Bringing Archives Online through the Archives Hub

Amanda Hill

Manchester Computing
Kilburn Building
University of Manchester
Oxford Road
Manchester
M13 9PL

Bringing Archives Online through the Archives Hub

Amanda Hill, Manchester Computing

Archivists all over the world are becoming increasingly involved with advancements in technology which are changing the ways in which users find out about our collections. The funding opportunities and technological developments of the past five years are twin drivers that have transformed the way in which UK archive services make information available. Traditional differences between repositories, sectors and domains are beginning to be eroded by the adoption of shared standards and by participation in collaborative projects. The vision of a National Archives Gateway outlined in the National Council on Archives 1998 report *Archives On-line* is now within reach.[1] A series of 'strands' of the National Archives Network have now been developed, which include A2A (Access to Archives, which covers repositories in England), AIM25 (Archives in the M25 area) and SCAN (the Scottish Archive Network). The Archives Hub covers archives that are held in the higher and further education sector throughout the UK.

This article describes the evolution of the Archives Hub service and highlights some of the technological advances that have been made during the development of the project. It then goes on to examine the future possibilities for resource discovery across all archive sectors and ultimately across the museum, archive and library domains.

History

The Archives Hub developed from the experiences of the National Networking Demonstrator Project (NNDP) which reported its findings in September 1998. This project was funded by the Joint Information Systems Committee's Non-Formula Funding of Specialised Research Collections in the Humanities Initiative (known as NFF), which had been granting funds for the cataloguing of archive collections in the UK's universities. One of the conditions of this round of funding had been that the resultant finding aids should be made available online. Consequently, the NFF Archives Sub-Committee was keen to investigate the feasibility of cross-searching multi-level finding aids that had been created by a variety of institutions and which existed in a variety of formats.

The majority of participants in the project were from the higher education sector, but local record offices and national institutions were also represented. A commercial firm, Fretwell Downing Informatics, was awarded the contract to:

> …demonstrate and report on how Z39.50 and ISAD(G) might be successfully employed to provide multilevel cross searching of a range of nominated archival catalogues prepared using a variety of software.[2]

Z39.50 is an information retrieval standard which was already being widely used for cross-searching library catalogues.[3] The 26 elements of the International Standard for Archival Description (General) - ISAD(G) - were becoming accepted as a basis for the creation of automated archival finding aids.[4] The Archives Sub-Committee wanted to see if a search and retrieval system combining the two standards could be applied to archival descriptions.

The project had an Archival Consultant, Keith Sweetmore of the West Yorkshire Archive Service, who was charged with the task of co-ordinating the contribution of data and acting as a channel of communication between the technical consultants and the archivists supplying data. The first meeting of potential contributors took place in June 1997 and the project was due to be completed within nine months. In their report, the technical consultants noted that

> With the benefit of hindsight it is clear that NNDP attempted to achieve in nine months what projects plan in two or three years in such programmes as eLib and the European Framework.[5]

The NNDP did successfully demonstrate the potential for a distributed archives network, based on the Z39.50 search and retrieval standard. It also demonstrated the enthusiasm of members of the archive profession for the idea of a national network of archive catalogues. However, the exercise also highlighted a number of areas which would require further investigation if the system were to function as a service. The key issue from the technical team's point of view was the need for finding aids to

> …conform to international professional standards…in a manner compliant with the key principles of ISAD(G) with clear record identifiers, well-defined levels and consistent entry of the core elements.[6]

This is echoed in the archival consultant's report:

> 'ISAD(G) compliance' is not enough, more explicit standardisation of practice is required, with attention to both structure and content.[7]

He went on to recommend that

> …a subset of NNDP participants from within the HE sector, capable of investing time and money in order to bring their data precisely to the required standards, should be established with a view to taking research and development forward.[8]

The results of the National Networking Demonstrator Project were sufficiently encouraging for the Archives Sub-Committee to propose the development of a pilot project for the higher education (HE) sector. Funding was made available by the Joint Information Systems Committee (JISC) to develop what was then called the HE Archive Hub, to populate it with content from HE institutions and to test the system. The

procurement document for the Infrastructure Provider considered that the Hub would become

> …an integral part of the Distributed National Electronic Resource. In addition it is the intention that the HE Archive Hub will be one element of a comprehensive networked archive catalogue…and so will be taken forward in co-operation with the NCA and the wider archival community.[9]

