

**NATIONAL NAME AUTHORITY FILE: REPORT TO THE NATIONAL
COUNCIL ON ARCHIVES**

by Peter Gillman
© The British Library Board, 1998

The opinions expressed in this report are those of the author and not necessarily those of the British Library.

This British Library Research and Innovation Report may be purchased as a photocopy or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK.

Contents:

1 Introduction	p 3
2. Terms of reference	p 4
3. Definitions	p 5
4. Background	p 6
5. Survey of repositories	p 8
5.1 Technical Capability	p 8
5.2 Findings Aids	p 9
5.3 Authority control	p 11
6. Interviews and Visits	p 13
6.1 Major National Institutions	p 13
6.2 Scottish and Welsh Archive Networks	p 13
6.3 Other Institutions	p 17
7. Form and function of a National Name Authority File	p 21
7.1 Passive file	p 21
7.2 Active File	p 21
8. Means of delivery and access	p 23
9. Construction and maintenance of a National Name Authority File	p 25
9.1 Examples	p 26
9.2 Data capture	p 26
9.3 System management	p 26
9.4 Host site	p 28
9.5 Data protection and intellectual property	p 29
10. Conclusions and recommendations	p 31
11. Bibliography	p 33
12. Appendix: Summary data from responses to questionnaires	p 34

1 INTRODUCTION

The National Council on Archives (NCA) was established in 1988 as a forum in which the interests of owners, custodians and users of archives in the United Kingdom could be represented and openly discussed, and matters of common concern be brought to the attention of the public, the government or relevant institutions. The NCA IT Standards Working Party was set up in 1990. *Information Technology Standards and Archival Description: report of a working party to the National Council on Archives* (March 1991) recommended (5.5.A(ii)) the establishment of “a standing committee to initiate activity and monitor progress. We suggest that this should bring together appropriate representatives of the national repositories, the Historical Manuscripts Commission and the Society of Archivists, with power to co-opt experts as required” and (5.5.C(ii)) that the national repositories and the Historical Manuscripts Commission should “examine the desirability of standardising name authority controls at national level”.

Standardisation of Name Authority Controls: a report to the NCA IT Standards Working Party from the Name Authority Project (June 1993) recommended (7.4) “that the National Council on Archives appoint a committee with authority to draw up the rules on which the National Name Authority File should be based and resolve the conflicts in current local practice, taking account of the points raised in the present report and the cataloguing rules of the various repositories which are supplied as supporting documentation”.

The *NCA Rules for the construction of personal, place and corporate names* was published in 1997. The International Council on Archives Ad Hoc Commission on Descriptive Standards publications, *General International Standard Archival Description* (ISAD(G)), 1994, and the *International Standard Archival Authority Record for Corporate Bodies, Persons and Families* (ISAAR(CPF)), 1996, provide a general framework in which the *NCA Rules* may be applied for consistency in archival descriptive practices.

The NCA IT Committee in promoting the acceptance of these standards recognises that archivists may be reluctant to abandon long-established, in-house practices unless they perceive a facility like access to National Name Authority Files (NNAFs), which could act as a common cataloguing resource. The parallel work of the NCA Networking Policy Committee, which is investigating the structure and management of an archival network in the United Kingdom, places importance on the creation of NNAFs as a central means of public access to the complex and confusing pattern of survival and deposit of private records and papers in record offices and libraries.

Accordingly a successful application was made to the British Library Research and Innovation Centre to fund a consultancy to investigate the extent of automated finding aids in archival institutions in the UK and the preparedness of archivists to co-operate in the creation of NNAFs. Peter Gillman of The Information Partnership was appointed as consultant. As an information specialist, rather than an archivist, he has brought a fresh and wider perspective to this work. This is his report to the NCA. A Steering Group², which met four times between 19 May 1997 and 16 January 1998, provided general and professional advice.

Dick Sargent, Historical Manuscripts Commission

2 TERMS OF REFERENCE

The invitation to tender required a *survey of the state of automation of archival finding aids in the United Kingdom and assessment of the viability of developing national name authority files*³.

The three elements to be investigated were:

1. the extent of automated finding aids in record offices and other archival institutions;
2. preparedness of archivists to co-operate with the creation of national name authority files; and
3. requirements and costings for establishing a central server to maintain and disseminate the authority files.

3 DEFINITIONS

During the early stages of this project, while the questionnaire was being drafted, the problem of terminology became apparent. Even with a Steering Committee consisting of experts in the field, differences of opinion on terminology were quite marked. This led to some difficulty in framing the questions. It was clear from a number of responses that, despite the Steering Group's attempts to standardise meanings that could be reflected in the questions, some respondents found the terms used to be ambiguous.

To avoid a similar confusion for the reader, the terms and definitions used in this report are, as far as is possible, drawn from the English language version of the ICA Dictionary of Archival Terminology. Terms likely to cause particular confusion are defined in footnotes within the text.

