

Serving the invisible researcher: meeting the needs of online users

Amanda Hill

Manchester Computing

Kilburn Building

University of Manchester

Oxford Road

Manchester

M13 9PL

Introduction

This article looks at the characteristics of users of online archival resources and analyses the particular needs of different types of users. It goes on to look at the ways in which technology has been, and will be, used to improve access to these resources, particularly by the elusive population of archives non-users.

Who are our online users?

Online users are anonymous, for the most part, compared to the users that most archive staff meet in the course of their work. Online users may never get as far as the front doors of our record offices; they may find all the information that they need within our web pages. They may be completely unaware that they are users at all, having found what they needed and moved on. It is important to recognise that they are still users of our services and that they need to be taken into consideration in the same way as the more tangible users who occupy the tables in our searchrooms.

This marks a fundamental shift of focus for many record offices. The archives profession has always been focused on the needs of its users, but these have principally been the users who have been visiting our searchrooms.¹ Physical users are easy to count, able to make their needs known in fairly straightforward ways, and can often be prevailed upon to complete user survey forms. Remote users are somewhat harder to find out about, but should not be thought of as in any way less important (or less countable) than the users who visit record offices. There are regular comments on mailing lists about the problem of falling visitor numbers in

archive offices, but this phenomenon should be seen instead as a positive consequence of improving online access to our materials.

The problem, for some services, is that there are no procedures in place for counting remote users, despite the government's insistence in its 2002 national strategy document that "all councils are now requested to plan, target and monitor the balance of delivery of services through direct internet, telephone and face-to-face channels".² In the follow-up report, published in December 2003, issues around measuring the take-up and impact of electronic services were identified as "perhaps the least developed element of [local authorities'] approach to e-government".³ A national Take-up and Marketing project was established to investigate this area. The project is due to report its findings in the autumn of 2004.

Although remote users are harder to identify, there is a fair amount of information that can be deduced about them. The current strands of the UK archive network, for example, are able to analyse the data recorded in their web server logs. So, as an example, the origin of users of the Archives Hub website over the course of 2003 breaks down in the following fashion:⁴

Origin	Proportion of total use
UK Academic (ac.uk)	24%
Other UK use	18%
Overseas use	15%
Other (e.g. .net, .com, .org)	43%

Table I. Origin of Archives Hub users, identified by their computer's location

These figures are obtained from the domain names of the computers used to access the website. They do not, of course, tell the whole story, as many of the .com and .net

users will be based overseas and there may be academic users of the service who are accessing it from home and who will not therefore have an ‘ac.uk’ location. The identifiable overseas searches can be further analysed to show the country of origin. Principle users of the Archives Hub (from over 100 countries in all) are shown in the chart below:

Overseas use: top ten	Proportion of total Archives Hub use
Australia	4.0%
Canada	2.8%
US universities (.edu)	2.5%
New Zealand	1.2%
France	1.0%
Netherlands	0.9%
Italy	0.9%
Ireland	0.5%
Germany	0.5%
Spain	0.5%

Table II. Overseas users of the Archives Hub

Other sources of information about remote users include online feedback forms and messages sent to website email addresses. The A2A website has a ‘new user’ information form, which invites users of the site to fill in some basic information about themselves and their use of archives and archive catalogues. Information gathered from these forms indicates that approximately 80% of respondents are researching their family history and that a majority of these new users are new to finding aids.⁵

As part of an information-gathering exercise in November 2002, Archives Hub users were encouraged to complete an online form, with the chance of winning a book token. The answers to the question about ‘How would you describe yourself’ are

represented in the chart below. However, it should be noted that the relatively high number of archivists and librarians in this chart is more likely to be a reflection of the way the feedback form was promoted on professional e-mail lists than an accurate record of the day-to-day composition of users of the service.

