

Amanda Hill (contact author)

University of Manchester

Oxford Road

Manchester

M13 9PL

amanda.hill@manchester.ac.uk

John Harrison

University of Liverpool

PO Box 123

Liverpool

L69 3DA

john.harrison@liverpool.ac.uk

This is a preprint of an article whose final and definitive form has been published in the *New Review of Information Networking* © 2005 [copyright Taylor & Francis]; *New Review of Information Networking* is available online at:
<http://www.journalsonline.tandf.co.uk/openurl.asp?genre=article&issn=1361-4576&volume=11&issue=2&spage=159>

Relinquishing control: developing the distributed Archives Hub

Abstract: This article examines the development of the Archives Hub service, particularly focusing on the recent move from a centralised to a distributed technical architecture. It describes the Z39.50-based solution which was adopted to enable such a system to be constructed and explains some of the technical and administrative challenges that have been faced during the transition.

INTRODUCTION

The Archives Hub is a union catalogue for descriptions of archive collections held in UK universities and colleges. It developed as a direct consequence of the creation of a large number of electronic catalogues of archives in the higher education (HE) sector in the mid 1990s. These had been funded from the JISC (Joint Information Systems Committee) Non-Formula Funding of Specialised Research Collections in the Humanities Initiative: often known as ‘Follett Funding’ after the Follett Report of 1993 (1) which recommended this disbursement. It was soon realised, however, that there was no easy way of searching across all of the separate catalogues that had been created, somewhat lessening the value of the work.

In order to test the technical feasibility of a cross-searching application for archives, the National Networking Demonstrator Project (NNDP) was established by the JISC with remaining Follett funds. This project ran in 1997 - 1998 and covered catalogues supplied by a range of local, national and university archives, in a range of different formats, to investigate the possibilities for cross-searching catalogues using the

Z39.50 search and retrieval protocol. While the NNDP project was running, the National Council on Archives published a key document called *Archives Online* (2) which set out a vision of a National Archives Network which would provide access to the UK's archival information and would be free at the point of use.

With only nine months in which to explore the possibilities, the NNDP was unable to deliver a sophisticated service, but the experiment was sufficiently encouraging for the JISC's Archives Sub-Committee to propose the development of a one-year pilot project for the higher education sector, then known as the HE Archive Hub. Funding for the creation of content was granted by the JISC to 15 universities, while the contract for providing the necessary infrastructure was awarded to a collaborative proposal from CURL (the Consortium of University Research Libraries), the University of Liverpool and MIMAS at the University of Manchester (3). The work to successfully extend the Cheshire II information retrieval system (which had originally been designed for use with bibliographic records) to work with Encoded Archival Description (EAD) files took place at Liverpool in 1999-2000. In 2000 further funding by the JISC was awarded to the partnership to develop the project into a service (4).

CONTENTS

The Archives Hub was officially launched as a JISC-funded service on 13 March 2001. At that stage, there were 3,000 descriptions of archives within the database and 15 universities' collections were represented. Three further tranches of funding were made available to universities and colleges for the creation of content, bringing the

total number of institutions represented on the Archives Hub to over 130, and the number of archive collections described to around 20,000 by July 2005.

The funding offered by the JISC to institutions was for collection descriptions only. In other words, they would pay for one summary entry to describe each of the collections held by an institution. This meant that the detail held within full archive catalogues was not included, but that there was now a high-level overview of many collections which had previously not been described on the Web in any way. Lessons learnt from the NNDP meant that the quality of these collection-level descriptions was high: the minimum standard for the metadata was considerably more stringent than the mandatory requirements of ISAD(G); the International Standard for Archival Description (General) (5).

To make the process of creating descriptions as straightforward as possible, an online template was devised for contributors. This allows the archivists, curators and librarians who compile descriptions to enter data according to the headings supplied by the ISAD(G) standard. The template then converts this information into XML form, structured according to the EAD format. The resultant EAD file is then ready for onward transmission to the Archives Hub, or for use within a local system. This latter option has not been widely used in the past, but this is beginning to change, as the next section of this article will show.

