentage distribution of referrer link by DNS organisation of user

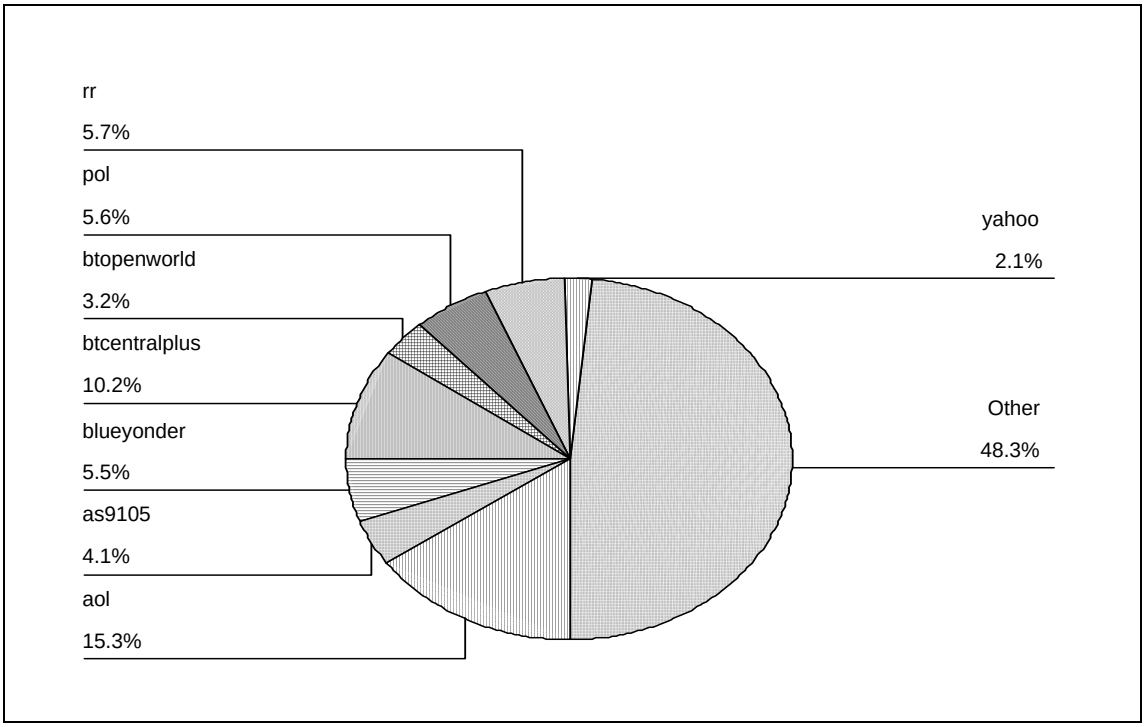


The following lists the DNS name of the referrer group 'Other unspecified' (i.e. the area marked 62% in the Academic bar in the above chart). The top 15 organisations accounted for 36% of the total were as follows. Once more, those likely to have been influenced by internal AHDS traffic are highlighted in green:

Figure 43: Top 15 academic institutions identified as referrer ('other unspecified')
pc094-016.odds.kcl.ac.uk
pc094-015.odds.kcl.ac.uk
pc094-026.odds.kcl.ac.uk
linux01.lib.cam.ac.uk
pc094-017.odds.kcl.ac.uk
pc094-010.odds.kcl.ac.uk
morse.ucs.ed.ac.uk
bottle.gla.ac.uk
pc094-030.odds.kcl.ac.uk
dozer.infodiv.unimelb.edu.au
farnham.surrart.ac.uk
pc168-21.UB.UU.SE
atticus.ahds.ac.uk
xena.lib.unimelb.edu.au
dahds7.essex.ac.uk