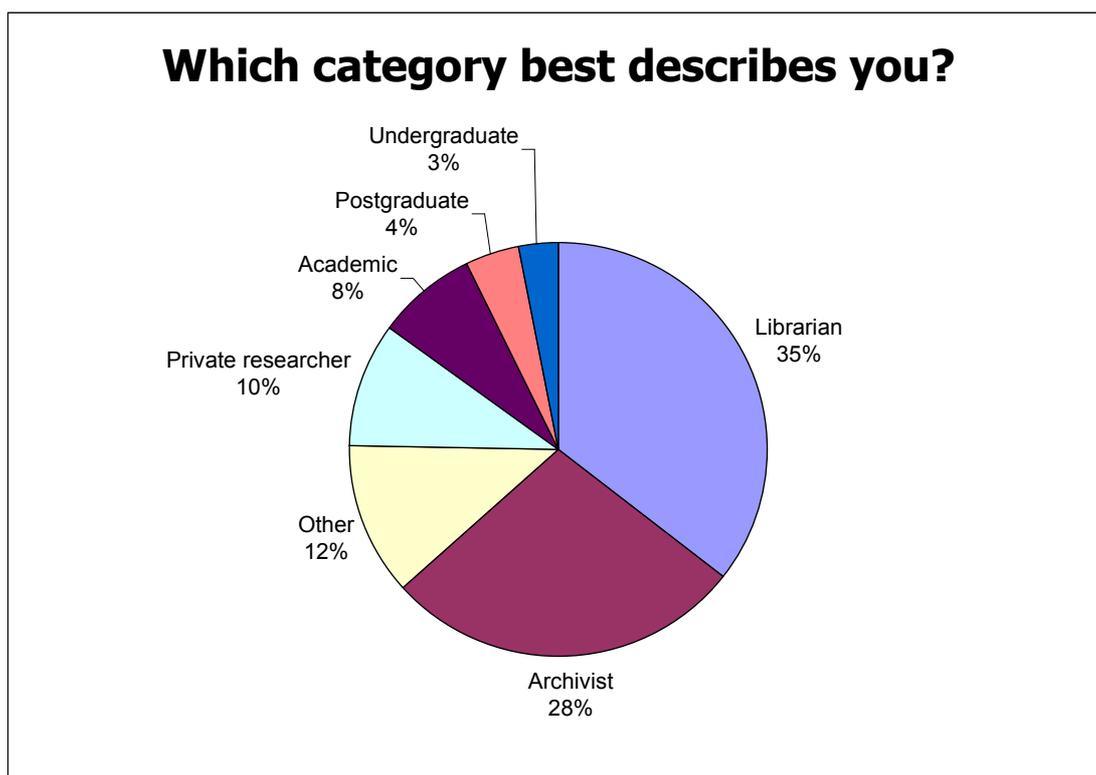


Figure 1. Breakdown of Archives Hub survey respondents, November 2002

As part of its investigation into requirements for online searching software, the LEADERS project conducted research into the searching facilities required by different types of users.⁶ Six archive repositories took part in a survey of users. These were the National Archives, the Wellcome Institute, Dorset Record Office, Birmingham City Archives, Glasgow University Archive Services and University College London's Special Collections.

LEADERS found that the majority of users (60%) fell into the category of personal leisure use. The next largest grouping (at 22% of users) was made up of people using archives as part of their job. This includes academics and professional researchers. The project went on to analyse the topic of research of these users. 64% were interested in looking for information about families, individuals or organisations, while 23% were looking for a particular topic. The correlation between these two sets of groupings was high, with 84% of personal leisure users looking for families, individuals or organisations and 85% of professional and educational users looking for topics.

This is interesting research, as it emphasises the importance of providing both detailed finding aids (which mention the names individuals and locations) and guidance as to the subject strengths of collections. In order for users to make sense of catalogues in an online form, where there are no archive staff on hand to identify collections likely to be of interest, the finding aids themselves need to carry additional information about the overall content of the archive they describe. Most of the current online archive networks already do this, to varying degrees. Thus the collection-level records within AIM25 and the Archives Hub are quite rich in subject terms. The A2A central team also encourage contributors to put index terms into the top level of description of the multi-level finding aids that they provide access to, although the level of subject indexing in A2A is variable.

The LEADERS research demonstrates the importance of access to detailed information of the type that is held in the lower levels of description, particularly for

leisure users who are looking for information on particular individuals or places. This group of users is particularly important in terms of the numbers of people it contains, or could potentially contain. They are best served by having access to detailed item-level information and clear routes to ordering copies of the material that is of interest to them. This level of service will also be of benefit to the professional researchers.