DISTRIBUTING THE DATA

It was recognised at an early stage of the development of the service that some institutions would want to retain control over their own archival metadata, while at the same time making the information available to gateways such as the Archives Hub. So although the Hub began life as a centralised database, there had always been the intention of examining ways in which the data could be held in a distributed manner. Software for searching and displaying EAD data is not currently readily available; most archive repositories in the UK use proprietary databases or word-processed files to hold their archival descriptions, with EAD chiefly being used in collaborative networking projects such as the Archives Hub, A2A (6) and JANUS (7).

The aim was to develop a version of the Archives Hub's software which would be suited to the needs of contributing repositories: providing a simple means of adding new descriptions and deleting existing data. The software would also need to provide a web search interface, for users of the repository, and access on a machine-to-machine level so that the Archives Hub could continue to search the data that was held in these remote servers. This software would also make it easier for repositories to add the lower levels of description which their paper or electronic catalogues contained, but which they had not yet put on to the Hub in EAD form, due to the nature of the funding from the JISC.

The main technical issues in designing and implementing a distributed search service are deciding on a method of access to the remotely held data, maintaining that access, and ensuring that the speed of response for the user searching the combined data remains acceptable.

Method of Access

The Cheshire information retrieval system has always included by default a Z39.50 server. Z39.50 is an international standard, which defines a protocol for information retrieval via machine-to-machine connections. The protocol allows several different operations to be carried out, the most commonly used being:

- **Search:** Creates a result set on the server based on the provided search criteria
- **Sort:** Sorts the records in a specified result set on the server side, based on the provided criteria
- **Fetch:** Fetches a specified number of results from the specified result set (or the last result set created during the current session by default)
- **Scan:** Returns a portion of a specified index, containing a specified number of terms, relative to the position of a specified term. This is often used to produce browsable lists of index terms.

Z39.50 has been used by the Archives Hub since its inception to provide an interface to the data held within the catalogue. Given the existing Z39.50 expertise within the development team, it was decided to extend the use of this standard machine-to-machine protocol to form the basis of a distributed search service. Larson (8) explains the overall approach to information retrieval that underlies the distributed Archives Hub. This involves gathering index terms from the distributed databases, forming them into a 'meta-index' at the centre and then using that information to identify the appropriate databases to target for the searches submitted to the Archives Hub. The process of ensuring access to the remote servers and harvesting their indexes is described in further detail below.

Maintaining Access to the Data

The problem of maintaining access to the remotely stored data is probably the most crucial in running a successful distributed search service. After all, there are alternative methods of accessing data across a network, but if no data is accessible, there would be no service. Much of the responsibility for maintaining access lies with whoever is hosting the data, but there are a few steps that have been taken to minimise the effects of inaccessible data on a combined search.

Information about the availability of the remote databases (known as Spokes) is maintained in a relational database. This relational database contains information about when each Spokes server was last accessed, when it was last updated and a flag for whether it is currently available. A script, scheduled to run every five minutes, checks the responsiveness of the remote servers, and updates the relational database accordingly. Once a server has been unresponsive for more than 15 minutes the Archives Hub service team is alerted by email, so that the problem can be investigated further. This means that any remote server problems can usually be corrected, and access restored, fairly quickly.

In case of longer-term problems with access to any of the remote servers, the Archives Hub offers repositories the option of storing a copy of their data on the central server at Manchester Computing. The link to the remote data can then be switched to the local copy in the event of prolonged down-time. This also provides institutions with the reassurance of a remote back-up of their data, which in its turn is backed-up. The process of transferring the data is set up to take place automatically, once a week.

Although the software was designed as a completely distributed service, many of the contributing repositories are not ready or able to host their own data and the Archives Hub continues to host this data on its main server. The data from each repository is set up in separate databases, known as 'Virtual Spokes', on the main server. This model means that it would be fairly trivial to move a Spoke out onto a remote server should the providing institution decide to change to this method of providing access to their data.

