If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data

Claire Warwick, Melissa Terras, Paul Huntington and Nikoleta Pappa
School of Library, Archive and Information Studies, University College London
c.warwick@ucl.ac.uk

Abstract

There are now many online, digital resources in the humanities, and their creation is funded by various governmental, academic, and philanthropic sources. What happens to these resources after completion is very poorly understood. No systematic survey of digital resource usage in the humanities has ever been undertaken – and the factors for use and non-use of digital resources are unknown. The LAIRAH (Log Analysis of Internet Resources in the Arts and Humanities)[1] Project is a 15 month long study into the factors which determine long term use and neglect of digital resources in the arts and humanities. Using quantitative Deep Log Analysis techniques to understand real time user behaviour, and qualitative user workshops to gain an understanding of user approaches to digital resources in the arts and humanities, the study identifies factors which may predispose a digital resource to become used or neglected in the long term. This paper provides an overview of the techniques used in the LAIRAH project, and presents some preliminary results which may be of use to both the creators of digital resources in the humanities, and the funders of these projects, to ensure that significant intellectual effort and time, and financial resources, are not wasted in the creation of projects which are then neglected by the user community.

1. Introduction

Digital Humanities may be a relatively young discipline but it is a very productive one. In its short history scholars have produced thousands of digital resources which

have been funded by governments, philanthropic bodies and universities. In the UK alone, over 250 digital humanities projects have been funded by the Arts and Humanities Research Council (AHRC)[2] since 1998. Yet what happens to such resources after completion is very poorly understood. (Warwick, 1999b) Anecdotal evidence suggests that some projects become well known but others have been relatively quickly forgotten. This must be regrettable since the non-use of a resource represents a waste not only of the considerable intellectual effort and time expended in its production, but potentially considerable amounts of funding. However, no systematic survey of digital resource usage in the humanities has been undertaken, and the characteristics of a project that might predispose it for sustained use have never been studied. For example, does the presence in an academic department of the resource creator, or enthusiast, who promotes the use of digital resources, ensure continued use? Do projects in certain subject areas tend to be especially widely used? Are certain types of material, for example text or images, more popular? Is a project more likely to be used if it has communicated with the user community during its design phase?

This paper presents the early results of a study carried out by the LAIRAH (Log analysis of Internet Resources in the Arts and Humanities) project (http://www.ucl.ac.uk/slais/research/circah/lairah/) which is based at UCL's School of Library Archive and Information Studies. We are undertaking a fifteen month study to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. The LAIRAH project is funded by the AHRC's ICT strategy projects scheme, which reports to the AHRC's strategic review of all ICT related activity. Thus our work is highly significant, since its conclusions will inform the future of funding for digital humanities in the UK for the foreseeable future.

This paper describes the qualitative and quantitative methods used to analyse use of digital resources in the humanities during the LAIRAH project, and provides preliminary results. These give recommendations for both those about to create a digital resource in the humanities, and institutions which may fund those projects, to ensure that the resulting resource will be as widely used as possible.

1.1 Previous work in the area

Although useful recent work on humanities scholars has been done by Barrett, (2005), Talja and Maula (2003), Greene (2000) Herman (2001) and Ellis and Oldman, (2005) this tends to be on the areas of either information seeking practices or information needs of humanities scholars. Seminal work done by Stone (1982) and Watson Boone, (1994) showed that humanities users need a wide range of resources, in terms of their age and type. This is still true in a digital environment, where humanities users continue to need printed materials, or even manuscripts as well as electronic resources, which by their nature may imply a much greater age of materials than those used by scientists as the most recent survey of humanities user needs showed. (British Academy, 2005) However, only Bates (1996) has systematically analysed the kind of activities carried out by humanist scholars in digital environments, and this work was carried out on the Dialog system, which predated the web. Thus, while extremely valuable, it is now somewhat dated.

Bates' work and that of Dalton and Charnigo (2004) and Whitmire (2002) has shown that those humanities scholars who use digital resources tend to be demanding of the quality of resources and are capable of constructing complex search strategies, given appropriate training. We are not aware, however, of any literature that has used quantitative methods, particularly deep log analysis, to measure the levels of use of digital humanities resources. Our research presented below is also innovative in that it attempts to investigate not just which resources users need, or how they search them, but their detailed opinions about such resources, such as the qualities that they appreciate and factors that may inhibit use. We have therefore combined by qualitative and quantitative research methods, to provide as full a picture as possible of this complex area.

2. Methods

We chose to use Deep Log Analysis as a metric to assess use levels of digital resources in the arts and humanities. This is a technique that has been used

extensively by the UCL SLAIS' CIBER[3] research centre in other areas such as health information and commercial publishing, (for example Huntington, et al. 2002) but has never been applied to digital humanities - as far as we are aware. This analysis allowed us to identify patterns in usage of digital resources in the humanities, and allowed us to present users with a selection of used and non-used resources. By using quantitative techniques such as Deep Log Analysis, and qualitative workshop techniques, our knowledge of which resources were useful or ignored, and why they were used or neglected, was increased.

2.1 Deep Log Analysis

All digital information platforms have a facility by which logs are generated that provides an automatic, real-time record of use. They represent the digital information footprints of the users and by analysing them using deep log techniques it is possible to track their information seeking behaviour. When enhanced, logs can tell us about the kinds of people that use the services. The attraction of logs is that they provide abundant and fairly robust evidence of use. Logs record use by everyone who engages with the system - there is no need to take a sample, thus it is possible to monitor the behaviour of millions of people, around the world. They not only have an unparalleled size and reach, but are a direct and immediately available record of what people have done: not what they say they might, or would, do; not what they were prompted to say, not what they thought they did. The data are unfiltered and represent both the users' behaviour and complement important contextual data obtained by engaging with real users and exploring their experiences and concerns.

Server log data are records of actual web pages viewed. These records occur as a result of requests made by the clients' computer and provide a record of pages delivered from the web server to the clients' computer. The client however is not identified by name or demographics in this process. However, the server does record the Internet address of the clients' computer. These addresses follow an Internet Protocol (IP number) and relate to registered domain name server (DNS) information. The DNS information gives information such as organisation name, organisation type (i.e. academic or commercial) and country registration. Neither the DNS address

information nor the IP number records information that can be used to identify the actual user. To preserve anonymity further the logs that we analysed were purged of any personalisation data.