None of the current strands of the archive network completely fulfils the needs of all users. A2A is the best at providing detailed item-level information, without a doubt. The statistics show that the most frequently accessed catalogues in A2A are those for Quarter Sessions, which contain large numbers of names of individuals.⁷ Searching by topic on A2A is more difficult. This is reflected in the frequency of use of the subject searching option in A2A, compared with that in the Archives Hub.

	Number of searches	Proportion of all searches
A2A		
Total searches	282,784	
Subject searches	111	0.04%
Archives Hub		
Total searches	21,724	
Subject searches (user-entered)	276	1.20%
Subject searches (clicked on from index headings within descriptions)	2,865	13.00%

Table III. Number of subject searches performed in October 2003 on A2A and Archives Hub

In order to satisfy more of our users it is clear that we need to be providing more information (more complete catalogues) and better quality information (catalogues

accompanied by subject indexes) which would allow easier access to our archives by topic. Our online services should, of course, be accessible to all users and they should also be subject to service level agreements in the same way as our office-based services.

Access to images of the archives themselves is another area which is becoming increasingly important to all categories of users, but perhaps particularly for those who would never consider entering the doors of a 'real' record office. Some of the websites funded through the NOF digitisation programme have delivered wonderful examples of the way in which archivists, other information professionals and educators can work together to produce resources that will attract lifelong learners of all ages.⁸ Many of these resources are essentially online exhibitions of material chosen to illustrate particular themes.

Archivists have been trained to describe their holdings, but the needs of the lifelong learner/non archive-user community are creating a demand for the interpretation of materials which requires a different set of skills. The task of writing easy-to-read text for the world wide web and weaving a narrative around individual items is closer to writing text for exhibitions than it is to writing finding aids. Museum professionals and educational experts often have more experience in the field of interpretation than archivists, and many of the cross-domain NOF projects illustrate the success of this pooling of expertise.⁹

The other key area of demand for images is for the leisure users described earlier, particularly those investigating their family history. The National Archives'

experience of providing images of the 1901 census returns gave additional proof that leisure users want to access detailed information about individuals, but it also gave clear evidence of the willingness of those users to pay for images of documents relating to their research. The Documents Online¹⁰ service at the National Archives is building up another online pay-to-view resource (£3.50 per image), divided into Family History and Other Records. The Scottish Archive Network's project to digitise more than 500,000 wills registered in Scotland between 1500 and 1901 is another example of a successful initiative that charges for access to images of documents (currently £5 per will). Interestingly, this service, which might have been considered a threat to the work of Edinburgh's record agents and researchers, has actually created a new income stream for them, as users often require transcription services in order to make sense of the originals.¹¹

Archivists have been hesitant to charge for access to information, but the online family history market is one that we are in a good position to serve, and which will willingly pay for the information which we can provide. It may be many years before genealogists can browse all of the UK's parish registers online, but goals like this need to be set now and planning for them implemented, at a national level, if they are ever to be achieved. The infrastructure needed for such projects is unlikely to be put in place by every individual record office.

New ways of getting information to users

How do we reach users who do not know about A2A or the other websites? Users who do not know that what they need is buried in one of our finding aids or images?

What is the best way of ensuring that these individuals become aware of our information?

The first projects to place archival descriptions online focused on building web interfaces to the data, meaning that users had to first know about the web sites in order to be able to find out more about their contents. The archives network strands were therefore part of the ‘deep’ or ‘invisible’ web, with their contents hidden from search engines and only accessible from search forms within their own websites.

The exception to this was AIM25, whose collection-level descriptions were available to search engines early on.¹² Usage of this service is correspondingly high, with many users coming straight from search engines to the descriptions. In December 2002 the Archives Hub followed this lead and allowed Google’s robots into the ‘news’ section of the website, where around 5% of the Hub’s descriptions exist in static web pages. The increase in use after this date was marked, and the profile of the types of searches carried out in 2003 differs fairly dramatically from the profile for 2002, as the graph below shows.