4 BACKGROUND

This study was commissioned by the National Council on Archives (NCA), with a view to continuing work started with the publication in 1993 of the NCA IT Standards Working Party on the Standardisation of Name Authority Controls. The latest outcome of this work has been the NCA *Rules for the construction of personal, place and corporate names* (NCA Rules) in 1997, published during the course of this project.

There has been a great deal of activity in the archives profession over recent years to formalise and standardise various aspects of archival recording. This has resulted in ISAD(G) for cataloguing and ISAAR(CPF) for authority file structures. There is now a great deal of goodwill towards the idea of implementing common standards for formats and, to a lesser extent, actual descriptions. Some of the work done has attempted to follow paths already established by librarians: often the results have been unsatisfactory or, at best, a poor compromise. The reasons are not difficult to see.

Librarianship is concerned with recording the physical attributes and data elements of documents - mostly of books, although latterly of video and audio tapes, CDs and the like. This physical description has concentrated on the actual appearance of a work (size etc.), but more importantly on established elements such as title, author(s), publication date and place, pagination, and ISBN. A subsidiary concern has been to record the locations of separate copies of a work to facilitate library inter-lending.

This activity led to the development of the MARC (Machine Readable Cataloguing) format for the inter-change of bibliographic records. MARC is a highly-structured format with fields and sub-fields designated for all possible data elements comprising a bibliographic description. MARC specifies the type of information to be entered in to these fields: it does not specify the form of that information. For this purpose librarians have developed tools such as the Anglo-American Cataloguing Rules (AACR2). Less standardisation is apparent in areas such as subject indexing and classification where, outside the sphere of the public and university libraries, specialist schemes proliferate to reflect the particular needs of user groups.

As a parallel activity there has been developed a wide range of library management software designed to carry out the routine functions of acquisition, cataloguing retrieval and control of stock lending. The emphasis in much of this software is to manage library transactions, and to provide coherent links with other software packages, and with external sources of bibliographic data.

The above description clearly places the management of archives well outside the concerns of librarianship. There remains, however, a sufficiently strong superficial similarity between the bibliographic recording of librarianship and the description of archival materials for many archivists to have been pushed in the direction of trying to use library management systems and techniques in their work. The results have not always been felicitous. Unfortunately in many areas this lack of 'fit' between needs and technical resources has been forced on archivists who have been required to use systems already acquired for library purposes in order to amortise the costs.

Library management software (LMS) is, in essence, a special-purpose version of a generalised database system, usually with features of Database Management System (DBMS) and text-retrieval technology. What has produced the mismatch between archival recording and LMS is when archivists have been constrained to use

the same record layouts and formats as librarians, because at this point the interests really do diverge sharply. Where archivists have been able to use LMS with purpose-built archival record structures the results have been generally successful.

The survey conducted during this project bears out this assertion. Very little special-purpose archive management software is in use; and much of the use of IT is at a low-cost level, represented by the extensive use of DOS-based word processing packages to automate the production of otherwise conventional finding aids.

Authority files have been well-established in librarianship since the production of the Sears List of Subject Headings around the turn of the century. As already indicated above, there has been relatively little overt use of authority files for the control of author names and book titles. There has been a considerable amount of covert use of very large catalogue resources as *de facto* authority files in the sense that large (multi-million entry) catalogues such as that of the OCLC information utility, designed as union catalogues⁴ to show the locations of multiple copies of the same works, become unintentional authority files by virtue of their sheer size. Size, even as a function of professional activity, does not guarantee accuracy: when OCLC celebrated the addition of its ten millionth record, analysis showed that around 30% of all author entries were duplicates. Cataloguers would search the files to find whether a particular book already had an entry, would fail to find it (for various reasons) and would therefore create a whole new entry. Finding, and clearing, this duplication was obviously a major task.

There are single archival resources in the UK with overall collections containing many millions of items, and for them their own records act as authority files to the extent that such things are required. There is no joint resource of the OCLC type representing the combined viewpoints of a number of institutions in agreeing to accept common forms of description. (This must be understood to be a totally different issue to that of accepting common formats to contain descriptions.)

Archivists, by the nature of their work, are concerned with descriptions drawn from actual documents, and recorded in their context. This naturally places a lesser emphasis on the use of authority files to establish how a name (for example) *ought* to be recorded: clearly it should be recorded the way it appears in the original. There is a tension here between the local description, working within the local situation, and the national description such as might be used to bring together research at a national level. Inevitably this raises the question of whether a NNAF offers enough utility to the profession to be worth the effort. The consideration of this point is in the Conclusions and recommendations section of this report.

5 SURVEY OF REPOSITORIES

The summary data extracted from the responses are attached as an Appendix.

This section presents the conclusions to be drawn from the analysis of the survey.