The following gives an example of the AHDS log file:

> 66.XXX.XXX.XX - - [24/Feb/2005:00:07:12 +0000] "GET /deposit/depintro.htm HTTP/1.1" 200 318 "http://ahds.ac.uk/copyrightfaq.htm"

(66.XXX.XXX.XX) is the IP (Internet protocol) address. (X indicates the presence of a number which has been removed for anonymisation purposes) This is an anonymous machine-to-machine address number used by computers to correctly send and receive data over the internet. (24/Feb/2005:00:07:12 +0000) is a date stamp and records the date and time of the file sent in response to the client's request. (GET /deposit/depintro.htm) records the file sent to the client and the directories where the file is stored on the server. (HTTP/1.1) is the record of the hypertext version communication between server and client. (200) is the status field and states if the request was correct and a file was sent and (318) records the size in bytes of the file sent. (http://ahds.ac.uk/copyrightfaq.htm) is the referrer log and states the address of the last site visited by the client.

We used the logs from the three main portals for digital humanities in the UK, the AHDS [4] central servers, the Humbul Humanities Hub[5] and Artifact[6]. In the case of the former we were able to analyse a year's worth of data. However, in the case of Artifact much less was available, due to the fact that they did not have the technical support to maintain their own logs. The data from Artifact become available when it merged with Humbul, but we had only three months worth and it appeared relatively late in the project's life. For the purposes of this paper therefore, we will concentrate on results gained from the Humbul and AHDS logs. Ideally we would have liked to use individual logs from the servers of digital humanities projects. However the project has a limited scope of fifteen months to allow us to report to the AHRC's strategy review. Gathering together log data even from the three service providers was a time consuming process, to do so from individual projects would have been unworkable given our deadline. However this is something that we would like to do in future work, given available funding.

Log data is however not a complete method in itself. We therefore chose to adopt another method successfully used by CIBER and mounted a questionnaire on the AHDS, and Humbul websites, and on that of the RePAH project, in which we asked about use patterns of resources.[7]

2.2 Qualitative Methods

The use of quantitative methods, primarily analysis of the log data, enabled us to identity a small sample of projects which could be studied in greater depth. We also studied the few projects that had been mentioned by respondents to questionnaires hosted by us and by the IHR Peer review project. (See below for details of this) We also asked those who work for the AHDS subject centres for their opinions about which five projects (per centre) were most frequently used and which five most neglected, based on the records of requests for data that they receive. We selected a sample of twenty five projects with varying levels of use, and conducted interviews about them. The interview data is too extensive to be reported in this paper, will be discussed elsewhere.

2.3 Neglected Resources workshop

2.3.1 Choice of resources

We were also wished to investigate the reasons why some projects appeared not to have been well used. Collecting data about the reasons for neglect of resources is much more challenging, and thus we held a workshop whose aim was to present users with a mixture of used and neglected resources to determine if there were any factors that users could identify that might explain the neglect of digital resources. Adams and Blandford (2002) have shown that those who do not use digital resources because they are unaware of their existence may express satisfaction with and wish to have access to such resources, once introduced to them. We therefore wished to test the

hypothesis that neglect is caused by factors other than inherent weaknesses in the resource itself, such as a simple lack of knowledge about resources

Our definition of neglect was that, from the evidence of the log data and what the AHDS subject centres told us, users did not appear to be making links to or requests for such a resource, as opposed to well used resources, which were being accessed repeatedly. We used a mixture of used and neglected resources, since we did not wish users to be prejudiced against the quality of resources that were presented as neglected. For the same reason we did not tell participants which resources were which, and chose resources about similar themes, including such areas as warfare and census data, which log data indicates were popular areas (see discussion below) The following list was therefore selected.

Neglected projects:

- Art and Industry in the Eighteenth Century[8]
- Collected Poems of Wilfred Owen[9]
- Correlates of War Project : International and Civil War Data, 1816-1992 [10]
- Exeter Cathedral Keystones and Carvings [11]
- Other Educated Persons[12]

Accessed projects

- GIS of the ancient Parishes of England and Wales, 1500-1850[13]
- Imperial War Museum concise art collection[14]
- Toronto Dictionary of Old English Corpus [15]
- Channel Tunnel Rail Link Archive[16]
- Designing Shakespeare[17]
- English Monastic Archives[18]

2.3.2 Recruitment of participants

A group of 20 participants was recruited to represent a cross section of humanities scholarship. Most had indicated interest in undertaking further research on questionnaires for various ICT strategy projects, or were these scholars' students or colleagues. Representatives of the AHDS subject centres and other digital humanities professionals were also invited. We wanted participants to be reasonably comfortable with the use of digital resources, to help ensure that a negative reaction to a project was not caused by a lack of confidence in using online resources.

2.3.3 Conduct of the workshop

Each participant had between five and ten minutes to investigate the resource and note their views of it. This is a relatively short period. However recent research has indicated that users tend to make up their mind about whether they will use web-based resources in a remarkably short time (Lindgaard et al. 2005). Nicholas et al (forthcoming) have also discovered that most visitors bounce out of websites very quickly after entering them .Thus we wanted participants to make judgements relatively quickly, as they might, if they discovered a resource for the first time, from a web search, library page or subject portal, and had to decide whether to pursue their interest any further.

Participants were given a sheet to indicate their views of each project, on which we asked the users, whether they had used the resource before; their views on its quality and possible usefulness in their research; and whether they would recommend it to a colleague or student. As a result of the wide subject coverage some resources were not suitable for all participants. It was also possible that a resource might be broadly suitable but not exactly right for their work (for example a literary resource but of the wrong historical period). In this case we asked for participants' opinions on the kind of resource, even if the exact example was not perfectly suited to them. Finally we asked whether participants thought a resource was used or neglected, and for their reasons for their conclusion.

3. Results

3.1 Findings from the Log data

The first finding emerged from the log research before we even began analysis of the data. We found that it is surprisingly difficult to extract log data even from large government-funded repositories. As outlined above, Artifact had insufficient technical help to collect more than basic reports until their merger with Humbul. We were allowed access to AHDS central server logs without difficulty. However, given the distributed nature of the service, logs are also held by the centres themselves, and so the data from these logs took longer to access and organise, and at the time of writing is still being analysed. The Humbul logs proved difficult and time-consuming to access, partly because of the presence of personalisation features in Humbul which caused concerns that individual users might be identifiable. It was therefore necessary to enter into a lengthy period of negotiation to determine how this data could be removed. We had not anticipated such problems and they undoubtedly delayed both our research and that of RePAH.

3.2 Names and subjects

The log data from the AHDS central site showed which links were being followed on the site itself, and it was possible to generate a list of pages that visitors used, which indicated which resources they were looking for, although in some cases more than one resource which might be found as a result of a query. In the case of the Humbul logs it is possible to identify which resource the user was interested in, and whether they followed links from a summary page about the resources, to its own website.

Certain names and themes recurred in the log data. Warfare, for example was a common theme, as was census data and terms relevant to family history. Place names such as Exeter, Canterbury, Gloucester, were noticeable in the AHDS data, as were terms suffrage and suffragette. Witchcraft and magic produced a large number of links through the Humbul pages; 53% visits to www.arts.ed.ac.uk (Edinburgh University), 85% of visits to Greenwich University's web page. Other popular areas

from the Humbul logs concerned popular subjects such as medieval monasticism, the English language and French Film stars. A project on Jeremy Bentham[19], based at UCL was especially well used, both in global terms and in relation to the UCL-based projects. This is understandable given his fame as a philosopher, and his well known link to UCL's foundation.