Pilot Project

The contract for developing the service was awarded to CURL, the Consortium of University Research Libraries, with the University of Liverpool as the lead site and Manchester Computing at the University of Manchester as a partner. Dr Paul Watry of Liverpool's Special Collections and Archives Division had experience in developing online cross-domain retrieval systems, while CURL and Manchester Computing were already co-operating in the development and delivery of the COPAC online union catalogue.[10]

The NNDP experience had a number of consequences for the pilot Hub project.   Most important was the decision to require contributing institutions to conform to an extended ISAD(G) mandatory element set. Inconsistency in data had caused problems for the NNDP and it was important to have guidance in place for contributors from the outset. The NNDP had used a proprietary format for the data it contained, but in the pilot project it was decided to use Encoded Archival Description (EAD) as the base format for the descriptions.[11] Some contributing institutions were already using EAD to create their finding aids and a mapping from ISAD(G) to EAD had recently been published by the EAD Working Group. EAD is a non-proprietary format that is increasingly becoming accepted as an international standard for archival finding aids, enabling the exchange of data between repositories.

Fifteen universities throughout the UK were funded to provide content for the pilot project. Pearl Romans and Chris Woolgar of the University of Southampton were the Data Editors with the task of overseeing consistency of the data. Archivists sent paper copies of their descriptions to Southampton for checking, before sending electronic versions to Liverpool for inclusion in the Hub. The electronic form could either be EAD SGML (Standard Generalised Markup Language) files or ISAD(G)-structured text files. During the pilot phase (August 1999 to July 2000) the contributing institutions created over 3,000 descriptions. Most of them were collection-level descriptions, but around 10% were full finding aids. Dorothy Johnston's article in a recent edition of this journal describes the data creation experience from a contributor's viewpoint.[12]

The pilot project achieved a great deal in a short period of time. The software developments that were required to implement the pilot were significant: the developers had to implement a Z39.50-compliant system that would support the retrieval and display of EAD-encoded finding aids. Z39.50 is widely used in the library domain as a way of cross-searching catalogues of different repositories, but it had not previously been used to

search finding aids in EAD. The 'Cheshire' online information retrieval system was enhanced and extended by the development team at the University of Liverpool in order to make such searching possible. The pilot project also developed the ability to 'drill down' and retrieve search terms from full multi-level catalogues.