Harvesting Indexes

Information about how to access the data held at the Spokes (remote and virtual) is contained within ZeeRex files, which are held at each server. ZeeRex is an XML representation of the Z39.50 Explain information, which includes details about the hosting machine, the Z39.50 port number, the name of the database, the number of records it contains, the types of searches that it supports and the date on which it was last updated (9). The central Hub maintains a copy of each ZeeRex file in a Cheshire database. For virtual Spokes, the ZeeRex files exist in this location only.

All the index terms for the remote and virtual Spokes are retrieved by using the connection and searching details from the ZeeRex file to perform a Z39.50 scan search on each of the indexes that the Spoke supports, (for example, title, subject, full-text). The lists of index terms thus gathered are then stored in a single XML file which contains all the index terms found in that Spoke. The individual XML files for each Spoke are then themselves indexed to create the meta-index. For any given

index term, the meta-index will contain information about the identity of the Spokes which contain that term. This allows searches to be limited to only those databases which hold matching records. If a user submits a search on 'table tennis' in the title index, this would significantly reducing the amount of time taken to perform a search.

The script which checks the availability of remote servers also checks the ZeeRex database for the 'last updated' information. If the date and time have changed for any database then a field within the relational database is changed to show that the database's indexes need to be re-harvested. Another script runs overnight to harvest the updated remote indexes and update the meta-index accordingly. This helps to ensure that the Archives Hub's meta-index is regularly updated and as synchronised with the information at the Spokes as possible.

Searching the distributed Archives Hub

A user submits a search through an HTML form on one of the Hub pages. The central server checks the relational database to see whether there are any Spokes which are inaccessible. If there are, the query is not passed on to them. The central meta-index is then searched to find which repositories have the submitted terms in their indexes and the query is then passed on to those which match. The results delivered by the Spokes are then merged, sorted and formatted for display in the Archives Hub's web interface.

Speed of Response

It is perhaps more important that users of the Archives Hub receive a correct response rather than a quick one. However in these post-Google times people have to come to expect a fast response from search engines. Even with the steps taken to minimise search times described earlier, it was found, when the distributed service was first implemented, that searches on the Archives Hub were taking an unacceptable amount of time. The service level agreement with the JISC stipulates that an average 'quick search' will take no more than five seconds, but these searches were regularly taking around 15 seconds. After a little investigation it was discovered that the majority of the time was being spent establishing connections to the Z39.50 servers. This was to be expected, as the search request has to be sent across the wire, a software server needs to be started up at the Spoke and then the connection has to be established. More surprisingly though, it was taking much longer to establish a connection with a local z-server, connecting to the virtual Spokes on the central Hub, than with the remotely held ones (e.g. Leeds and Liverpool): consistently around 164 milliseconds for the local server compared with 12-50 milliseconds for the remote ones.

This suggested that the bottleneck was caused by the databases held on the central server. In other words the problem was due to the centralised implementation of a system designed to be distributed. This did in fact prove to be the case. The locally hosted z-server was responsible for serving over 130 databases, which meant that every time the search client tried to establish a connection, the z-server had to get the configuration of each of these databases, even though we would only ever try to search one of them.

There were a few potential measures that could be taken to reduce the start-up time for the central z-server. Only one however was in keeping with the idea of a distributed service and this was to distribute the virtual Spokes further. Ideally each would be held on a separate machine and so would be served by a separate z-server, this however was not feasible. The compromise was to split the virtual Spokes amongst a number of z-server configurations: effectively distributing them. Ten separate z-server configurations were created to handle requests for connections on different ports (2100 – 2109). The virtual Spokes were separated into these ten different z-server configurations arbitrarily by name.

Having made these modifications the time taken to establish a connection to one of the local z-servers fell in line with that of remote servers, at around 25 milliseconds. This had the effect of reducing the average time taken by a quick search on the Archives Hub from 10-15 seconds to around the promised five seconds. To further reduce the response times of the web interface it is proposed that in future, when the central Hub's software is upgraded to the Cheshire3 software, the remote searches will be performed in parallel, rather than sequentially (10). This should have the effect of reducing the response time to correspond directly with the time taken by the slowest Spoke, rather than the sum of the time taken by all the Spokes searched.