Information provided by the AHDS service providers helped underline this phenomenon. PADS (the Performing Arts Data Service)[20] told us that Designing Shakespeare is one of their most used projects. VADS (the visual Arts Data Service)[21] told us that the Imperial War Museum Concise Art Collection was often requested, as was the London College of Fashion Archive[22]. War, Fashion and Shakespeare are all very popular subjects of study. The list of well used and neglected projects provided by AHDS History demonstrated an interesting contrast. The Great Britain Historical Database: Census Data: Occupational Statistics was, unsurprisingly, well used but a similar project, entitled Enumerator Returns for County Antrim was neglected. This suggests that not only the subject but the name of the resource is significant. The latter project may not have been found because, when searching, the more intuitive term to use is 'census data' and not 'enumerator returns'. A project archived by VADS called 'Other Educated Persons' is also neglected, from which title it is almost impossible to infer the content of the resource. It concerns art in the East End of London, but the search terms used to retrieve information on this area would be unlikely to include the words "other educated persons".

However, information from the service providers at times contradicted the evidence of the logs. Literature Languages and Linguistics felt that a collection of Wilfred Owen's poetry was little used. Yet the AHDS logs showed the term 'Wilfred' relatively high on the list of pages viewed. Visual arts felt that the Exeter Cathedral Keystones and Carvings Project was seldom used, nevertheless the term 'Exeter' was significant in the list of pages visited. (Although we cannot be certain that this refers to the same project)

3.3 Names and resource description

Regardless of the usefulness of the subject matter the issue of naming and description of a resource is important, if potential users are to understand what it may be used for. In the arts and humanities, print publications often have witty, unusual and eye-catching titles, partly to attract the readers' attention and pique curiosity as to their contents. In the case of a book or article title, there is usually an elucidation of meaning after a colon so for example 'If you build it will they come?' is explained by the rest of our article title, which allows readers to comprehend its subject matter.

However, in the case of digital resources an unimaginative but accurate description is more advisable. While humanities scholars have complex models of information processing in the world of print, these are not yet as easily transferred to the digital realm. (Buchanan et al. 2005, Blandford et al. forthcoming). Thus users find it easier to guess the contents of print publications, given their greater experience of them, but find it much harder to guess the contents, purpose and scope of digital resources. Therefore if a user is browsing a digital collection and finds a resource whose title does not accurately describe it, they may become confused, and discouraged, and may not explore any further. (See the results of the workshop, described below). We must therefore be aware of the need to provide signposts for humanities users of digital resources, the most obvious of which is the project title.

3.4 The importance of information resources

The respondents to the questionnaire were enthusiastic about the usefulness of digital resources. Our respondents were regular users, 89% of whom used the web every day, and 40% for more than four hours per day. 81% identified themselves as extensive users of digital resources, 83% either agreed or strongly agreed that they could not do their work without digital resources.

Figure 1. Use of Digital Resources

Figure 2. Impact of digital resources

77% of them agreed or strongly agreed with the statement that digital resources had changed the way that they did their research.

Figure. 3. Digital resources and research

In order not to influence users too much we decided not to offer a definition of digital resources. Instead we asked them to list their three favourite resources. We found that most users perceive digital resources not to be specialist research resources for humanities scholarship, but generic information resources. This is supported by previous work carried out by Sparke Jones and the UCIS project (British Academy, 2005, Rimmer et al. 2006)

A wide range of resources were mentioned, but the most popular was the university library web site, which 14% of the users identified as the most important resource. Google, in comparison gained 8% of the votes.

Figure. 4 Most useful digital resources.

Many different resources are classified as 'other' in Fig 4. figure 4. The vast majority of them are what might be termed information resources or gateways, such as libraries, archives and subject portals, whether these are publicly funded or commercial. For example, the British Library[23], the National Library of Scotland[24], the National Archives[25], JSTOR[26], the AHDS or Humbul, SOSIG[27], Literature Online[28], the e-DNB[29]. Specialist subject centres like Palatine[30] were also mentioned, and privately constructed information portals such as Voice of the Shuttle[31] and the ORB[32], as well as subject-based digital libraries such as Perseus[33] . It therefore appears that most of our users regard digital resources primarily as a way to access information, which in the analogue world might be compared to the library or archive, rather than specialist research resources which we might compare to a monograph or a literary text for primary study.

It is significant that most resources fall into the 'other' category, which suggests that there is a very wide range of resources being used, and very little agreement as to which are most useful. This underlines the diversity of humanities research. Thus in the case of content as well as interface it is naïve to suppose that one size will fit all, unless we are discussing generic resources like library catalogues. It is therefore inevitable that only a few digital resources available will be suitable for any given scholar- as is the case for example with print journals, where a medieval economic Historian would be most unlikely to read the Journal of American Studies. It may therefore be that for the foreseeable future generic resources suit the needs of the majority of humanities scholars, because the specialist resources have not yet been created.

This finding is supported by our log data especially from Humbul, where we found numerous links to information sites. From the logs we generated a list of the forty top level domains that were accessed.

Table 1. 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%

Half of the domains listed above are for sites of libraries, archives, e-text collections, portals or publishers. We then extracted details of the sub directories belonging to the UK universities in this list: in order of popularity, Edinburgh, Sheffield, UCL, Greenwich and the School of Advanced Study (University of London). Information resources were again high on the list of resources linked to. Almost all of the School of Advanced Study pages were for the web pages of subject research centres, such as the Warburg Institute, and the Commonwealth Institute. Three of the most popular resources at Edinburgh (29% altogether) were the Centre for the History of the Book (second) the Dictionary of the Older Scots Tongue – (fifth) and the Edinburgh Journal of Gadda Studies. (The last two sites do not give access to the resource, but information about it). At Sheffield six such resources were present in the log data, Assemblage (an archaeology journal), which was the second most popular resource, The Association for Low Country Studies, CAPRA- an archaeology journal, The

Centre for the Study of the English Cultural Tradition, The International Band Dessinee Society and the Hegel Society of Great Britain.

The questionnaire recipients identified only three UK funded research projects: The Old Bailey Online[34], the PARIP project[35], and the Powys Digital History Project[36]. There were also two further USA funded research project, the CHILDES corpora[37] website and the Perseus digital library- both of which are information aggregation and reference sites, and the Photographic Exhibitions in Britain site, which is, perhaps oddly, Canadian funded although it is archived by VADS. The same question, asked by the IHR ICT Strategy programme study on Peer Review obtained similar results; in this case the Old Bailey Online was the only publicly funded project mentioned. This does not mean that respondents never use specialist digital resources, since we only asked about the three most commonly used, but they obviously do not use such resources as frequently as information aggregators, portals and libraries, whether digital or physical.