The initial survey form was sent out to a pilot run of 32 names and addresses. These were chosen equally from the National Register of Archive's (NRA) records of the Greater London Archive Network (GLAN), and Joint Information Systems Strategy (JISC) members. After modification resulting from helpful comments received from the pilot names, and from the project Steering Group, the modified survey form was sent to a further 238 names and addresses. These were selected by the NCA to cover all areas of archive work.

The pilot survey yielded 30 responses, and the final version 102 to produce a total of 132 returns. The differences between the pilot and final versions were not very great, and therefore the returns from both exercises could be combined and jointly analysed.

With a relatively small sample to work from, it would not be appropriate to attempt to draw conclusions based on statistical significance. The comments made against each section analysed; and the conclusions drawn; are thus subjective and qualitative. There may be further useful cross-correlations to be made between different answers.

The following comments and conclusions have been extracted from the survey results.

5.1 Technical capability

While most repositories have some IT capability, the bulk of the equipment reported is relatively old (at least in computer/PC terms). The pattern is of a mix of stand-alone and networked machines, often running incompatible software that prevents any effective data transfer between machines, even within a single institution. Stand-alone PCs reported number 115, while 86 networks were reported. Thus about two-thirds of the respondents have access to networked systems and stand-alone systems.

Approximately 45% of PCs are fitted with 386 or lower specification processors. This means that these machines cannot run Windows applications since they lack the processor power, memory size and hard disk capacity for this.

The bulk of the networked systems are not used to carry archival records or management systems, but are the networks of parent institutions (for instance local authorities). In a number of instances these networks provide access to library management systems which are used with varying degrees of success to manage archival records.

The predominant software application on all IT equipment reported is the word processor (WP). Stand-alone PCs account for 92 users of a total of 12 different software packages; networked PCs account for 82 users of a total of 9 different software packages. From the responses, amplified by visits and interviews, it is evident that one of the principal uses of IT in repositories is to automate the production of otherwise conventional finding aids. Typically this is by using a word processor to hold the text of lists and descriptions. Instead of printing out this text, it

can then be searched using the character-string 'search' function of the WP. In some instances the finding aids are printed as well, for use in public search rooms and the like. WP files are not generally made available for searching other than by staff, since it would be possible for users to amend or delete text.

The form of IT application described above is adequate when its products are confined to the originating body. However as a means for data exchange it is a dead end. The 'flat' WP file generally has no structure in the sense of fields to hold the various types of data associated with a record. The only structure discernible in a WP file is generally the range of input dates covered, or the particular collection dealt with, or some similar grouping. The main problem that this causes is that names (of whatever type) are simply embedded in the narrative and descriptive text: they cannot be readily extracted as entities. The consequences of this for the creation of a NNAF are considered in Section 9 (Construction and maintenance of a NNAF). Briefly the concern is that there are only two ways in which names can be extracted from such a 'flat' file: the text is scanned by eye and names are noted; or the text is loaded in to a database system from which an alphabetic list of words can be drawn, out of which list the names are picked by eye. Either method is extremely time-consuming.

5.2 Finding aids

This question was designed to provide some feeling for the likely size of an NNAF - the order of magnitude, if not the actual size in numbers. The question provoked more enquiries and discussion than any other in the survey. Despite this, the results were very useful since large numbers of respondents had no problem understanding the question and providing good answers.

Names and functions of finding aids One hundred and seventy six different finding aids (many representing hundreds or even thousands of specific entities such as the Public Record Office (PRO) Class Lists) are identified. Many of these are highly specific, but a significant number are generic. The following table summarises the results:

Class of finding aid	Numbers reported
Accessions register	5
Catalogues	14
Handlists	7
Indexes	10
Library catalogues	4
Lists	18
Subject indexes	31
Corporate names indexes	2
Personal names indexes	36
Place names indexes	28

All of the remaining specific named finding aids are assumed to fall in to one or more of the above categories.

Physical forms and methods of production

Forty-three different types are listed, with 740 applications between them. The most notable are: Card indexes (142); Manuscript (84); Typescript (121). Of all the forms, 275 are largely computer based, meaning that a computer is used in some way during the production process. The largest single group within this set is the 95 word processor applications.

Volumes

The numbers given in volume calculations are to be taken as indicative only. In many cases no effective count could be made: under guidance from the consultant respondents were asked to estimate numbers from samples, and then gross up.

The total sum of entries in all recorded aids is 23.3 million, with just 6 institutions reporting more than 1 million entries each. These six institutions account for just over 40% of the total of entries.

Rates of addition were frequently estimated in the same way as were the total numbers of entries. It would appear from the total figure reported (1.4 million) that either there are phenomenal growth rates represented here, or that the volumes have been greatly over-estimated.

Figures for index entry deletions (2,362) and amendments (21,917) are both quite low, probably indicating that for most institutions the emphasis is on recording and indexing new material, rather than re-visiting entries already created.