EARLY IMPLEMENTERS

The Archives Hub's Contributors' Forum was closely involved in the development of the Spokes software. The Forum comprises a group of archivists and librarians who have been involved in supplying the Archives Hub with descriptions of their archives

and who maintain an active interest in the improvement and development of the service. Many of the individuals in the group were interested in running the Spokes software themselves, so had a particular motivation for ensuring that the software met the needs of their users, as well as of their own staff.

The first Spoke implementations were at the University of Liverpool, followed by the John Rylands University Library of Manchester, the University of Leeds and the Modern Records Centre at the University of Warwick. The Rylands' Spoke was first installed in December 2002 and the ones at Leeds and the Modern Records Centre in February 2003. There then followed a lengthy period of amendments and developments to the Spokes software, to ensure that the software behaved as required by those who would be using it. Edinburgh University Library installed the Spoke software on a test server in October 2004.

The final version of the Spokes software to use Cheshire II was known as Cheshire for Archives 2.3. This software was fairly close to the requirements of the Contributors' Forum, with the exception of the highlighting of search terms, which was proving difficult to implement. The software was now easier to customise, with part of the administrative interface allowing the insertion of institutional names, logos and links to home pages. More substantial modifications could be made by anyone familiar with HTML and Cascading Style Sheets to alter the appearance of the web interface to suit the requirements of the repository concerned. The screen shots below show the different front-ends that were developed by staff at Liverpool, Edinburgh and John Rylands.