3.5 Creation versus reuse

Most of the pages viewed on the AHDS website, and those highest in the frequency list, concerned deposit and creation such as 'how to deposit', staff contact details, and information about issues such as copyright. However, this is contradicted by the views of those who use digital resources, since our questionnaire data indicates that only a minority of users (32%) believed that data archiving was central to their research.

Is data archiving once your research is complete central to your work?

16.11% 16.11% 22.82% 22.15% 12.75% 10.07%

1 2 3 4 5 6
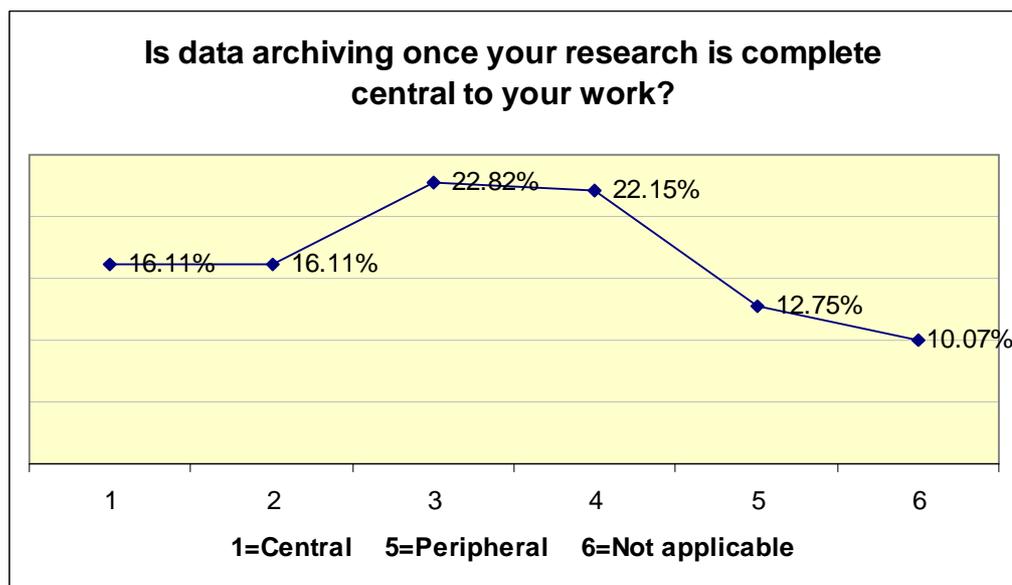
1=Central 5=Peripheral 6=Not applicable

Figure. 5. Data archiving as part of research

It is possible therefore that there is a scholarly bifurcation between those who create specialist digital resources as part of their research, but do not tend to reuse those of others, and those who prefer to use more generic information resources, but are less concerned with deposit and archiving.

3.6 Academic versus commercial users

Analysis of the logs showed whether users visited from an academic domain or from a commercial ISP. This must be treated with some caution, since if an academic user is accessing resources from home and not using a proxy server, they will be counted as using a commercial ISP. Nevertheless, we found that academic users tended to be more persistent in their information seeking patterns. While those using an ISP and commercial users were most likely to view just one page in a session (63% and 60% did so) academic users were least likely to do so, although 54% of them only viewed one page, academics were recorded as having longer sessions; 31% had sessions lasting over 3 minutes compared to 21% of commercial session users and 19% ISP based sessions. Academic users tended to persist in their search. They were least likely simply to use an external search engine, and tended to use a combination of the

on site search facility, navigation menus and browsing. As a result they tended not only to access the summary information about a resource, but to make a link to the resources itself: 39% of those who used a combination of methods (which academics were more likely to do) linked to a resource, as opposed to 25% of those who used a search engine.

3.7 Workshop findings

3.7.1 Recruitment

Our first significant finding was the relative difficulty of recruiting participants. We contacted people who were either known to the research group as users of digital resources, or those who had replied to surveys indicating an interest in further research. While digital humanities professionals were eager to take part university lecturers were harder to recruit. We initially tried to recruit a mixture of humanities computing professionals and traditional academics, and to keep a balance between subject specialists. However, this proved impossible. Many of the academics did not reply to our (repeated) email invitations, and in a notable case reacted with hostility and a demand for payment. We therefore accepted all those who were interested enough to volunteer. This meant a potential lack of subject balance, with a population skewed towards historians and archaeologists, graduate students and humanities computing professionals.

This problem may have been caused by a simple lack of time on the part of busy academics. However, it may also be evidence of a lack of interest, or perhaps confidence, in the use of digital resources amongst the mainstream academic profession. Had the subject matter been of genuine interest it is likely that more academics would have been keen to participate. There was also a marked contrast with the response from those whom interviewed about their research projects, who in most cases replied swiftly and made time to talk to our researchers. The latter were to some extent digital enthusiasts, since they had directed the development of a digital research project. This supports the impression of a divide between the enthusiastically digital (who appear to be a minority) and the majority of the academic profession.
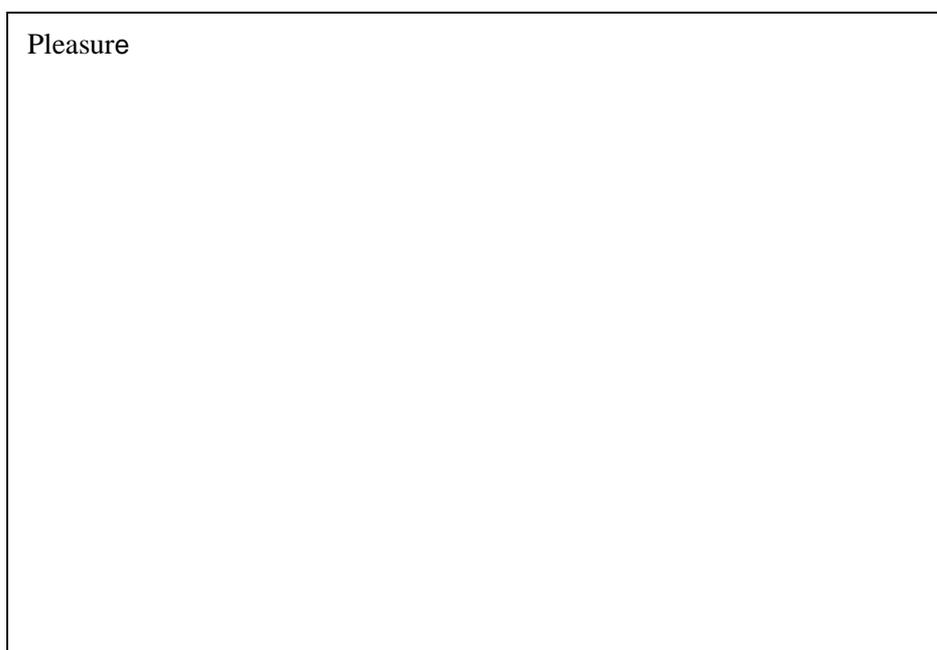
This is worrying, since there is a danger that digital humanities may therefore become ghettoised rather than further integrated into scholarship.