The real totals for the UK as a whole are, of course, unknowable. Best estimates are that the total number of personal and family names that might be contributed if all archives were able to supply their entries would be about 50 million, the number of corporate names around 25 million, and the number of place names around 5 million. These estimates are based on off-the-cuff estimates arising during the interviews, and represent a formidable challenge to data collection and maintenance, to say nothing of storage and searching.

In reality a number of things will conspire to keep the numbers down:

1. Not all repositories will be able to, or will wish to contribute any or all of their entries.
2. Some natural level of prominence in relation to personal and family names will be found. The Public Record Office for Northern Ireland (PRONI) experience of 200,000 personal names yielding 10,000 entries (i.e. 5%) for the Prominent Persons Index is a useful benchmark.
3. The entries for the 6 major national institutions are already included in the calculation of 2.3 million (about 40% of the total) and it is highly unlikely that all the rest of the repositories in the UK put together could equal or exceed this figure.

Whatever the final figures are, their sheer size already rules out a number of technologies for storage and dissemination on the grounds of a mis-match of capacity against requirement; or by reason of the maintenance workload involved.

Attitudes to sharing records

In general the attitudes to sharing records are positive, in the sense that archivists are willing to contribute records from their own systems to those of others', or to union or shared resources. What is less clear is the extent to which archivists would

be willing (or able) to adapt their own finding aids in order to conform to common intellectual standards. (This discussion is concerned with the actual content of finding aids; not the formats used to contain the content).

It seems to be in the nature of archival work, and is certainly reflected in the profession's internal view of its purpose, that description from the top level down is something that is very particular to the repository, and that finding aids are about the collection and its view on the world. There are thoroughly justifiable reasons why this should be so, given the nature of the material that is handled.

Local difficulties arise when, over time, different standards or systems of description are applied to the same body of material. The difficulties are magnified if it is to merge in some way the finding aids from two or more repositories. At this point it is not simply a matter of inconsistency of description that is the stumbling-block, but the fact that each finding aid is an expression of a very specific purpose, and the union finding aid will not have that same purpose.

A parallel problem has been found in many bibliographic and text-based applications, where attempts have been made to construct common thesauri for what appear to be similar organisations. A case in point is health care where the printed *DHSS Data* thesaurus has been applied, with little success, to the subject indexing of health-care collections outside the Department. The 'view', represented by the structure of the terminological relationships, was all wrong for them, and for many such attempted applications the thesaurus no longer related to 'their' world.

Only 9 organisations in the survey claimed to be supplying their finding aids to union resources, although 104 organisations supplied them to others. This seems to be more in line with the way archivists work: share finding aids, but do not create combined ones. One way of trying to understand the reasons for this imbalance is to ask whether, if the resources were available, archivists would already have created such combined finding aids. This point was introduced during the interviews in various ways. It is notable that none of the archivists questioned had the creation of union finding aids on a 'wish list', or mentioned it as a desirable objective. There are obvious implications here for the form and function of a NNAF: these are dealt with in Section 7.

5.3 Authority control

Thesaurus management systems (closely linked in concept to authority files) are used by 8 respondents. Three of these are dedicated stand-alone systems; the other 5 are modules of library management software packages.

The full tabulation of responses is presented in the Appendix. This shows the types of authority accepted for the four classes of name (personal, family, place and corporate) against five possible sources of authority (published works, documents, in-house lists, in-house rules and external rules).

In almost every case the use of internal, or in-house rules outweighs the use of external rules by about 2:1. The largest single group applying external rules, taken from all respondents, is the 29 who use them for the spelling of personal names. The smallest single such group is the 21 which use external rules for alternative forms of place names.

The forms in which rules are held is overwhelmingly manual. Cards (176), hard copy (17), manuscript (1), sheaf binder (3) and typescript (16) give a total of 213. Computer based forms covering databases (9), online (9), printout (47, whether from

WP or other software) and screen-based (23) account for a total of 88. This is a ratio of 2.5:1.

In general very little use is made of external rules or thesauri. This has obvious implications for the work that might have to be done to standardise and normalise input received from these sources.

6. INTERVIEWS AND VISITS

6.1 Major national institutions

Public Record Office (PRO)

The Public Record Office (PRO) houses the national archives of England and the United Kingdom: that is, records created by the actions of central government and of the Courts of Law of England and Wales. The PRO's vision for the 21st century, described within the *Archives Direct 2001* Programme, aims to provide an automated catalogue of all the Office's holdings (PROCat), and to make on-line finding aids available to the public via the World Wide Web by the year 2001.

The principal output format for PROCat is intended to be Encoded Archival Description (EAD), a Document Type Definition (DTD) within the Standard General Markup Language (SGML). A pilot for PROCat, the Core Executive Project - so called because it incorporates those records of central 'Whitehall' departments - aims to anticipate editorial issues arising in the development of the automated catalogue, and to test both the utility of EAD and the software for its support. This pilot will be available to the public in 1998 initially from a website at the University of Virginia, USA, and later in the year from the PRO's own website.