The complex background work on software developments became visible to the contributors with the establishment of the first version of the Archives Hub interface on the Liverpool server in the autumn of 1999. By February 2000 the software had been transferred to Manchester Computing. Below is an image of the home page as it looked in December 2000. This interface evolved rapidly in response to comments from the data contributors, given at regular meetings of the contributors and developers and through the Archives Hub mailing list.
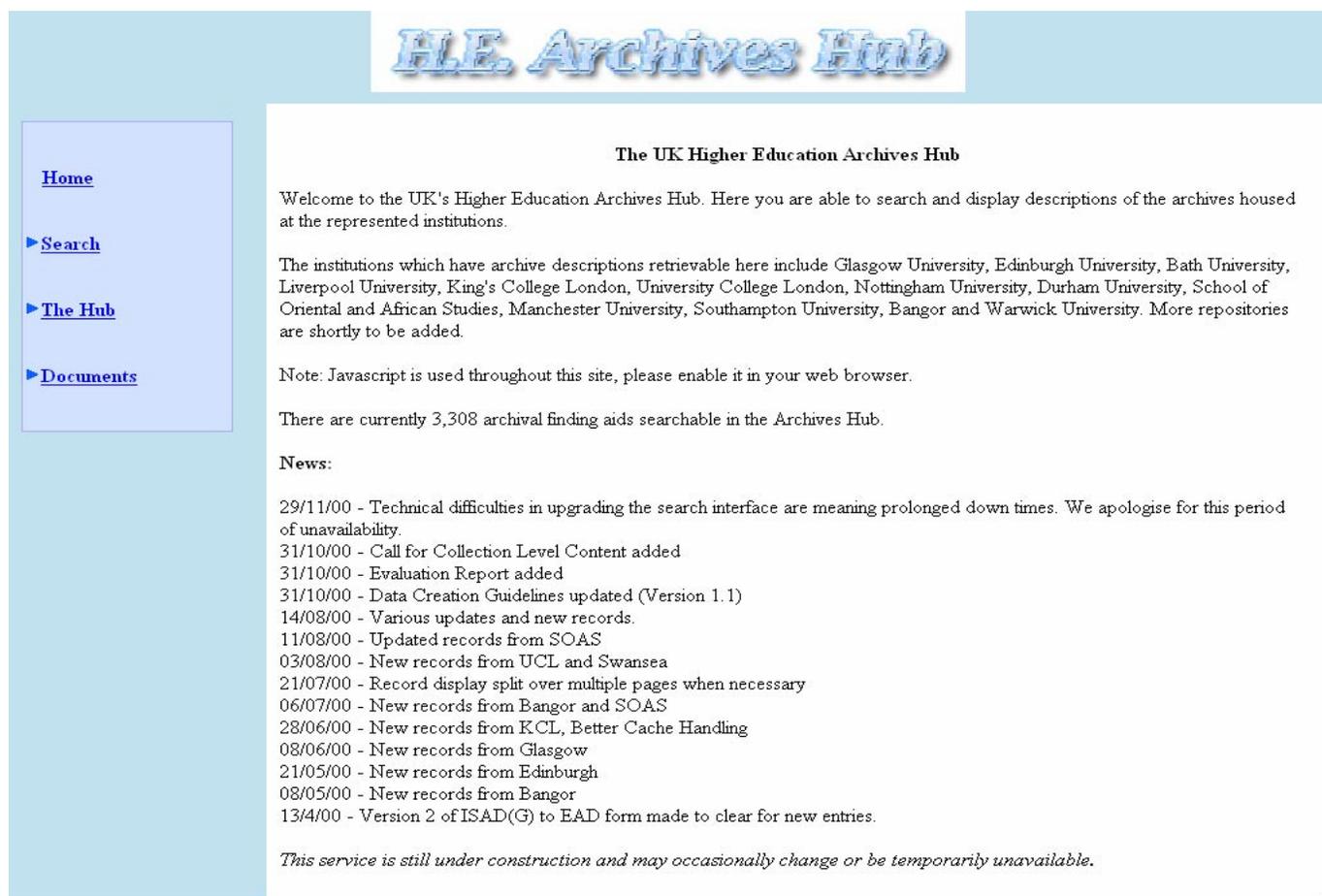
Figure 1. Archives Hub Home Page, December 2000

The pilot project had been a success, at a cost of £218,000. £143,000 had been awarded to the data editors and contributors and £75,000 to the software developers. In July 2000 the Joint Information Systems Committee approved funding for a further three years to develop the pilot into a full service.

From Pilot to Service

The creation of a service generated a need for training, marketing and co-ordination roles that had not been part of the pilot project. The team at Liverpool continued to be funded for software developments, but they were joined in January 2001 by the service team, based at MIMAS, a section of Manchester Computing in the University of Manchester. MIMAS is a JISC-funded National Data Centre and already runs a number of national services such as COPAC, zetoc and JSTOR. The service team's first task was to redesign the Archives Hub website (www.archiveshub.ac.uk), which in March 2002 looked like this:

Figure 2. Archives Hub Home Page, March 2002

Funding for new collection-level content had been granted to 24 institutions in early 2001. Between them they promised over 13,000 new descriptions for the Archives Hub. In December 2001 the remaining funds for content were allocated to a further 11 institutions. Some of these repositories were acting as co-ordinators for collaborative projects between several local institutions. A number of other institutions with relatively few collections (or whose descriptions are already in the required form) are providing their records without the need for funding; bringing the total number of institutions that will be included in the Archives Hub to over 50. By June 2003 the number of collections available for searching through the website and through the Z39.50 protocol will be over 20,000. In total £500,000 has been awarded to UK higher education institutions to provide for the creation of EAD-encoded collection-level descriptions for the Archives Hub.

JISC's priority has been to achieve comprehensiveness at collection-level, even though the Hub's software is capable of searching and displaying multi-level descriptions. In 1997 the NFF Archives Sub-committee commissioned a 'Survey of Needs' of archive-holding institutions in the HE sector.[13] Of the 70 institutions (rather than repositories) that responded to the survey, all but 14 are represented on the Archives Hub and/or the AIM25 service run by King's College, London.[14] However, it should be noted that there is still a significant amount of material (both catalogued and uncatalogued) in UK higher and further education institutions that is not yet covered by either service. The Archives Hub's last 'Call for Content' was oversubscribed by 70% and it is clear that there is still a lot to be done before any claim to comprehensiveness can be made.