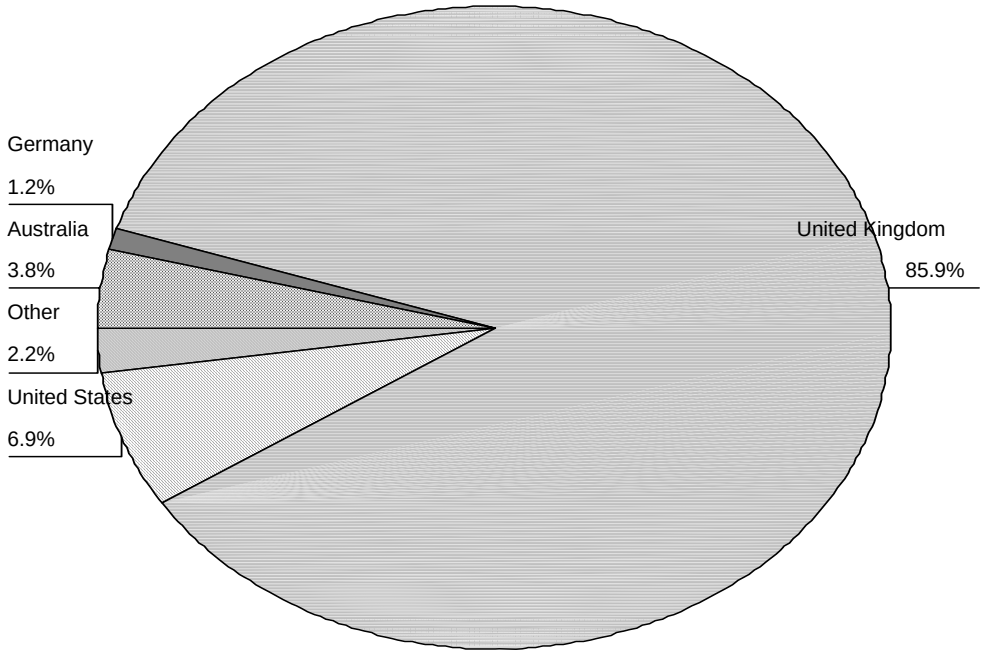
These 'unspecified' users and specified 'commercial users' are most likely to enter the site using a search engine:

Figure 44: Commercial referrer group (other unspecified & unknown)



The following gives the country location of academic user sessions only. Academic institutes are said to be less likely to mis-register their location. As can be seen most academic sessions (86%) comes from the UK.

Figure 45: DNS country distribution of user sessions – academic institutions only



Other: Countries accounting for less than 1% individually

When referrer information is broken down into the 7 highest user institutions it shows that Cambridge (23%), Essex (24%) and Oxford (22%) have a relatively high percentage of users coming in via their university servers. Once again, however, the statistics for Essex may be influenced by internal AHDS traffic.

A5.4.3 Site Penetration

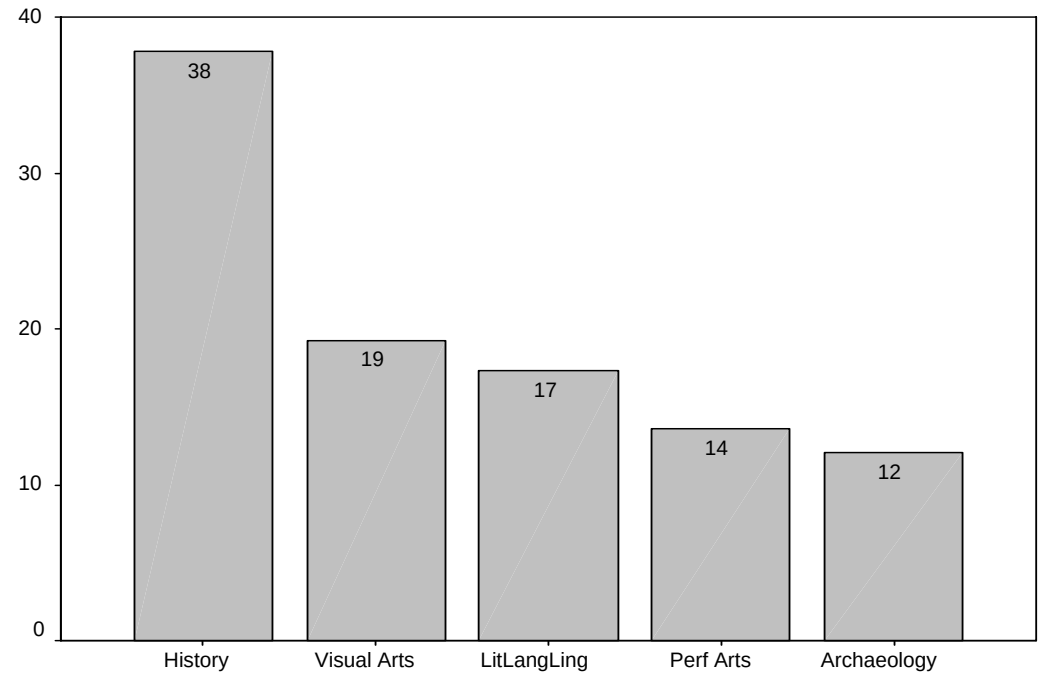
How many pages on the AHDS site were viewed in a 'session'? The percentage distribution of number of pages viewed in a session show few users viewed more than 2 pages in a session. Three quarters (72%) viewed one page. Viewing one page and then exiting is described by CIBER as 'bouncing'. It delineates the user-scenario in which a search engine facilitates information-gathering for a user, who views a site, realises that it is not for them, and leaves. This behaviour is more apparent with search engine/directory listings, where there is little cost in cycling through the first 10 - 20 hits. In terms of the overall referrer-group 78% of search-engine users left after viewing one page. Commercial users were most likely to view one page in a session (71%) whilst academic users were least likely to do so (58%). Those users who viewed more pages were more likely to conduct an internal site-search. Half (56%) of those sessions viewing 11 or more pages did so. 28% of all sessions viewed more than one page, 19% viewed 2-3 pages; 6% 4-10; and 1%, 11 or more.