From the outset, it has been recognised that an authority file, composed of authority records relevant to persons, places, corporate bodies and subject terms, is essential to PROCat, to guarantee efficient searching and data exchange. To this end, a pilot authority file will accompany the Core Executive Project and will facilitate the searching of those classes on the World Wide Web. The authority records composing this file will be consistent with the *NCA Rules*. and will not be incompatible with ISAAR (CPF).

As PROCat develops, authority records comprising the PRO's authority file will be created to accompany the online catalogue. This authority file will provide the PRO's contribution to the National Name Authority File (NNAF). Although the PRO's contribution to the NNAF will therefore, be some time in development, it will be substantial.

Scottish Record Office (SRO)

The SRO is the national archive for Scotland. Automation has moved in the direction of free-text searching of narrative records, which provides for relatively simple access to unstructured records. This approach allows searchable databases to be built quite quickly on the STATUS systems used. The drawback, at least as far as the NNAF project is concerned, is that the data elements are not fielded. This will make the identification and extraction of individual names of all types extremely complicated.

Most classes of finding aids are automated: Government records (of Government institutions in Scotland), private papers (500 large and 4,000 small collections), plans (100,000), NRAS surveys (4,000). There are also indexes to testaments (wills).

The SRO is a major player (and the instigator of) the Scottish Archive Network SCAN. Other SRO resources are discussed in the section on SCAN.

Public Record Office of Northern Ireland (PRONI)

From March 1998, PRONI will fully implement its Public Record Office Management System (PROMS), a multi-user on-line system which will revolutionise all aspects of its archival administration. However, any computer system is only as good as the data held on it and, from the beginning, PRONI has sought to quality assure information before capture, including that from the manual indexes.

One of the most used indexes is that for personal names, which consists of 200,000-plus entries. Since most census records for the island of Ireland and many other vital records identifying individuals have not survived, PRONI sought over the years to fill the gap by indexing documents which gave any personal names. A recent development is the creation of a Prominent Persons Index (PPI), a selection of 'the great and the good' from among the general personal names index (about 15 per cent of the entries fell into the new category). The opportunity has been taken to correct errors, conflate multiple references into single ones and insert epithets. Presently at around 4,500 entries, this relatively small database will soon increase markedly with the inclusion of figures of importance in trade and manufacturing. All the work has been undertaken in conformity with the NCA Rules

Corporate names are located in two indexes, personal and subject. The former was combed during the PPI exercise and relevant cards extracted to form a separate index. The next stage will be to check the extracted entries against those in the subject index so as to create a comprehensive corporate names index. There is a relatively small number of very large corporate bodies, and a large number of much smaller ones. This means that basic information on corporate bodies is often difficult to trace. Nonetheless, it would be PRONI's intention to create a corporate names database which conforms to the NCA Rules.

A major element of PROMS is the geographical names index, which consists predominantly of townland names (the townland is the smallest geographical unit in Ireland). The published Topographical Index for Northern Ireland (1961) was used as a quality assurance control for the data capture, and all variant entries were checked (e.g., those for townlands prior to the Ordnance Survey's standardisation of names in the 1830s).

The current PROMS-related concern is with the transfer of word-processed catalogues to the database. Completion of the transfer will mean online access to 12 gigabytes of descriptive information. The logical next stage is the setting up of links between index entries (subject, personal, etc.) and relevant catalogue entries, thus allowing the user to move effortlessly from one searching-aid to the other.

A NNAF is seen as a very important tool for researchers and enquirers, that really acquires value when archive location information is built in. The observation is made that enquirers are little concerned with the niceties of the archival discipline (provenance, etc.) but essentially want an answer to their query which gives a complete picture of the resources available. These resources will frequently extend beyond the confines of any one institution.

PRONI has developed a Web site which contains basic information (opening hours, etc.) but also a great deal of archival information (e.g., introductions to archives of national and international importance). The emphasis is on making the site user-friendly and attractive to the eye, and the latter means the inclusion of prints and other illustrations.

There is no equivalent in Northern Ireland of the Welsh and Scottish Archives Networks since PRONI provides an integrated archival service for the whole of this part of the country and for both private and official records. However, PRONI is eager

that its records should be linked extensively with those of other institutions, thus allowing off-site enquirers access to information across the whole range of institutions.

National Library of Wales Department of Manuscripts and Records (NLW)

The Department of Manuscripts and Records, in collaboration with one of its two sister departments, is adopting the CAIRS TMS package already in use elsewhere in the NLW, and has been using the accessions module of this since December 1996. The cataloguing module, which is being developed to conform with ISAD(G) and EAD, will include facilities for the construction of authority files and thesauri. A corporate strategy on authority files and thesauri, which aims to meet the requirements of both librarians and archivists and to achieve the consistent application of access points throughout NLW, is taking shape.