Innovations

A number of reports were written after the completion of the Archives Hub's pilot phase. These summarised the lessons learnt and made recommendations for the next phase of the project. The report of the Data Editors (unpublished) emphasised the need for more detailed guidance for contributors on the creation of data, particularly in relation to subject indexing. As a consequence, Richard Higgins of the University of Durham's Archives and Special Collections was commissioned to produce a set of 'Data Creation Guidelines'.[15] These guidelines gave much more specific advice on creating descriptions for the Hub. Some of this guidance has now been superseded by changes to the way index terms are marked up, but the document remains a useful source of information for Archives Hub contributors. Full training on the creation of descriptions is now available for contributors.

Both the Data Editors' report and a technical review by Kevin Ashley of the University of London Computer Centre mentioned the desirability of a mechanism for contributors to preview their finding aids before they went 'live'. The contributors had also found the process of having to first submit data in a paper form frustrating. A priority in the subsequent phase of the Archives Hub was to simplify the data creation procedure. The paper stage of the process has been removed and the job of Data Editor now involves mounting new descriptions on the Hub as well as checking their consistency. In September 2001 a new template was introduced which allowed contributors to preview their descriptions as they would appear in the Hub and then edit them, but without making them 'live'. This template is known as the EAD Wiki and allows archivists to create and edit their records on the server in Manchester from their web browser.[16] Records created in this way can then be gathered directly from the server and put into the live system. Many current contributors are now using the EAD Wiki to test their records before submitting them and also to allow colleagues to proof-read them easily.

In the pilot phase a decision had been taken to allow contributors to use either the Anglo-American Cataloguing Rules (AACR2)[17] or the National Council on Archives' Rules for the Construction of Personal, Place and Corporate Names (1997).[18] Subject terms could be taken from the UNESCO thesaurus or from the Library of Congress Subject Headings. This reflected existing practice within the contributing institutions and was seen as an

experiment to determine whether plural approaches could be accommodated in an online service. The table below shows the different combinations of rules and subject schemes that were used by the pilot phase contributors. In the current phase of the service there is a similar distribution, except that contributors are often using LCSH subjects where there is no equivalent term in UNESCO.

Standards	Number of Contributors
NCA & UNESCO	7
NCA & LCSH	5
AACR2 & LCSH	3

Table I. Standards used by pilot phase contributors.

The differences between AACR2 and NCA rules are small but significant in an online context. The treatment of compound personal names is the main problem. Following AACR2 John Nicholas Blashford-Snell would be indexed under Blashford-Snell, whereas using the NCA rules he would be listed under Snell. Consequently, browse searches of indexes may return two entries for the same individual. The use of name authority files would resolve this problem and work is taking place now in the United States on the development of an XML (eXtensible Markup Language) Document Type Definition (DTD) for defining name authority files. The DTD is known as Encoded Archival Context (EAC) and will be compliant with the International Standard Archival Authority Record for Corporate Bodies, Persons and Families.[19] According to Daniel Pitti, "EAC is intended to extend and complement EAD".[20] EAC would provide a basis for an international name authority file for archives that could be linked into EAD descriptions.[21]

Subject indexing is a skill with which many archivists are unfamiliar. 'More training on subject indexing' is a consistent request on the feedback forms from Archives Hub training courses. In brief, contributors are advised to 'Summarise Significant Stuff' in their index terms, using international standards to do so. The Archives Hub website provides the ability to search both UNESCO and LCSH to help contributors find appropriate terms.[22] A technological solution to the issue of searching finding aids based on two different thesauri has been developed by the team at Liverpool and is now available from the 'Browse Indexes' page of the Archives Hub website.[23] This feature is called 'Subject Finding' and it allows a user to enter a natural language word or short phrase in order to see how the Archives Hub contributors might have indexed the term. A 'Subject Finding' search on 'atom bomb' retrieves the following index terms:

Subject Browsing

Your request was submitted and has matched the following subjects.
Click on the name to see the finding aid, or to list the finding aids if it occurs in more than one.