A5.4.4 Subject Analysis

The AHDS five subject-areas were used as the unit of analysis here: History, Visual Arts, Performing Arts, Literature, Language and Linguistics, and Archaeology. Most user sessions (75%) did not view a subject page. Of those that did, most (23%) just viewed one subject.

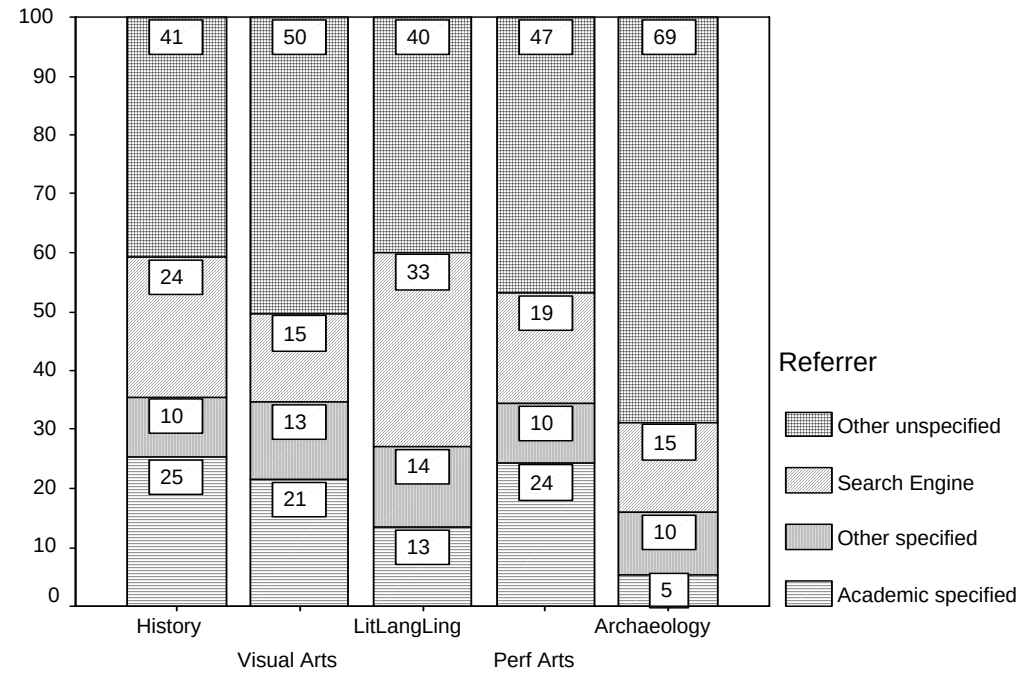
In terms of subject, History was the most popular: 38% viewed one page from History; 19% viewed a Visual Arts page; 18% viewed a Literature Language and Linguistics page; 14% Performing Arts; and 12% Archaeology. But the existence of the independent website access for each of the services renders this analysis very tentative – many users of (for example) the Archaeology site will have accessed it directly, and not through the AHDS central server.