The Department intends to contribute to the NLW authority files and thesauri in accordance with emerging standards such as the NCA Rules and ISAAR(CPF), and is interested in collaborating on a NNAF.

With retrospective conversion in mind, the Department is engaged in transferring its archival descriptions from paper to disk. It is recognised that retrospective conversion will be necessary because, unlike the library applications, the use of archival material does not tail off over time.

Card and printed indexes to the Department's holdings are estimated to include about 30,000 personal names, 2,000 family names, 10,000 place names, and 2,500 corporate names. Other indexes to such particular areas of the holdings as probate records, marriage bonds, manorial records, criminal records, and poetical works, which contain many thousands of personal names, place names and subject terms, are already searchable in-house by computer.

National Library of Scotland Department of Manuscripts (NLS)

Early attempts at IT in the NLS were concerned with automating the production of traditional finding aids. Over recent years it has become appreciated that IT offers possibilities in its own right, and increasingly the thinking is moving in the direction of new forms of finding aid - principally electronic, and delivered over the Web.

There will only be one more volume of the printed catalogue produced, and then all further cataloguing will be delivered electronically. There are 8 volumes at present, the current 9th volume will be produced for WP files which have been coded for print format and layout. There is sufficient material extant for a further 6 or 7 printed volumes, but these will probably not be printed, and instead the entries will be made available online.

The contents of the printed book-form catalogues have been available online as TelNet files for some years, but it is now recognised the Web access is both better, and inevitable. The speed and availability of the Web, combined with user expectations of its use, are directing this development.

The general layout of entries follows the British Library pattern. Standardisation of names is routine, based on the alphabetic index as the authority. The index as a whole contains many hundreds of thousands (perhaps millions) of names. Most have epithets, some have lifetime dates. There are no subject terms apart from a few generics (e.g. music, writs, diaries). The range and diversity of collections represented means that many names occur infrequently, and it would be unhelpful not to standardise.

Searchers expect to be able to search for material independent of its location, which indicates a need for institutions to share their resources at the level of access to indexes and finding aids.

The printed text of the Catalogue of Manuscripts has been digitised and is now being edited. The main problem is that in the digitisation all cues to names, epithets etc., (such as emboldening, italics, brackets) have been lost and have to be re-entered. This has created difficulties in trying to extract names automatically. Extraction for electronic purposes may be based on the MARC format, but SGML/HTML may prove to be more useful.

British Library Department of Western Manuscripts (BL)

The Department has been cataloguing material for over 200 years. The published catalogues provide a rich and substantial source of information about its holdings in reasonably consistent format, though they present a number of problems in the varying standards and levels of description and depth of indexing over that period; and in the use of internal rules and conventions which have only been written down recently and in retrospect. Since 1986, current cataloguing has been entered on computer, although hard copy catalogues are still produced for publication. Converting over 30 older printed catalogues to automated form is a formidable project in progress (about 50% completed by the end of 1977).

The material catalogued includes illuminated manuscripts, maps, charters, seals, papyri, single volumes and collections of literary and historical papers and music. Cataloguing requires narrative description at volume level, with a structured layout for collections reflecting the arrangement and hierarchical links, supplemented by detailed indexing at piece (i.e. leaf or folio) level.

This type of catalogue has proved unsuitable for combining with bibliographic records for printed materials and it is not proposed to make it conform to the 'core record' in the British Library's new corporate bibliographic system. An independent online enquiry system with Internet compatibility is being developed specifically for Manuscripts data as part of the Retrospective Conversion Project and it may be possible at a later date to give access via the Library's 'single gateway' OPAC.

An index heading generally consists of personal, family, corporate, government or geographical name, appropriately coded. Subject indexing is at present discretionary and selective. The fullest form of a name is preferred, not necessarily the form found in the manuscript. This means in practice including all forenames which can be discovered, with the latest or highest title or office as epithet and cross-references for change of name. The form of name chosen for the index is generally reflected in the narrative description, although it may be shortened if repeated. The current indexing is held on an *Advanced Revelation* database. The authority file contains around 58,000 entries. For cataloguers, an exact-match search on a keyword displays all occurrences of the name. A cataloguing manual describes the principles and methods for direct data entry.

Epithets or descriptive adjuncts are considered essential to identify and distinguish names and are supplied wherever possible, a practice also followed by other national libraries. In the case of personal names, the Department does not generally indicate lifespan dates - this is because in piece level indexing there are a large number of obscure individuals for who such information cannot easily be discovered, and because in a hard-copy catalogue the heading is always seen in conjunction with the body of the entry which must contain a date.