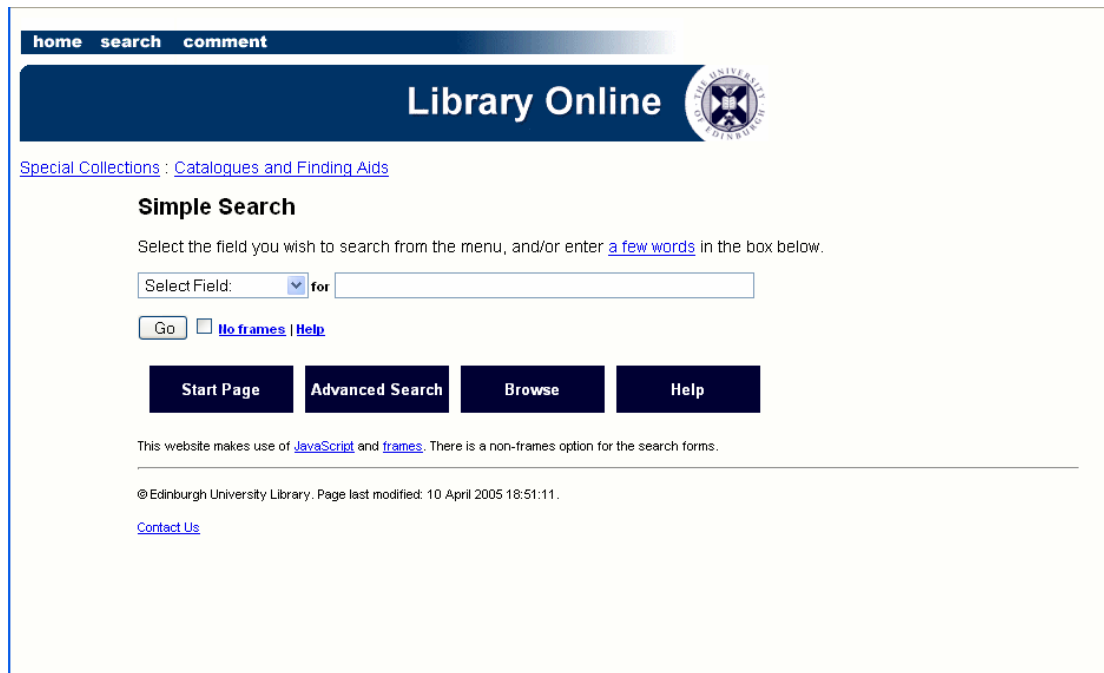


Figure 1: Home page of Edinburgh University Library's Spoke

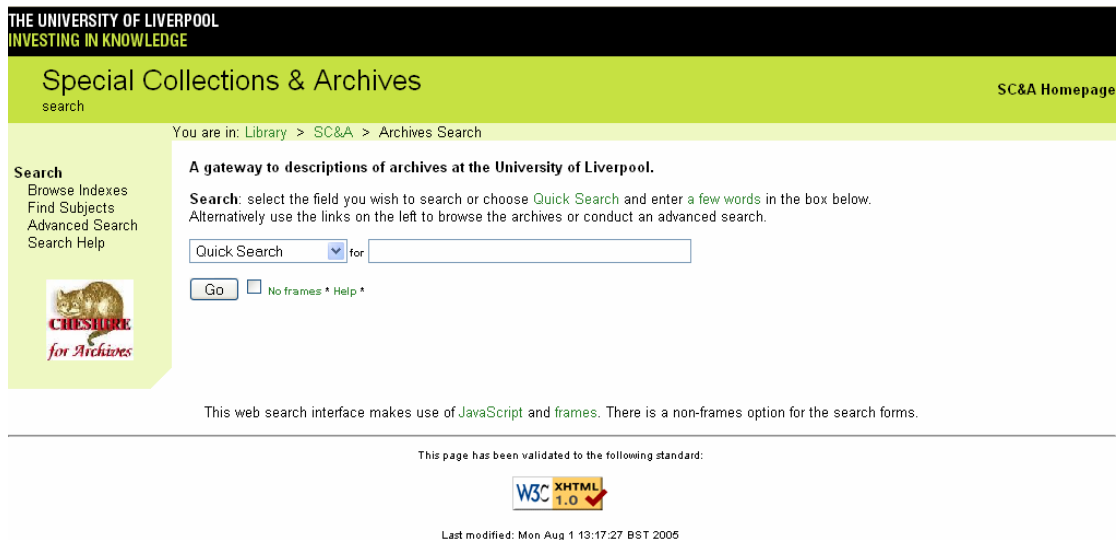


Figure 2: Home page of the University of Liverpool's Spoke

Library Home	A gateway to descriptions of archives in The John Rylands University Library.
Archives Home	Search: choose Quick Search and enter <u>a few words</u> in the box below, or select the field you wish to search from the drop-down menu.
Search	Alternatively, use the links on the left to browse the archives or to conduct an Advanced Search.
Browse	Quick Search <input type="button" value="v"/> for <input type="text"/>
Advanced	<input type="button" value="Go"/> <input type="checkbox"/> No frames Help Getting started
Help	

This website makes use of [JavaScript](#) and [frames](#). There is a non-frames option for the search forms.



Figure 3: Home page of the John Rylands' Spoke

CURRENT SITUATION

The next version of the Spokes software, based on the new Cheshire3 software, has been designed by the Cheshire Development Team at the University of Liverpool and is now nearing completion. The specification for Cheshire for Archives 3.0 is available on the Archives Hub's website (11). This new version of the software will allow further customisation of the display of records, which will be based on XSLT rather than the current custom-built CGI script. It will also include support for Unicode and will provide additional machine-to-machine search interfaces through SRW (Search and Retrieve Web Service) and SRU (Search and Retrieve URL Service).

The distributed version of the Archives Hub went live on 12 July 2005, searching the remote data created by Liverpool, Leeds and John Rylands and the remaining data in the virtual Spokes held in Manchester Computing. It is planned that Edinburgh's Spoke, which will cover the University Library and the Lothian Health Service

Archive, will be added to the list of remote databases in the autumn of 2005. A number of other potential Spokes are evaluating the software to see whether it will meet their requirements. Significant advantages for the institutions include the open-source nature of Cheshire for Archives, its use of XML standards and the support that is available from the Archives Hub service and development teams.

ISSUES RELATING TO THE MOVE TO A DISTRIBUTED MODEL

The move from a centralised database to a (partially) distributed one has raised a number of management issues and resulted in some changes to existing procedures. One important change has been the removal of central control over the quality of the metadata that is held on the remote servers. Part of the quality assurance procedure for the centralised metadata is a manual check by the Hub's Data Editor on all new descriptions that were submitted to the service. This process includes ensuring that the metadata descriptions conform to the EAD 2002 Document Type Definition (DTD) in the correct encoding, that they meet the Hub's Data Creation Guidelines (12), and that the controlled access fields of place, personal, family and corporate names and subject headings have been formulated according to national or international standards.