Figure 46: Frequency distribution over first subject viewed



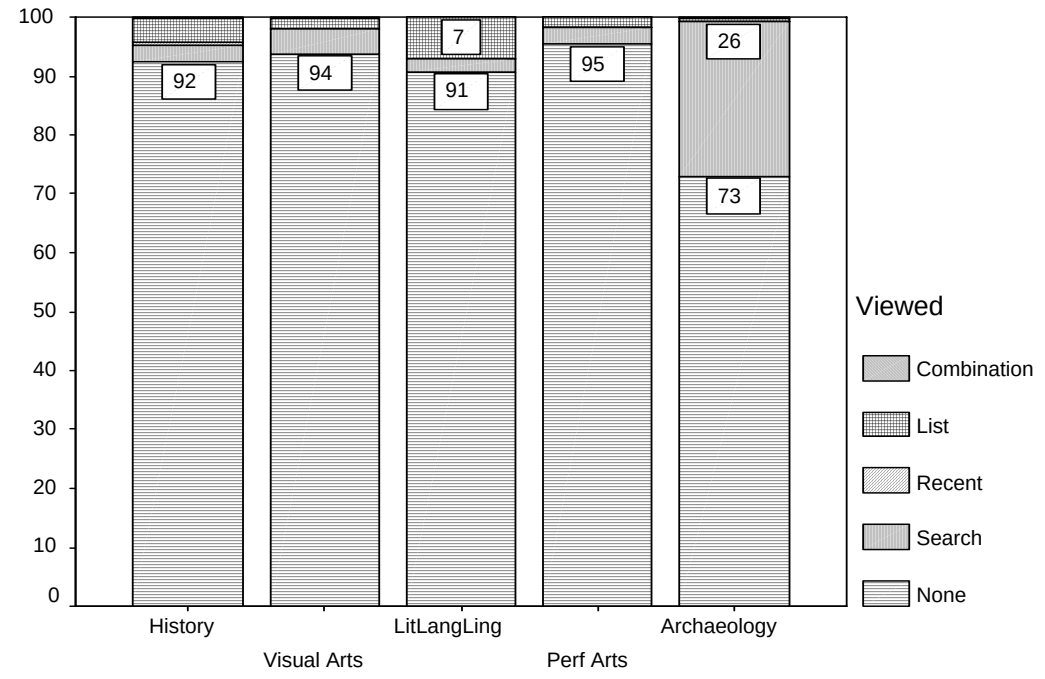
There is a greater use of search-engines by users viewing Literature, Language and Linguistics (33%), as compared with users viewing Visual Arts (15%) and Archaeology (15%).

Figure 47: Percentage distribution share of referrer link by first subject viewed.



When in a subject-grouping, users could either search the database view, a listing or visit a page listing recent items. The following gives the percentage share of these activities, broken down by subject. It appears that few users go on to search the database. Less than 5% do that in each subject, save for Archaeology, where approximately 33% of users go on to do a search. Of course, it is impossible to establish by this kind of analysis the extent to which site-design is a factor in the behaviour of users at this point.

Figure 48: Percentage share of type of view by subject-grouping .



The following gives an idea of the other non-search pages users were looking at. Without a more detailed analysis of the pages in question (difficult to provide because of the diversity of the AHDS site), this is of limited utility since the most frequent category is ‘other’.

Figure 49: History – frequency of pages viewed.