3.7.2 Lack of confidence

Many participants, especially those from a more traditional humanities background showed a marked reluctance to commit themselves as to the quality and usefulness of resources, especially in areas in which they were not subject specialists. Although we made it clear that all findings would be anonymised and that we particularly wanted to know their views about whether and why a resource was used, some still preferred to say they did not know, or not to provide further comments. This suggests a lack of confidence in expressing views about digital resources; especially in areas where participants were not specialist, or resources they had not so far encountered. One participant even argued that it was wrong of us to offer opinions or judge the work of others in this way. Yet it is likely that s/he will do so when reviewing a book in a scholarly journal, examining a PhD or refereeing articles for publication. However, these activities are a familiar part of the analogue scholarly world and it appears that digital resources are still too remote from the experience of the majority of humanities academics for them to feel confident in their opinions of them.

This may be explicable by reference to the Wundt Curve, which is a concept from psychology which seeks to model the relationship between familiarity and pleasure. (Saunders and Gero, 2004).
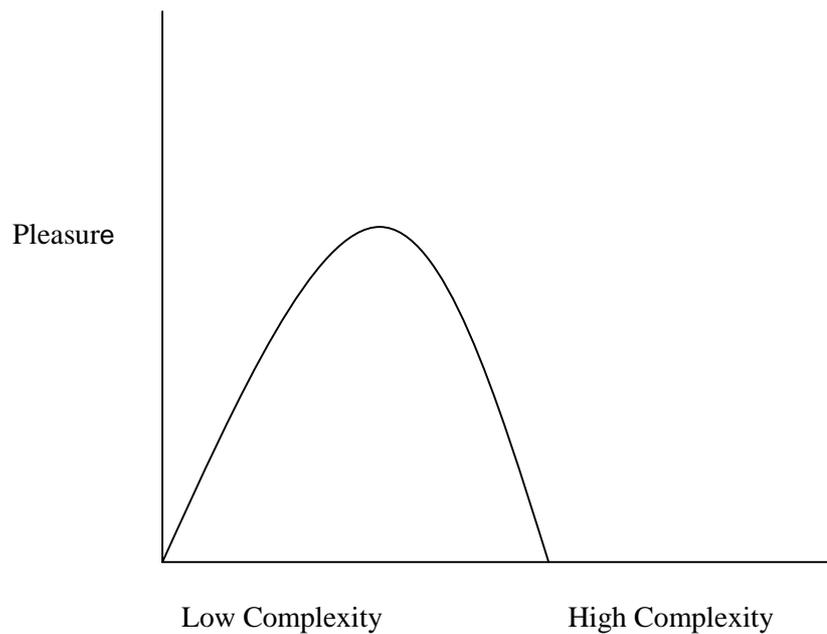
Figure 6. The Wundt Curve

Pleasure

Fig. 6. The Wundt Curve

This theory argues when something is very unfamiliar we tend to dislike it because of the cognitive effort necessary to comprehend the concept or enjoy something like an unusual art work or musical composition. However, if we are too familiar with something we do not enjoy it because there is too little cognitive effort involved, and thus the concept or object becomes banal. Ideally there should be enough cognitive difficulty to stimulate the brain without overstretch or boredom. It is arguable therefore that for most humanities academics, specialist digital humanities research resources are too unfamiliar, and thus this causes them to feel uncomfortable and unwilling either to use them or to express opinions about them. Whereas informational material, journals and subjects centres, even if accessed by a web page, are sufficiently familiar that they are more easily comprehensible, and place the user at a more optimal point on the curve. To test this hypothesis we ran the workshop again with a group of MA students from UCL SLAIS, all of whom are familiar with digital humanities resources. We will report in the findings in detail elsewhere, but they were more confident in offering opinions of the resources, as compared with the first group.

3.7.3 Critical judgement

Participants were highly critical of the resources offered, and none met with universal enthusiasm or approval. Even in cases where a resource might be useful for their work, participants provided informed critiques of its strengths and weaknesses. Participants tended to assume that a resource was not used, and thus identified half of the used resources as neglected. This was perhaps the opposite result from what we had expected, since we had assumed that if shown resources of good quality, they might recognise them as potentially useable, even if not for their personal research.

Problems noted by the participants concerned content, interface and ease of use. They recognised, and indeed required high quality resources, and tended to find resources that do not live up to this standard disappointing. This may particularly be because, as the questionnaire data showed, many of the resources that participants used most regularly are commercially produced, and thus the content and interface are usually of a high standard. On the evidence of this sample, users seem unwilling to allow for any lesser standards, even if they know a resource is not commercially produced.

There were exceptions to this, for example in the case of Other Educated Persons. Participants tolerated the rather old-fashioned interface and limited scope of the resource as it was clear that it was a Masters student's project of a relatively early date which would have been quite radical for its time in terms of its subject matter and functionality. They therefore saw it almost as a historical object that should be preserved for its own sake, and not necessarily as something that might be reused.

3.7.4 Names

Names proved to be significant to workshop participants, especially as a way of providing clues to the resource's purpose and provenance. For example, some participants were uncertain about the contents of the Channel Tunnel Rail Link Archive. Although it is a collection of records of Archaeological digs along the route of the rail link for the channel tunnel, some of the participants assumed it might be about railway engineering. They therefore concluded that it might not be used, because its purpose was not obvious.

Participants also commented favourably on naming, in the case, for example, of the Imperial War Museum Concise Art Collection. Its name described the resources

accurately and the Imperial War Museum also provided the reassurance of a trusted brand, reassuring them about the quality of the resources. This was one of the main reasons that they identified the Concise Art Collection as well used, and confirms previous findings about the importance of trusted brands on the internet, such as the BBC for provision of news. (BBC, 2006)

A participant who was a digital humanities professional objected to this emphasis on names as a descriptor, and insisted that metadata such as that provided by the AHDS would give users ample information about the purpose of the resource. Nevertheless, other participants observed that if a name was confusing, they might not be motivated to look at any further details.

The participants also commented on subject matter as a potential reason for neglect. They suggested that the Exeter Cathedral Carvings Project, for example, might not be well used because the subject might only be of interest to a minority of scholars. Conversely they were reasonably confident that the Designing Shakespeare project would be used because of its popular subject matter. Nevertheless they also expressed concern that resources that were well regarded and used in a small community should not therefore be seen as inferior to those that were relatively superficially used by a larger community.