When repositories move to hosting their own descriptions this centralised quality assurance is no longer an automatic part of the data loading process, meaning that there is a risk of poor-quality metadata being loaded on to the remote servers. In some cases the size of EAD files can also cause problems: some archival descriptions can translate into individual files of over ten megabytes in size. These can have an

adverse effect on performance at the central service. The servers themselves are another potential problem: the Hub team have no control over the security and maintenance of the remote servers. Any failure at a Spoke will result in data being unavailable for retrieval through the Archives Hub.

Some of these issues are being tackled by a Memorandum of Understanding which is signed by each institution that plans to supply its data to the Archives Hub. The Spokes software is freely available for installation and use by non-profit-making organisations, so no restrictions can be imposed at the point of download. This would be undesirable anyway, in that it might discourage potential users from testing the software. The Memorandum of Understanding only needs to be signed if an institution is ready to supply its archival descriptions to the Archives Hub through its Spoke. As is the case with much of the Hub's development, the memorandum was a community effort, having been drawn up with the assistance of the contributors to the service.

The memorandum (13) covers topics such as the security and availability of Spokes servers, recommendations as to data quality and back-up, and staff training. It also outlines the support that institutions running Spokes can expect to receive from the Archives Hub team.

One consequence of the move to a more distributed architecture which is already obvious is the changing role of the Hub's service and development teams. With the centralised model, contact with contributing institutions was principally between the Data Editor and the staff creating descriptions in the repositories. It is expected that

this interaction will continue, both for contributors to the virtual Spokes and archivists at real Spokes with EAD-related queries, but the Hub teams are now also communicating with systems support colleagues in the Spokes' institutions. This often begins with answering general queries about the software, followed by giving advice on the recommended hardware and software environment for a Spoke, helping them to install the Cheshire for Archives software and then helping to resolve any technical hitches.

FUTURE PLANS

In June 2005 the Archives Hub ran a prize-draw questionnaire with the aim of gaining feedback from users about the future direction of the service. Respondents were asked to rate five possible activities as high, medium or low priorities for the future, and were also given the opportunity of suggesting alternative aims. The graph in Figure 4 shows the number of responses that were given to each of the activities.