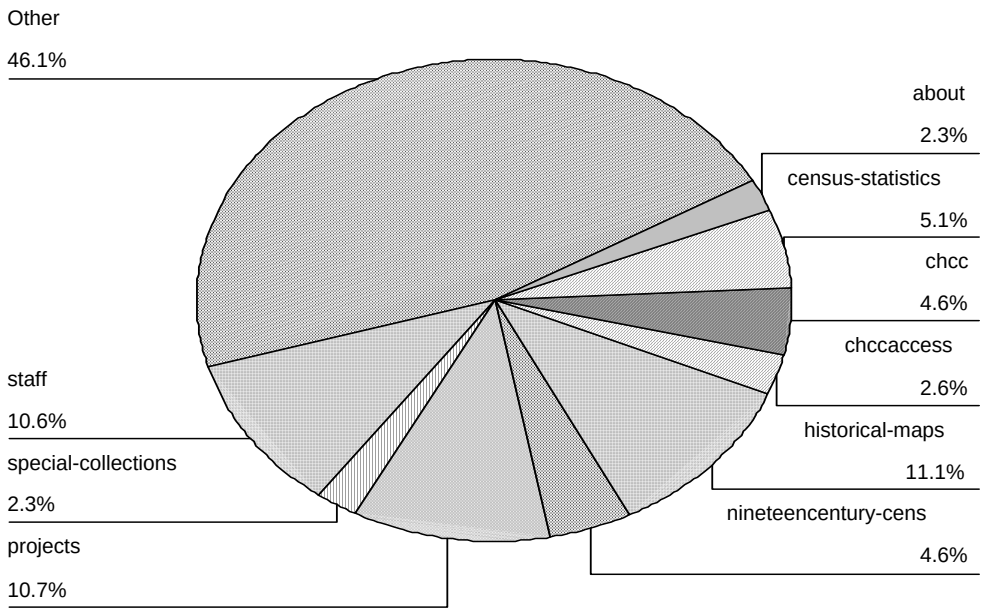


Figure 50: Visual arts – frequency of pages viewed.

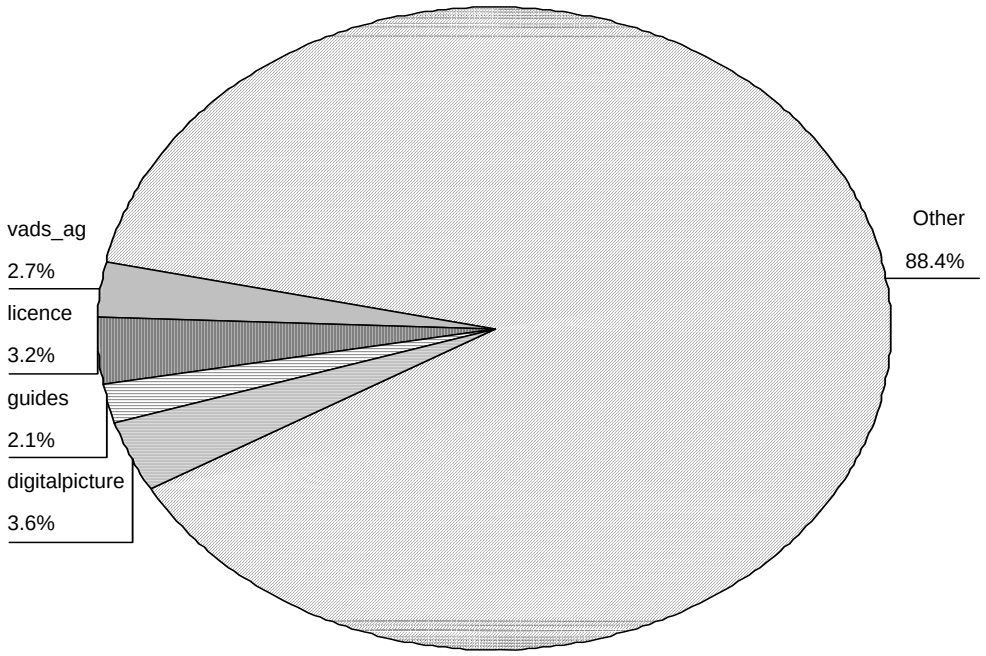


Figure 51: Literature, Language Linguistics – frequency of pages viewed.