Subject	Relevance
atomic bomb blast effect	100%
atomic bomb physiological effect	99%
bombing aerial	91%
operation overlord	89%
world war 1939-1945 aerial operations american	89%
world war 1939-1945 campaigns france 1944-1945	88%
world war 1939-1945 aerial operations	88%
world war 1939-1945 aerial operations british	88%
operation buffalo 1956	7%
blast effect	6%
nuclear weapons testing	6%
methodism	1%
atomic energy moral and religious aspects	1%
medical physics	1%
pantelleria battle of 1943	1%
world war 1939-1945 africa north	1%
v-1 bomb	1%
v-2 rocket	1%
intelligence service	1%
strategic defense initiative	1%
asphyxia	1%

Figure 3. 'Subject Finding' results for 'atom bomb'

Sometimes the results may seem surprising, as in the return of 'Methodism' here, but the feature does mean that researchers can be more confident that they have not missed any significant resources. 'Subject Finding' emphasises the importance of using index terms to record and summarise the main themes of collections. Consistent use of standards in creating the index terms means that searching across the records of different repositories is more efficient. If a user clicks the word 'Methodism' (a Library of Congress Subject Heading) in the example above, they will retrieve records from the universities of Durham, Sheffield and Swansea, all of which have significant relevance to Methodism, rather than just passing references.

Increasingly, online archive services will be sharing information about their descriptions by using 'metadata harvesting' protocols such as that developed by the Open Archives Initiative.[24] In the harvesting process, key elements of descriptions such as titles, dates, references and index terms are made available for other services to 'harvest' and re-use. Consequently, the information gathered is not the full text of a catalogue but an outline of it. In such abbreviated descriptions the use of appropriate index terms will be of particular significance.

The Future

The Archives Hub is now established as a key element of the UK National Archives Network.   The development team are now working on the creation of software that will enable Archives Hub contributors to host their own data. This will create a 'distributed' service using the software's Z39.50 capabilities and will give the contributors greater control over the creation and editing of their descriptions. Many existing contributors have undertaken to continue sending new descriptions to the service after the end of their funded projects, so that sending an entry to the Archives Hub will become as automatic as sending a copy of a new finding aid to the National Register of Archives.

Close connections are maintained between the Archives Hub, the Historical Manuscripts Commission and the other networking projects. In December 2001 a joint evaluation exercise took place, funded by Re:source. The report of this evaluation is now available on Re:source's website[25]. This report showed a high demand for an option to cross-search all of the services from one interface. Technological progress on the ability to cross-search all the strands is continuing apace; one example being the way in which the National Register of Archives now carries links directly from its summary descriptions to the fuller records within the various networking projects.[26]

The various network projects have demonstrated the capabilities of current technology in providing access to the UK's finding aids online, but it is clear that pulling the strands together and extending the comprehensiveness of their coverage will need to be a priority if the National Archives Network is to reach its full potential. The simple truth is that soon, for many researchers, if a resource is not listed online, it does not exist and will not be used. Securing the resources to convert hard-copy finding aids into the required electronic form and to tackle cataloguing backlogs will be essential.

Our aim should be that within the next few years all new catalogues are created in a form that can be mounted for retrieval and cross-searching on the Web. With software manufacturers increasingly offering EAD export as an option, this will soon be achieved. Training archivists in the use of the necessary standards must also be a priority. The post-graduate archive training courses are already incorporating EAD and ISAD(G) training, but this will need to be extended throughout the profession. Subject indexing is another area that is extremely significant in the context of online retrieval but for which there is little training currently available.

It is too early to assess the impact of the networking projects on the use of archives at repositories, but the assumption is that, at the very least, demand for reprographic orders will increase. It is also likely that the number of personal visits to archives will rise as information about their holdings becomes more widely disseminated. With developments such as Z39.50 cross-searching and sharing of metadata information through the Open Archives Initiative it is highly likely that archival finding aids will be retrieved alongside descriptions of museum collections and library materials. As a

consequence, our catalogues will reach a wider audience than has ever been possible before, both nationally and internationally.