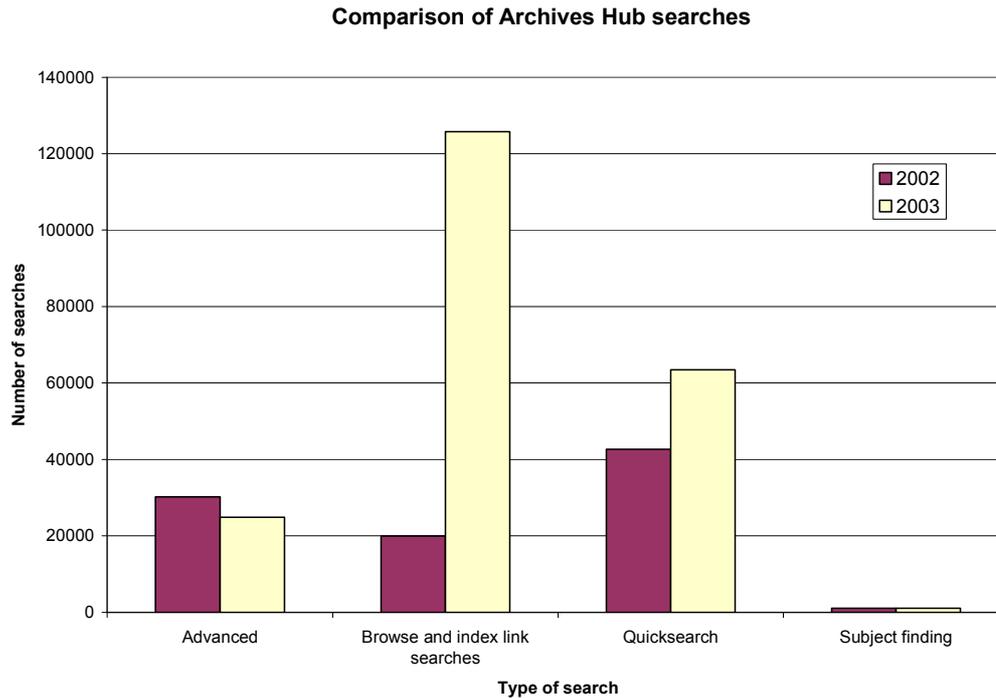


Figure 2. Profile of Archives Hub searches performed in 2002 and 2003

A recent investigation into the ways that users reach the Archives Hub showed that 84% of people came directly from search engines to those static pages. Many of these users then navigate the site using the index links associated with the descriptions. This is the reason for the dramatic rise in the ‘index links’ searches in 2003. The conclusion from this seems to be that users like clicking, but not typing! Opening up content to search engines ensures that it will reach the widest possible audience. There is a side effect of doing this that should be mentioned: after widening access to Archives Hub descriptions there was a tenfold increase in the number of queries coming in to the Archives Hub helpdesk. Most of these relate to locating or accessing archives, but a small proportion have nothing whatever to do with archives: requests for plough parts and boot laces are examples of these.

The figures provided by the A2A team in the table below show that the proportion of search engine referrals to the A2A website is comparatively low. This is because A2A currently requires users to enter at the front page and perform a search in order to reach its finding aids; it has no direct access to descriptions for search engines.

Referring site	Percentage of referrals
.gov.uk (not pro.gov.uk)	34%
.co.uk & .org.uk (not a2a.org.uk)	27%
.com & .net & .org	16%
Google	14%
.ac.uk	7%

Table IV. Sites transferring users to A2A

AIM25 has made information about its collections available through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as well as through search engines. OAI developed in the academic community as a way of sharing information about e-print journal articles, but as it requires simple information about resources, encoded in Dublin Core metadata, it can also be used to share information about other types of resource. For the AIM25 team, this has involved mapping the ISAD(G) data fields used for AIM25 records into the metadata elements needed by Dublin Core (DC). Inevitably, in this process, some of the richness of the ISAD(G) is lost. Of the ISAD(G) fields only the title, creator, scope and content, subject and access restriction fields were mapped into the DC version. Here, they are associated with the URL of the full ISAD(G) record. An example of the DC version of an AIM25 record is given below:

Title	BEVERIDGE, William Henry, 1879-1963, 1st Baron Beveridge of Tuggal, economist: Coal Crisis Papers
Author/Creator	Beveridge William Henry 1879-1963 1st Baron Beveridge of Tuggal economist
Publisher	British Library of Political and Economic Science
Year	2002-05-20
Resource Type	text
Resource Format	text/html
Note	Three bound volumes relating to problems in the coal industry, containing: Vol. 1 Printed and stencilled pamphlets and memoranda. Vol. 2 Miscellaneous letters. Vol. 3 Press cuttings from The Times 8 Apr 1921-19 May 1921.
Subject	Coal mining
Subject	Labour disputes
Subject	Labour relations
Subject	Mining
Subject	Strikes
URL	http://www.aim25.ac.uk/cats/1/5750.htm
Rights	CLOSED APPLY TO ARCHIVIST
Institution	Archives in London and the M25 Area (AIM25)

Table V. AIM25 record in Dublin Core metadata.

OAI metadata is made available through the HTTP protocol to OAI service providers, who may provide searching facilities over any number of OAI content providers.¹³ A list of OAI service providers is maintained by the Open Archives Initiative.¹⁴ One example is the OAIster search engine at the University of Michigan, where the metadata from AIM25 may be retrieved alongside that of over 200 other content providers.¹⁵

OAI is particularly suitable for records at collection-level such as those maintained by AIM25, as there is no hierarchical structure involved. AIM25 records are held in a database structure, which made mapping the data fields into Dublin Core relatively straightforward. Rendering full finding aids in EAD into OAI is more complex. A project at the University of Illinois, funded by the Andrew W. Mellon Foundation,

investigated this issue and reported on the complexities of trying to represent all the information within full EAD finding aids using OAI.¹⁶ One of the main problems identified in this report is that EAD is very flexible, creating large variations in the use of particular tags between different institutions. This means that it would be extremely difficult to create a single tool that would be able to convert the varying content of these EAD files into OAI records.

One way of allowing access to complete finding aids without having to make changes to the structure of the files is to use the library search and retrieval protocol known as Z39.50 (the name comes from the reference number for the US standard which defined the protocol). Z39.50 was used in 2002 to set up a prototype gateway that cross-searched records from the A2A and Archives Hub services.¹⁷ The disadvantage of Z39.50 is that directly searching across many databases at the same time is inefficient, meaning that cross-searching using the protocol is not scalable.

The approach used in developing the software for the distributed version of the Archives Hub was a hybrid one, where Z39.50 is used to harvest information from the remote EAD databases every day. The indexes to the data are then combined into a central 'meta-index' at the Archives Hub, which is what the users of the system access when they perform a search. The 'Spokes' distributed software can be installed at repositories, allowing staff to add, edit and delete their own EAD files, while still making them available for searching through the main Archives Hub website. It also allows the repositories to provide a local web and Z39.50 interface to their EAD files. The Z39.50 interface means that the EAD can be searched by other

systems, thus allowing cross-searching of, for example, bibliographic and archival metadata.¹⁸

The next big thing in the world of online resources is the advent of ‘web services’. This phrase refers to an architecture which supports the use of individual applications (services) directly by other applications. This means that resources can be constructed which will automatically interact with other resources. At the moment, Google¹⁹ and Amazon²⁰ are two of the big names who have started to provide services of this nature. These applications are, at the time of writing, freely available for incorporation within other websites.

The way that these services interact with each other is by passing messages structured in XML, a highly flexible mark-up language. The messages are usually passed from one system to another over HTTP, the protocol of the World Wide Web. Each service contains a machine-readable file which describes the format of the messages that can be received by that service and the responses which will be returned. One of the key things about web services is that they are a layer on top of existing applications, which may be running on any computing platform, as long as it is able to ‘talk’ XML. Likewise, the services which make use of those applications may be running on a completely different software platform, but if they are able to issue and process XML requests, it does not matter.

The UK government’s e-GIF framework is mandating the use of XML as a means of exchanging information between systems from 2005.²¹ Its view of the future is that the technologies of web services will become increasingly central to processes at all

levels of government. The impetus behind the use of web services is significant, with web standards organisations W3C and OASIS and industry giants such as Microsoft and IBM involved in pushing them forward.²²

In the archive world, it would, theoretically, be possible to add web service interfaces to existing systems and then to build applications which could provide cross-searching capabilities across any number of systems. It remains to be seen how scalable this approach would be in practice, but, as with the Distributed Archives Hub, it would have the advantage of retaining the rich metadata of the existing finding aids without having to map them to Dublin Core.