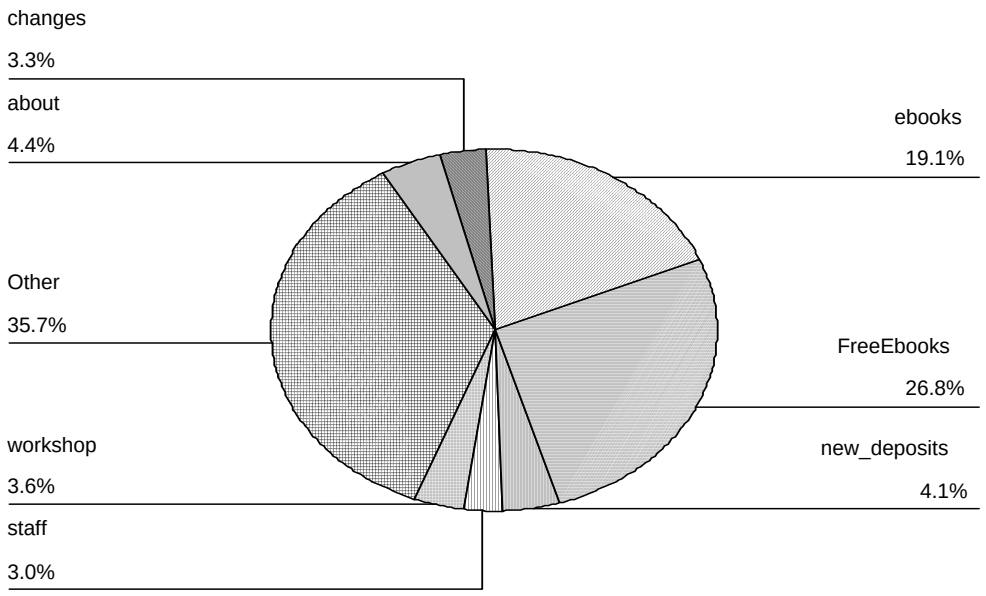


Figure 52: Performing arts – frequency of pages viewed

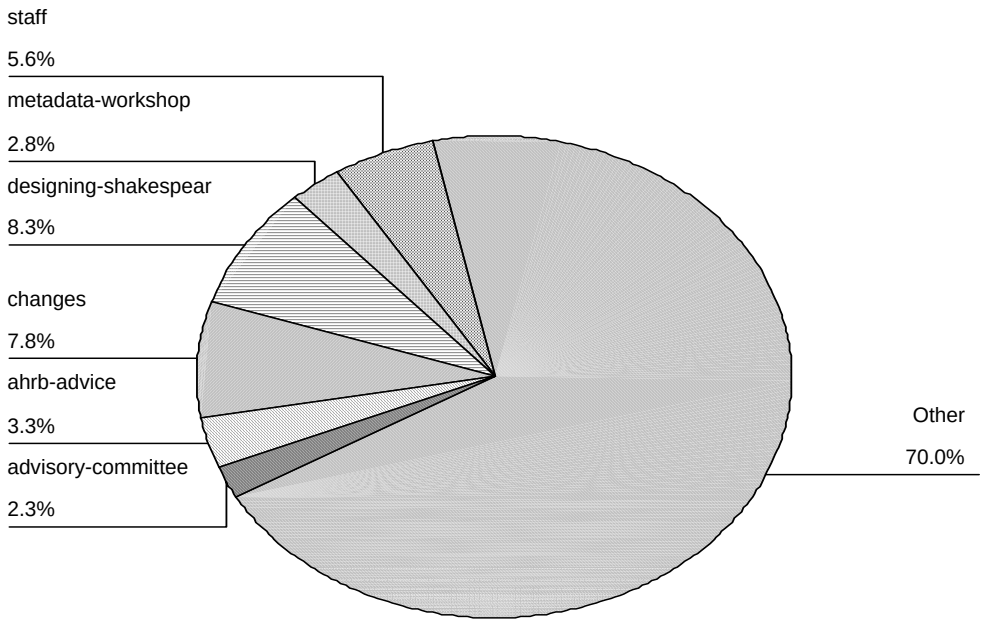
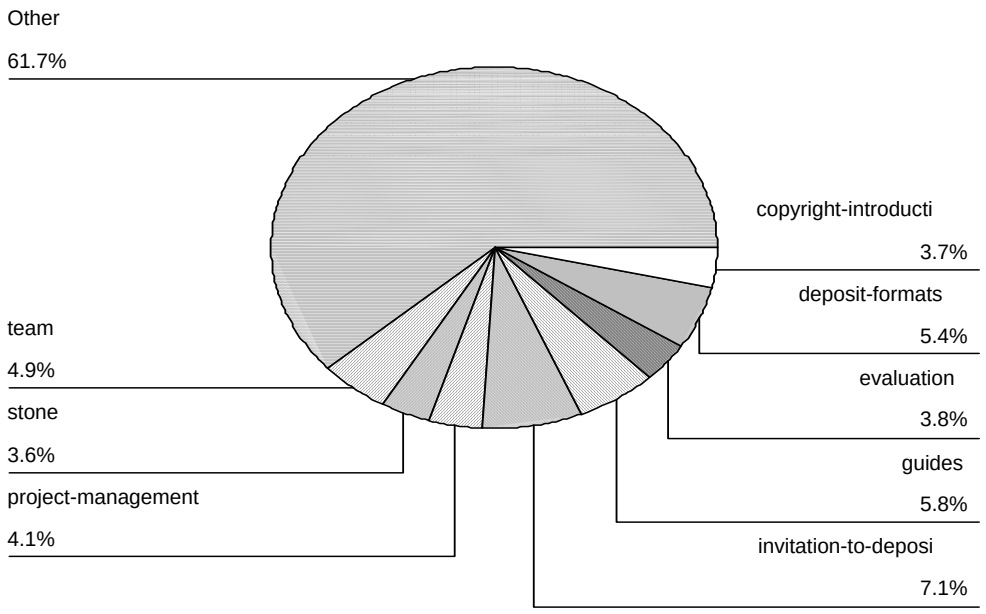