Future exploitation of the catalogues will depend upon conversion of the existing information to machine-readable form. Once converted, new records and old will have to be run together. At present new index entries are added immediately to the database but are checked weekly by the Name Authority Manager. From early 1998 work will begin on merging the two sets of data. This will be examined, a preferred form selected for each, and variants removed. Much preparatory work has already been done in the amalgamated Index of Manuscripts, published in the mid-1980s, where entries for the same person are bracketed with the preferred heading placed first, and this will itself serve as authority.

In future IT might provide new functions to assist searching: Soundex ('sounds-like') name matching and fuzzy searching have already been explored; enhanced cross-referencing, scrolling pick-lists and hypertext links all promise benefits. One of the greatest advantages of automation is the ability to perform keyword searches on narrative text. It will obviously be important to ensure that in the transfer from manual to machine technologies there are gains in functionality, with nothing lost of the richness of the original.

6.2 Scottish and Welsh Archive Networks

6.2.1 Scottish Archive Network (SCAN)

The Scottish Archive Network will be formed contingent on a successful application for National funding. At the time of writing (December 1997) this application was still pending. The following represents the intended shape and content of the SCAN.

The major components of the network will be:

1. Top-level finding aids of every Scottish archive repository available electronically in ISAD(G) format, searchable with text retrieval software.
2. Hypertext links from these finding aids to the appropriate archive, to item/bundle level.
3. Suitable hardware and software and training to be provided to archives that will need them in order to put their material on the network.

The network will carry a variety of tools and services in addition to the top-level finding aids: standard responses to frequently asked questions (FAQs), initially those of the SRO, but other archives will be encouraged to contribute their own; 'How to' guides to subjects such as tracing family trees; exhibition material digitised for wider distribution and local use; bookshop and archival publications; a 'coffee machine' password-moderated discussion group; gazetteer of Scottish archives, local history groups etc.; 'What's New'.

There are also plans to offer certain classes of document (such as wills) in digitised form.

The philosophy of SCAN is to link collections and their finding aids rather than try to present complete indexes and catalogues.

6.2.1 Welsh Archive Network (WAN)

There are 15 local Record Offices, 3 National repositories (including the NLW) and 2 Universities in the network - the Welsh Archives Council (ACW) membership. A Scoping Study (*Cyngor Archifau Cymru/Archives Council Wales: A Welsh archives network* (November 1997)) on national archives has just been completed, which

included an assessment of IT hardware and software in use. A Web site has been established by NLW for the ACW, containing contact and directory information on the various repositories.

The medium-term intention is that the WAN should provide, in accordance with recognised standards of archival description, high-level pointers to national resources.

6.3 Other institutions

In addition to the major national institutions a number of institutions, archival and otherwise, were included in the interview process. It would be invidious to identify and then summarise the viewpoint of each. Instead, the comments received have been grouped under topics. The following organisations were interviewed and, in most cases, visited in connection with the project. These institutions were chosen for two main reasons: they were thought to be particularly representative of their type; and the nature of their response indicated that there were areas of work where more detailed exploration of their experience or view would be of value to the project.

Ceredigion Archive
Cornwall Record Office
East Sussex Record Office
Gloucestershire Record Office
Institution of Electrical Engineers
Museum of London
Post Office Archives
Royal Air Force Museum

6.3.1 Construction of finding aids

It has been noted elsewhere in this report that the largest group of IT applications represented in the survey related to the use of WP for the production of 'flat file' conventional finding aids. The interviews emphasised this conclusion. Very few of the indexes mentioned are in machine-readable form: the commonest forms are index cards, either typescript or hand-written, and frequently with hand-written annotations.

One County Record Office (CRO) is seeking funding for a pilot project to computerise the records of a local family and estate, and the Quarter Sessions archives. The CRO faces the prospect of re-keying all of the material, which is presently in WP 'flat file' format. The result of this pilot will be a useful indicator of how likely it is that a Record Office (RO) could generate name entries from a fielded database for submission to a NNAF. The resource to be built during the pilot will deal with only a few hundred names, will be designed to be delivered over the Web, and will offer different levels of search and access for different types of use and user.

The survey indicates that there is only a small number of repositories likely to be able to produce machine-readable files of name indexes. Even for these it would be necessary to carry out work on the field structures; and many data elements might be missing, to be filled in later.

A number of repositories have placed their finding aids on to microforms, with indexes on the same media, or in index card form.

6.3.2 Indexing of names

The main interest is in local people, since, particularly in ROs, anything up to 60% of all enquiries may be for family history, and enquirers are wanting to follow up people

with regional interests. In most cases the research will begin with place names (suggesting that many archivists are covert geographers) and expanding to take in the family names associated with places. For national collections the focus is on personal and family names, with place being dealt with in an epithet.

The indexing of all types of name is complicated by their variants over time. It has been suggested that a personal/family name index might be inflated four-fold or more by the inclusion of all variant forms and the cross-references between them. In the case of place names in one RO, there are ten or more variant spellings for each place name. Some County place names can have twenty or more variant spellings: some personal names from the 1600s can have ten or more variants.