Strategies will have to be developed to cope with the likely increase in use of our resources and to capitalise upon it. For heavily-used materials a programme of digitisation may be required. The Public Record Office's experience of providing the 1901 census returns online has shown that demand for digitised images of key resources is extremely high. This was highlighted in the Full Disclosure Retrospective Cataloguing Framework document:

> 2.4 There is an approaching future where the public user will expect, as a matter of course, to retrieve not only the record, but in certain cases an electronic surrogate of the document or artefact they require.[27]

But piecemeal retroconversion and digitisation projects will be increasingly unlikely to attract funding: collaboration between repositories and across sectors and domains is becoming essential. Funding bodies will want to be assured that there is an overall strategy in place and that their money is contributing to a widely-shared vision.

Conclusion

The advent of the World Wide Web has had an enormous impact on the ways in which people discover information. This has generated a corresponding impact on the ways in which archivists, curators and librarians deliver information about their resources. It is now up to the professionals in these fields to work together to exploit the opportunities that this new medium is offering. The archives domain is now in a strong position to build on the work of programmes such as A2A, AIM25, SCAN and the Archives Hub.

[1] National Council on Archives, **Archives On-Line: The Establishment of a United Kingdom Archival Network** (Birmingham, 1998), available online at
http://www.archives.org.uk/publications/archivesonline/aolintro.asp

[2] D. Kay and I. Ibbotson, **National Networking Demonstrator Project: Technical Consultants' Report** (December 1998), available online at http://www.kcl.ac.uk/projects/srch/reports/fdi.htm

[3] For more information, see the Z39.50 maintenance agency pages at http://www.loc.gov/z3950/agency/

[4] The second edition of ISAD(G) is available online at http://www.ica.org/biblio/com/cds/isad_g_2e.pdf

[5] **NNDP: Technical Consultants' Report** (1998), section 3.4

[6] **NNDP: Technical Consultants' Report** (1998), section 7.2

[7] K. Sweetmore, **National Networking Demonstrator Project for Archives: Report of the Archival Consultant** (November 1998), available online at
http://www.kcl.ac.uk/projects/srch/reports/ksfinhtmldoc.html

[8] **Report of the Archival Consultant** (1998)

[9] Higher Education Funding Councils' Joint Information Systems Committee, **HE Archive Hub: Infrastructure and Management Procurement (1999-2000)** (March, 1999) available online at
http://www.kcl.ac.uk/projects/srch/actives/coreact/inpro.htm

[10] http://www.copac.ac.uk/

[11] For more information on EAD, see the official EAD website at http://www.loc.gov/ead/

[12] D. Johnston, 'From Typescript Finding Aids to EAD (Encoded Archival Description): a university case study' **Journal of the Society of Archivists**, vol 22 (2001), pp 39-52

[13] L. and S. Will, **Survey of Needs** (January 1998) available at
http://www.willpower.demon.co.uk/JISC/Survey1.htm

[14] http://www.aim25.ac.uk/

[15] Available at http://www.archiveshub.ac.uk/dcguide1.pdf

[16] More information on the EAD Wiki is available from http://www.archiveshub.ac.uk/eadwiki/

[17] **Anglo-American Cataloguing Rules**, 2nd edition (London, 1978)

[18] National Council on Archives, **Rules for the construction of personal, place and corporate names** (London, 1996). The NCA rules are also available online at http://www.hmc.gov.uk/nca/title.htm.

[19] International Council on Archives, **ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families** (The Hague, 1994). The ISAAR (CPF) is available online at http://www.ica.org/biblio/com/cds/isaare.html

[20] D. V. Pitti 'Creator Description: Encoded Archival Context', paper presented at the University of Sydney's Computing Arts 2001: Digital Resources for Research in the Humanities conference. Available online at http://setis.library.usyd.edu.au/drrh2001/papers/pitti.pdf

[21] See the EAC website at http://www.library.yale.edu/eac/

[22] The search forms are available at http://www.archiveshub.ac.uk/unesco and http://www.archiveshub.ac.uk/lcsh

[23] http://www.archiveshub.ac.uk/searchsub.shtml

[24] For more information on the Open Archives Initiative, see the website at http://www.openarchives.org

[25] http://www.resource.gov.uk/information/research/respubs2002.asp#nanurg

[26] The National Register of Archives is available from the Historical Manuscripts Commission's website at http://www.hmc.gov.uk

[27] Full Disclosure Implementation Group, **Retrospective catalogue conversion and retrospective cataloguing: a framework for funding bodies** (Revised September 2001). Available from the British Library Cooperation and Partnership Programme website at http://www.bl.uk/concord/otherpubfulldisc1.html