3.7.5 Metadata and citation

Several participants raised doubts about the quality of the content, for example their only reservation about the Monastic Archives project was that it is still incomplete. This caused some participants to worry about how reliable searches of it might be. However, they welcomed the fact that the team had been clear about the state of their data. Participants were worried nevertheless that there was insufficient evidence about the provenance of the data for several resources. One participant who was an archivist worried that in many cases it was difficult to identify where original data had come from, and the methodology for selection and digitisation. S/he would have welcomed more information of the kind that in print is provided by citations and bibliographies.

3.7.6 Access problems

Participants found that anything that made it hard to access a resource was unwelcome and could deter them from using it. They also found that interfaces that helped them to manipulate the data were welcome.

Judging by the metadata, the datasets in history were all thought to be little used because it is so difficult to access the actual resources. We did not ask users to access the data in the workshop, but some more expert users tried, and found that the process of gaining permission to use resources, applying for passwords and downloading the data was difficult and time consuming, and thus a deterrent to use. As a result The GIS of the Ancient parishes of England and Wales was thought to be neglected, although it is one of the most popular resources held by AHDS History[38].

In contrast participants liked the interface to the Monastic Archives project, because the initial screen was simple and easy to use, and its subject matter obvious. The next page contained multiple search boxes to aid users to search the database. Thus the process of accessing and interrogating historical data was made as simple as possible. One historian argued that the downloaded data from the rejected datasets could be interrogated in a more complex way than that which is allowed by the front-end to the Monastic archives. However, it was necessary to be an expert user to do so.

3.7.7 Interfaces

The interfaces to the material were very varied. In general we tried to link to the view of a resource that a users would be presented with if accessing it through the AHDS. However as one participant noted, the same, or very similar data can be viewed differently if searched for directly though Google. For example we chose the Wilfred Owen poetry collection because we knew it to be a popular subject matter- concerning warfare, and by a widely known writer. However, the form in which it was presented was unattractive and limited. One of the reasons for lack of use may be that there is no information about which edition of the poems has been used, which is important to academic users. However, another reason for neglect may be that there is an attractive, usable electronic edition of Wilfred Owen's poems which can be accessed via http://www.hcu.ox.ac.uk/jtap/ . This also provides information about the edition used, and manuscript sources, as well as being part of a much larger multimedia

archive on Owen. It is therefore likely that this resource is being used in preference to the AHDS editions because of its superior content and functionality.

Given that one of the most attractive and useful interfaces, belonging to the Imperial War Museum Concise Art Collection, was designed by the AHDS Visual Arts Data service and not the project itself; we discussed who should be responsible for interfaces. Projects themselves might not have access to the design expertise possessed by AHDS centres, and one of the AHDS representatives expressed the view that allowing projects to design their own interfaces could create potential problems when resources were deposited, and for potential users, because of variations in functionality and ease of use. However s/he felt that the AHDS centres themselves did not have sufficient expertise about the project contents to design interfaces to them, and that one size could not be expected to fit all. The ideal arrangement would therefore be for projects to work much more closely with the AHDS to design interfaces, which has happened in the case of some archaeological projects, who have paid ADS (Archaeology Data Service) to design an interface for them. VADS staff also stressed their willingness to work more closely with project creators, but told us that they typically have little contact with resource creators between the initial proposal being approved and the final deposit of the resource.

3.7.8 Importance of the material

Respondents made clear that the factors described above might deter them from using a resource if they were unfamiliar with it, and not certain of its value to their work. However, they stressed that if they already know that such a resource would be vital to them, or if it were the only such resource available they would be likely to use it despite the potential problems or disadvantages. For example, linguistic corpora are vital to linguists and thus the participants thought that the York Toronto Corpus would be widely used, because of the nature of the resource. They pointed out that the interface was not especially attractive, but that it had the functionality necessary to make it usable, and thus there would be no real obstacle to a linguist who wanted to use it.

4. Conclusions and recommendations

4.1 Log availability

Many of our conclusions are relevant to both the findings of the log data and the workshop. However our first conclusion relates to the importance of log data itself. Our research shows that log data is potentially a valuable research resource, however, it is often undervalued and not maintained or made available. We recommend as a result that publicly funded centres and research projects should be under an obligation to maintain log data for an agreed minimum period, for example three years. Sufficient funding should also be made available to hire technical support to maintain the log data and make it available to publicly funded research projects, subject to a written agreement with the research centre or project. If necessary there should be the provision for a confidentiality clause, specifying that individuals may not be identified in published research output. This would remove the need for complex anonymisation of log data.

4.2 Nomenclature

The importance of naming and description of projects emerged from both the logs and the workshop, and appears to be a significant factor in determining whether a resource is used. It may seem obvious that resources addressing popular subject matter tend to be used more than those on more obscure subjects; however it is potentially significant in terms of future research funding. There are also undoubtedly excellent projects whose subject matter is highly specialist and not well known outside the immediate research community. Yet they may be vital to the work of that research field. When deciding on issues of funding, this kind of use must be weighed against the likelihood that a resource on a popular subject is more likely to be re-used although its use might be broad and shallow. For example we do not know whether the witchcraft resources are being used by fans of Buffy the Vampire Slayer rather than for research on seventeenth century popular culture. There is also a danger of a kind of 'electronic canon' being formed, (as Warwick predicted in 1999) where less well-known authors or subjects are marginalised by yet more census data and Shakespeare. (Warwick, 1999a) Funding bodies must therefore make decisions about

the value of study of more recherché topics. In effect they must develop the sense of collection building that has hitherto been the concern of librarians in the print world.

It is equally important that we recognise the value of experimental research, where a resource is created to test a technical or intellectual issue in knowledge representation. This kind of conceptual research is valuable, and it may be that no reuse can be envisaged. However, when bidding for money to produce digital resources those who propose their creation may suggest that they will be widely used, in order to justify the level of funding sought. To draw a clear distinction between material intended for a wide audience and that created to solve a research question it might be possible to ask for more evidence of usage, or to make it clear that for 'pure research' projects re-use should not necessarily be a prerequisite for funding.

The actual name of a resource also emerged as vital and potential resource creators must be aware of the importance of what they decide to call their resource, since it may have a profound effect on its future use or neglect. Although metadata can help to describe a resource it appears that non- expert users may not progress as far as accessing this. It is also important to stress the importance of organisations like universities, museums, libraries and archives brand names that users trust to produce good quality resources.

4.3 Information resources

The preference amongst users for information resources over specialist research resources has various consequences. In terms of funding priorities it suggests projects which collect together large collections of information resources for reference, whether generic or subject based are likely to be well used. Our findings also demonstrate the importance of traditional scholarly structures in humanities research and the use of the web for information about journals and academic conferences. This confirms Barrett's (2005) findings that humanities users still need traditional, generic resources and value personal knowledge repositories and face to face meeting as highly as digital resources.