Figure 53: Archaeology – frequency of pages viewed



A5.4.5 AHDS Web-Log Analysis Conclusions

Site Usage

The AHDS site averaged 1-3,000 hits per weekday with an average of 600-900 sessions per day. These figures, however, are not an adequate measure of overall site visits since so many users accessed the AHDS at its individual sites rather than through its central server. The impact of commercial DNS access to the AHDS site is even more pronounced than for Humbul; but it is more accessed by UK users than Humbul.

Site Penetration

Academic users, as with Humbul, tended to make more 'serious' use of the site than those identified as coming from non-academic origins. Of the academic users, half visited 11 or more pages/views on the site during a visit. This suggests a satisfying depth of penetration to the site's resources.

Subject Distribution

In comparison with our overall Arts and Humanities research profile (**A2**), History and Archaeology significantly out-perform their cohort size. Visual Arts and Performing Arts perform in accordance with their profile. Languages, Literature and Linguistics under-perform in accordance with their profile.

A5.5 Artifact Web-Log Analysis

Despite repeated requests to do so, CIBER has not submitted its analysis of this material that has been supplied to them from this service provider. We shall submit an addendum to this report if it arrives.

A5.6 Individual AHDS Service-Provider Web-Log Analysis

Despite repeated requests to do so, CIBER has not submitted its analysis of this material that has been supplied to them from this service provider. We shall submit an addendum to this report if it arrives.

A5.7 Overall Conclusions

This analysis of the available web-log statistics provides a good deal of circumstantial detail about the traffic patterns of the service providers. But the conclusions that we can draw from it are disappointing modest and frustratingly inconsequential.

Although we have some indicative measures of **overall usage**, we cannot satisfactorily isolate academic and non-academic usage in the data. What is more, it cannot be reliably used in a comparative context. Many users accessed the AHDS individual sites rather than going through the central server. It is likely that the AHDS overall statistics significantly under-record its overall usage, whilst some traffic it records may well be internal to the service itself. The high volume of Oxford referrals in the Humbul statistics may also relate to traffic internal to its service. Any direct comparison of the impact of the AHDS as compared with the RDN subject centres is impossible on the basis of our evidence. The AHDS overall site-visits and session statistics may well have been higher than Humbul's if we include individual site traffic. Equally, the RDN subject centre traffics may be higher than that of the AHDS if we include the unknown usage statistics of Artifact. We have no way of knowing, such is the measure of our uncertainty about the reliability of the data to hand.

The indications of **subject distribution** provide some limited, but useful conclusions about the discipline-specific patterns of access to these services. In both instances, the most active users were from History – both in absolute terms and in comparison with their research cohort. In both instances, Languages, Literature and Linguistics were the least active users – in comparison with their research cohort. Philosophy, Law and Theology also appeared to be relatively inactive users. We hesitate, however, to draw more specific conclusions about other subject domains from the rather fragmentary analysis furnished by the data.

The indications of **site penetration** reveal, unsurprisingly, that academic users tended to be more serious 'users' of the sites, both in terms of the numbers of pages/views visited and in terms of the amount of time spent upon the site. The statistics for those users that went on to link to an external resource directly from the Humbul site should perhaps be a cause for concern. On the other hand, we should probably place the statistics of AHDS site-usage alongside the downloads of collections from its sites [**A3, above**]. Taken together, this

suggests that Arts and Humanities users in 2005 were finding more materials, and doing more with them, than they had done previously.

Anyone expecting to arrive at a picture of user-behaviour from web-log analysis is likely to be disappointed. It is a blunt instrument for analysis without complementary, detailed evidence of the user demographics in question. We therefore regard this evidence as being best adduced as part of a 'triangulation' approach, using it to confirm, strengthen or nuance, the conclusions we arrive at through our online questionnaire, focus groups, interviews, and Delphi analysis.