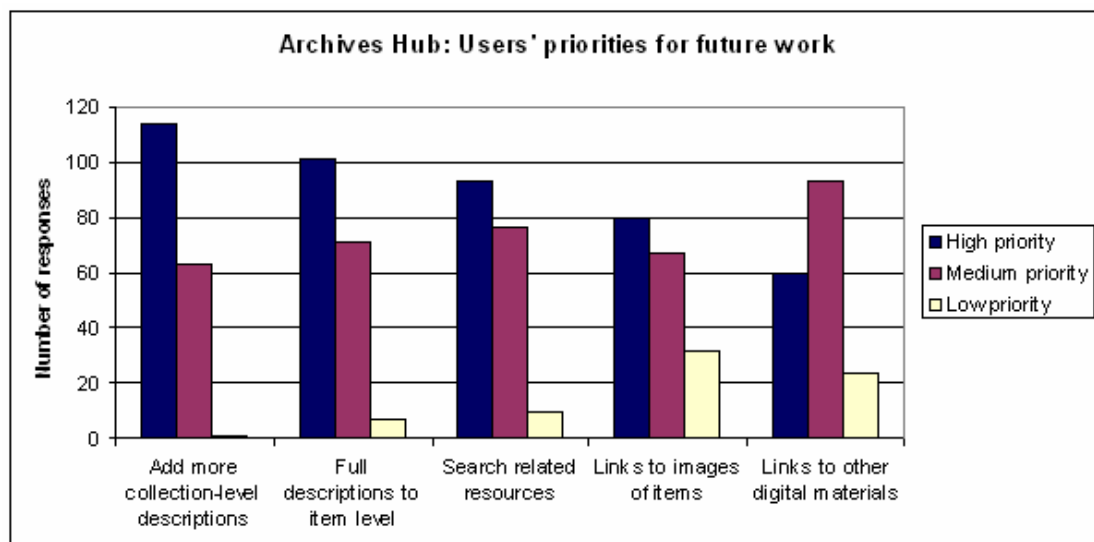


Figure 4: Users' priorities for future work

The highest priorities for users of the service were the addition of further descriptions, both at collection-level (which is the basis for much of the current content, as described above) and in the form of full catalogues. The inclusion of item-level information within the Hub has been a consistent request from users from the earliest evaluations of the service (14). The Spokes software now makes it possible for repositories to add these lower levels of descriptions to their existing Hub collection level entries. One example of this is the Spoke maintained by the staff of the Special Collections and Archives division at the University of Liverpool (15). A project of converting paper catalogues into full EAD descriptions has now been completed, enabling detailed information about individuals and places to be retrieved by users, as described by Lumb (16). It is hoped that other repositories will be able to expand their Spokes in a similar manner.

The Contributors' Forum will continue to be an important conduit for Spokes users to feed back their experiences of using the software and to discuss future enhancements. One of the strengths of this project has been the commitment and involvement of the HE archives community in the development of the distributed system.

CONCLUSION

The distributed version of the Archives Hub had been in development for four years before its successful launch in 2005, but is now demonstrating its advantages over the centralised approach. The technical architecture, built around international archival and interoperability standards, enables archivists and librarians to take control of the editing and display of their own metadata descriptions *and* to make that information

immediately available to external services such as the Archives Hub. Plans for a national archives network are still under discussion, seven years on from the publication of *Archives Online*, but this model will enable archives to expose their descriptions to any future network development. For UK archives, the distributed Archives Hub is a first step away from their stand-alone catalogues or centralised, often out-of-date, union databases and towards the service-oriented future of the internet.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of Ray Larson, Rob Sanderson Clare Llewellyn, Steve Tattersall, Paddy Collis and Jane Stevenson in the development of the distributed Archives Hub and in the writing of this article. The support of the JISC, the Archives Hub Steering Committee and the Contributors' Forum are also gratefully acknowledged.

1 <http://www.ukoln.ac.uk/services/papers/follett/report/>

2 NATIONAL COUNCIL ON ARCHIVES, *Archives On-line: the establishment of a United Kingdom archival network* (Birmingham, 1998), available online at <<http://www.ncaonline.org.uk/materials/archivesonline.pdf>>

3 The March 1999 call for proposals for the HE Archive Hub service can be found at http://www.jisc.ac.uk/index.cfm?name=funding_3_99

4 The experiences of the NNDP and the early development of the Archives Hub is described more fully in A. Hill, 'Bringing Archives Online through the Archives Hub', *Journal of the Society of Archivists*, Vol. 23, No. 2, 2002

5 ISAD(G), available online at <http://www.icacds.org.uk/icacds.htm>

6 A2A can be found at <http://www.nationalarchives.gov.uk/a2a/>

7 JANUS, a network for Cambridge, is at <http://janus.lib.cam.ac.uk/>

8 LARSON, R. R. Distributed IR for Digital Libraries. Paper presented at the European Conference on Digital Libraries (ECDL), available at <http://cheshire.lib.berkeley.edu/ECDL2003.pdf>

9 For more information on Zeerex, see <http://explain.z3950.org/>

10 More on Cheshire3 can be found at <http://cheshire3.sourceforge.net/>

11 The specification for Cheshire for Archives 3.0 is at <http://www.archiveshub.ac.uk/arch/spokes3.shtml>

-
- 12 Available online at <http://www.archiveshub.ac.uk/arch/dcg.shtml>
- 13 <http://www.archiveshub.ac.uk/arch/memorandum.shtml>
- 14 The evaluations of the Archives Hub are available from <http://www.archiveshub.ac.uk/introduction.shtml>
- 15 <http://archives.liv.ac.uk/>
- 16 LUMB, R. 'If it's not a Spoke, fix it!' available at <http://www.archiveshub.ac.uk/arch/livspoke.shtml>