It also underlines the importance of physical information resources such as libraries, archives and research centres. Access to research centres, such as those at the School of Advanced Study is obviously still very significant in the research of humanities scholars. However, more attempts might be made to integrate their physical function with that of a digital mission, following the example of the IHR (Institute of Historical Research)[39] in London and the HRI (Humanities Research Institute)[40] in Sheffield. University libraries are the primary point of access for digital resources for many users, and national and specialist libraries and archives are also highly valued. Digital resources have also not replaced the library as an important research resource. If anything their function as information gateways have increased their importance. It is therefore vital that they should be funded appropriately and that funding for digital resources should not therefore be diverted from library and archive budgets.

In a separate study (Pappa et al, 2006) we have tried to find specialist digital resources for humanities research, beginning with either the departmental home page or the university library. We found it relatively hard to find such resources, even for an expert information seeker. This might help to explain why so many of the resources being used are information collections, as it may be that these are the kind of resources that librarians, as information specialists themselves, consider most valuable, and therefore create links from the library web page to them. Thus the users tend to follow the links provided. If these do not include specialist digital humanities resources users may not look further for them, since they trust the judgement of librarians. It is therefore important that librarians should be aware of specialist digital research resources and to provide links from the library site. This may require more specialist training for subject librarians, which suggests that modules on electronic publishing and digital resources in the humanities should be available to library school students, as is the case at UCL SLAIS. For mid-career professionals it might be possible to pride training in collaboration with organisations like the AHRC Methods Network[41] a body set up to help disseminate information about the use of digital resources in the UK Higher education research community.

4.4 Critiques of digital resources

Humanities scholars are capable of providing detailed, informed critiques of digital resources. They demand the highest possible quality, both in terms of content and interface and may be relatively easily deterred from using digital resources if they are not convinced of their value. They therefore assumed that more resources were neglected than was actually the case, and this was generally a judgement made as a result of their concerns about resource quality. This may be because users have become accustomed to the kind of high quality content and interfaces that is provided by commercial publishers or organisations like large libraries and archives, and are disappointed when resources produced by academic teams do not appear to match such standards. This has important implications for the creators of specialist digital resources, since it is clear that to the majority of users, inherent scholarly value is not generally a compensation for a digital resource that is produced to lower standards than those used by commercial organisations.

In terms of content, users require as much information as possible about the quality and provenance of a resource and whether or not it is comprehensive in coverage. Users may find it difficult to comprehend the extent and coverage of digital resources since they lack the clues that we are used to in the physical world. Scholars can browse a library shelf or journal issues and quickly determine the approximate extent of the resources available, and thus be sure that they do not miss anything important, but this is much for difficult in the case of digital resources. This concern was also found by Bates (1996) when she interviewed scholars using the Getty project resources, and has also been noted by Duff et al (2004) and Dalton and Charnigo (2004) when studying historians. It is clearly therefore a long standing problem which is far from being solved. However it is relatively easy for producers to provide information about the source of the material and how it was selected. It is therefore important that producers of electronic resources should make clear the source of their data and their methodology for digitising and marking it up, and that this should be easily accessible from the web page of the actual resource as well as with data deposited with the AHDS. The extent of the resource should also be indicated, especially if it is selective or incomplete. An excellent example of this is the Powys Digital History Project where this kind of documentation is available directly from their website and written in easily understood, non-technical language.

For most users, ease of access is vital and the more hindrances placed in the path of a potential user the more likely it is that they will give up and not access the resource. A front end which makes data more easily manipulable is also very welcome. Users are clearly aware than one of the great assets of digital resources is to enable users to manipulate data in different ways. If this potential is limited, or relies on the use of separate software, scholars may be deterred from using the resource. Producers of digital resources must be aware of how easily deterred a user may be from using any given resource. Thus unless necessitated by copyright regulations, registration screens or similar hindrances to access should be avoided if at all possible. Resources should be designed to help users manipulate data directly from the web interface, to avoid users having to download and use it with separate software.

4.5 Interfaces

If an interface compares badly to the kind of professional front end that scholars are accustomed to from commercial products, this immediately creates an adverse impression on potential users. However, at present it appears that the issue of the design of interfaces to digital humanities material is managed in a somewhat random fashion. If a project is lucky enough to benefit from a good designer then the interface may be very good, however it seems as if many projects do not take this aspect of their work as seriously as the design of the back-end materials themselves. Yet if the interface makes it hard to access such material, a great deal of effort in its creation may be wasted. One participant at the second workshop likened this to producing an expensive car but neglecting to provide a steering wheel, thus rendering it impossible to drive. To remedy this situation, project creators must be aware of the importance of good interface design and spend sufficient time on developing and testing interfaces, perhaps bringing in expert advice. This activity must also be cost appropriately in terms of time and funding. Another interesting possibility is to encourage much more collaboration with the AHDS in interface construction, although at present there is insufficient funding to allow this to be done on a significant scale.

4.6 Summary

We have seen that in some cases, when scholars are aware that a resource is vital for their work, or they cannot access information without digital resources, they are willing to be extremely persistent and grapple with difficult interfaces and barriers to access to get the data that they need. The log data shows that users from academic domains tended to be more persistent and use different search strategies to reach their goals. This is important, since it suggests that academics are keen to find and potentially to use certain resources. Nevertheless, we would argue that the kind of scholar who is likely to know they need such a resource and persist until they find it is the kind of early adopter who is already using specialist digital resources. If funding bodies wish to increase the take-up of resources then they need to urge producers to take into account the kind of barriers to adoption that are described above, since these are likely to be off-putting to the less determined user, who is likely to represent the majority of humanities scholars.

As we have shown, where resources fit the needs of the research community, as in the case of information aggregation sites, scholars will enthusiastically adopt digital resource use. However at present too many digital resources require users either to struggle with unfriendly interfaces or to be technical experts even to begin to use them. It is also important to realise that humanities users are highly critical of the quality of research resources themselves, thus content must not only be of excellent quality, but must advertise this fact, by making clear what kind of material it contains and how this has been selected. If producers of digital resources can satisfy these entirely reasonable scholarly requirements in their building of resources, encouraged by the appropriate requirements specified by research councils, then we may be far more optimistic that the users will adopt digital resources for humanities research.

5. References

Adams, A. & Blandford, A. (2002) 'Digital libraries in academia: challenges and changes.' In E-P Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs & T. Costantino (Eds.) *Digital Libraries: People, Knowledge , and Technology.* 392 - 403. Springer LNCS 2555.

Barrett, A. (2005) 'The information seeking habits of graduate student researchers in the humanities.' *The Journal of Academic Librarianship.* 31(4), 324-331.