By allowing other computer systems to interact directly with our finding aids we are opening up possibilities of presenting archival data within any number of other applications and portals. These could be a world-wide archival network, a subject-specific gateway, a corporate or institutional portal, or a local search service such as the Mersey Libraries example mentioned earlier. Other possibilities for improving work flow would be allowing archives to access and update centralised databases such as one for CARN reader's tickets or another for name authority files.

Conclusion

Users of our online services are just as important as users who enter our record offices and, if we are to form a clear picture of the overall use of our services, we need to ensure that there are processes in place to count those users. In order to engage those people who are unlikely ever to venture inside the doors of the searchroom, we need

to be creating attractively designed websites with exhibition-like content. To best serve those users who do want to access original materials (or who would be willing to pay for surrogates), the information held within our finding aids needs to be as detailed and of as high a quality as we can afford. It also needs to be made as widely available as possible. Raising the visibility of finding aids through search engines is one approach, but we also need to be thinking of allowing other systems direct access to our search interfaces, either via Z39.50 or through web services. Doing this will allow the inclusion of archival data in a wide range of interfaces, aimed at a range of different users, ensuring that even people who have no awareness of archives at all can locate the information that is of relevance to them.

Archive offices are not tourist destinations, but information centres. Physical visitors are one sort of user, but will form an increasingly small proportion of our overall user profile. Providing services for the much larger body of invisible users (and generating income from them) needs to become a priority, rather than an afterthought.

¹ Remote users are often catered for with research services, but these are usually aimed at genealogists who have clearly defined aims and who have determined which records should be searched.

² Office of the Deputy Prime Minister, **The national strategy for local e-government**, (2002) p.26

³ Office of the Deputy Prime Minister, **www.localgov.gov.uk : one year on : the national strategy for local e-government**, (2003) p.34

⁴ In this example, a user is defined as someone who has performed a search on the service and viewed results.

⁵ These statistics are held on the A2A website at <http://www.a2a.pro.gov.uk/stats/NewUserStats2.htm>

⁶ <http://www.ucl.ac.uk/leaders-project/Papers/papers.htm>

⁷ <http://www.a2a.pro.gov.uk/stats/statistics.htm>

⁸ There is a portal to all the NOF digitise projects at <http://www.enrichuk.net>

⁹ The Theatre Museum's 'PeoplePlay UK' website is a particularly good example of this (www.peopleplay.org.uk).

¹⁰ <http://www.documentsonline.pro.gov.uk/>

¹¹ Information from Alan Borthwick of the Scottish Archive Network

¹² AIM25 can be accessed at <http://www.aim25.ac.uk/>.

¹³ For further information OAI see Chris Prom's report to the Society of American Archivists www.archivists.org/saagroups/tsds/OAIReport.pdf

¹⁴ <http://www.openarchives.org/service/listproviders.html>

¹⁵ <http://oaister.umdl.umich.edu/o/oaister/>

¹⁶ Christopher J. Prom and Thomas G. Habing, 'Using the Open Archives Initiative Protocols with EAD' in **JCDL 2002: Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries, July 14-18, 2002**, Gary Marchionini and William Hersch (proceedings editors). New York,

Association for Computing Machinery, pp. 171-180.

<http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>

¹⁷ www.ang.org.uk/ang/index.jsp

¹⁸ The Mersey Libraries website is an example of this kind of cross-searching, using this software

<http://www.merseylibraries.org>

¹⁹ <http://www.google.com/apis/index.html>

²⁰ http://www.amazon.com/gp/browse.html/ref=sc_bb_1_0/104-0685465-8816719?%5Fencoding=UTF8&node=3434651&no=3435361&me=A36L942TSJ2AJA

²¹ <http://www.govtalk.gov.uk/interoperability/egif.asp>

²² A good (brief) introduction to web services for businesses can be found at <http://www.webservices.org/index.php/article/articleview/429/1/61/>.