Bates, M. J., (1996) 'The Getty end-user online searching project in the humanities: Report no. 6 Overvew and conclusions.' *College and Research Libraries,* 57 (6), 514-523.

BBC (2006) Media holds its own in trust poll, available at http://news.bbc.co.uk/2/hi/technology/4965550.stm. Accessed 30/06/06

Blandford, A., Makri, S., Buchanan, G., Gow, J., Rimmer, J., Warwick, C. 'A Library or Just another Information Resource? A Case Study of Users' Mental Models of Traditional and Digital Libraries.' *Journal of the American Society of Information Science and Technology.* Forthcoming.

British Academy (2005), *E-resources for Research in the Humanities and Social Sciences - A British Academy Policy Review.* section 3.5 Available from http://www.britac.ac.uk/reports/eresources/report/sect3.html#part5.   Accessed 30/06/06

Buchanan, G, Cunningham, S. J., Blandford, A., Rimmer, J. Warwick, C. (2005) 'Information Seeking by Humanities Scholars.' *Proceedings of European Digital Library Conference, September 18-23,* Vienna, Austria, Springer Verlag, Lecture Notes in Computer Science. Volume 3652 pp. 218-229

Dalton, M. S. & Charnigo, L. 2004, 'Historians and their information sources', *College & Research Libraries*, 65 (5): 400-425.

Duff, W., Craig, B., & Cherry, J. 2004, 'Historians' use of archival sources: Promises and pitfalls of the digital age', *Public Historian*, 26 (2): 7-22.

Green, R. 2000, "Locating sources in humanities scholarship: The efficacy of following bibliographic references", *Library Quarterly*, 70 (2): 201-229.

Herman, E. 2001, 'End-users in academia: meeting the information needs of university researchers in an electronic age Part 2 Innovative information-accessing opportunities and the researcher: user acceptance of IT-based information resources in academia', *Aslib Proceedings*, 53 (10): 431-457.

Huntington P, Nicholas D, Williams P, Gunter B.(2002) 'Characterising the health information consumer: an examination of the health information sources used by digital television users'. *Libri* 52(1), 16-27

Lindgaard, G., Dudek, C., Fernandes, G. & Brown, J., (2005) 'Attention web designers: you have 50 milliseconds to make a good first impression', *Behaviour & Information Technology* , 25: 115-126

Nicholas, D., Huntington, P., Jamali, H., Dobrowolski, T. 'Characterising and evaluating information seeking behaviour in a digital (consumer) environment: spotlight on the "bouncer".' *Information Processing and Management.* Forthcoming.

Pappa, N., Warwick, C., Terras, M., Huntington, P., (2006) 'The (In)Visibility of Digital Humanities Resources in Academic Contexts'. Paper presented at *Digital Humanities 2006*, Paris Sorbonne, 5-9 July 2006.

Rimmer, J., Warwick, C., Blandford, A. and Gow, J (2006) '*User Requirements for Humanities Digital Libraries*'. Paper presented at *Digital Humanities 2006*, Paris Sorbonne, 5-9 July 2006.

Saunders, R and Gero, JS (2004) 'Situated design simulations using curious agents'. *AIEDAM* 18 (2): 153-161

Stone, S. (1982) 'Humanities Scholars-Information needs and uses' *Journal of Documentation*. 38 (4): 292-313

Talja, S. & Maula, H. 2003, 'Reasons for the use and non-use of electronic journals and databases - A domain analytic study in four scholarly disciplines.' *Journal of Documentation.* 59 (6): 673-691.

Whitmire, E. 2002, 'Disciplinary differences and undergraduates' information-seeking behavior.', *Journal of the American Society for Information Science and Technology.* 53 (8): 631-638.

Warwick, C. (1999a) 'The lowest canonical denominator: Electronic literary texts, and their publication, collection and preservation.' In M. Klasson, B. Loughridge and S. Loof, (eds.) *New Fields for Research in the 21st Century, Proceedings of the Anglo Nordic Conference 1999.* pp. 133-141.

Warwick, C. (1999b) 'English Literature, electronic text and computer analysis: an unlikely combination?' *Proceedings of the Association for Computers and the Humanities- Association for Literary and Linguistic Computing, Conference,* University of Virginia, June 9-13.

Watson-Boone, R. (1994) **'**The Information Needs and Habits of Humanities Scholars.' *Reference Quarterly*, 34, 203-216

---

[1] http://www.ucl.ac.uk/slais/research/circah/lairah/. All URLs were last accessed 20/02/07.

[2] http://www.ahrc.ac.uk/

[3] http://www.ucl.ac.uk/slais/research/ciber/

[4] http://ahds.ac.uk/

[5] http://www.humbul.ac.uk/. Humbul and Artifact have now merged to become Intute Arts and Humanities - http://www.intute.ac.uk/artsandhumanities/

[6] http://www.artifact.ac.uk/

[7] The RePAH project (http://repah.dmu.ac.uk/) is another of the ICT Strategy Projects, which is examining user requirements for subject portals. They are using the same three portals, and thus we are undertaking the log analysis of the data on their behalf.

[8] http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-4635-1

[9] http://www.ota.ox.ac.uk/ text 2216

[10] http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-3441-1

[11] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-ECKC-1

[12] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-OEP-1
[13] http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-4828-1
[14] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-IWM-1
[15] http://www.ahds.ac.uk/catalogue/collection.htm?uri=lll-2462-1
[16] http://www.ahds.ac.uk/catalogue/collection.htm?uri=arch-335-1
[17] http://www.ahds.ac.uk/catalogue/collection.htm?uri=pa-1018-1
[18] http://www.ucl.ac.uk/history/englishmonasticarchives/
[19] http://www.ucl.ac.uk/Bentham-Project/
[20] http://ahds.ac.uk/performingarts/index.htm
[21] http://ahds.ac.uk/visualarts/index.htm
[22] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-LCFCA-1
[23] http://www.bl.uk/
[24] http://www.nls.uk/
[25] http://www.nationalarchives.gov.uk/
[26] http://www.jstor.org/
[27] http://www.sosig.ac.uk/. Sosig has now become Intute Social Sciences –
http://www.intute.ac.uk/socialsciences/
[28] http://lion.chadwyck.com/marketing/index.jsp
[29] http://www.oxforddnb.com/
[30] http://www.lancs.ac.uk/palatine/
[31] http://vos.ucsb.edu/
[32] http://www.the-orb.net/
[33] http://www.perseus.tufts.edu/
[34] http://www.oldbaileyonline.org/
[35] http://www.bris.ac.uk/parip/introduction.htm
[36] http://history.powys.org.uk/history/intro/entry.html
[37] http://childes.psy.cmu.edu/
[38] http://ahds.ac.uk/history/index.htm
[39] http://www.history.ac.uk/
[40] http://www.shef.ac.uk/hri/
[41] http://www.methodsnetwork.ac.uk/