The ancient and modern civil parish names are part of this problem, along with the legacy names of merged, split and taken-over corporate names.

Many name indexes have been created for very particular purposes. Thus in Cornwall the mine maps are indexed according to the place names derived from a comprehensive capture of all names on the relevant 1906 Ordnance Survey 6" maps.

Local complications of place names include the very common prefixes of Llan- (Wales) and St (Cornwall), which would each produce large and unwieldy clumpings in the relevant indexes if included for each instance.

Museums and learned and professional societies do not have the same interest in place names, or at least not to the same degree. Here the interest is likely to be in identifying a single prominent individual, perhaps with the objective of establishing an authoritative version of a family name, or of honours and titles. This calls for a true NNAF, where epithets and lifetime dates are of considerable importance. University Class Lists and enrolment records are frequently consulted in connection with this kind of work.

6.3.3 Submission of name entries to a NNAF

The most immediate question is: What to submit? Would a NNAF include every name, such as every tenant farmer and miscreant from Court records, or would it only include prominent people? The example of the PRONI PPI should be noted here, where 200,000 name entries yielded 10,000 names of Prominent Persons. Based on the same proportions, one CRO with around half a million names in its personal name index would yield around 25,000 name entries. Most ROs could provide some local equivalent of the PPI for their areas. Criteria for inclusion have been suggested: landowners; families with holdings in more than one County, or overseas; people of achievement.

Scanning or digitising technology can cope with a great deal of variation in forms of input, but index cards are notoriously difficult to capture in this way. The experience of converting library card catalogues to machine-readable form has been that re-keying has often worked out slightly cheaper than scanning. A number of the firms specialising in conversion routinely send bulk shipments of catalogue overseas for keying.

It is often necessary to edit entries in indexes and catalogues before they are scanned or re-keyed, and also to mark the entries to indicate field structures to be used. These are important activities but they can add greatly to the cost. For this reason they are equally often not carried out in the hope that the editing and housekeeping will be made simpler once the data are loaded in to the computer. This hope usually turns out to be ill-founded, resulting in poor-quality data to which are added the inevitable errors that occur in the conversion process.

6.3.4 Potential use of a NNAF

There is little direct interest in a NNAF as a source on how to catalogue or index a name. There is however considerable interest in verifying the existence of a name, along with any associated titles honours and epithets. This defines a NNAF as being more a source of reference than a source of authority. If collection location information could be added in to enrich the entries, then an extremely useful finding aid would have been produced.

The use of a NNAF to direct cataloguing and indexing practice might well happen if the resource were to be provided, but does not have warm enough support to indicate this as a primary reason for creating such an authority file.

The heavy CRO interest in local names, both place and personal, is unlikely to be satisfied by a NNAF, being at too fine a level of detail. This implies that the bulk of the traffic with such institutions is likely to be one-way - enquiries made of the database, but little contributed to it. Where such institutions would find great value in a NNAF would be in tracing the collections relating to families whose land-holdings cross county, or country borders. In this connection some CROs have as much as 20% of their enquiry traffic originating abroad.

7. FORM AND FUNCTION OF A NNAF

Evidence collected from the survey and during the interviews and visits does not provide a strong case for a NNAF as a passive look-up tool. There is evidence of such a requirement, but not enough to indicate a justification on its own for the effort and investment required.

The justification for a NNAF only begins to become apparent when an active system is considered. The differences between the two approaches are discussed below.

7.1 Passive file

A passive file would be for reference only. Entries could be consulted, but users could have no interaction with it. It would be expected to be internally consistent, so that place names in personal name entries would be drawn from, and included in, the place name entries themselves. It is assumed that the forms of entry would conform to the NCA *Rules*.

Such a file would offer the possibility of searching on any one field, and on one or more fields in combination. The technical means for handling such a file are suggested in Section 9.

It is clear from the survey and interviews that different sorts or repository have quite different kinds of primary needs: place, then person for ROs; person for identification in a learned society. This indicates that for the RO type of use there should be considerable attention paid to capturing place information; while for identification purposes dates, qualifications, honours and titles are likely to be more important. Neither type of use is paramount, and either may be applicable to either situation.

If a passive file contains pointers to the collections represented, this introduces a further type of use, and one which is a more compelling reason for building a NNAF. Pointers to collections turn the file in to a research tool for use by archivists and enquirers alike. In the examples below this possibility is indicated by {Collections}.

7.1.1 Personal / family names

A passive personal / family name authority file would have a relatively simple structure, such as:

- Surname
See preferred form, from variant
See *also*, e.g. "Smythe see *also* Smith"
Family name
Forenames
Titles, honours, qualifications
Epithet
Life-time dates
Notes
{Collections}

and the majority of entries would be expected to have data for all or most of these fields..

7.1.2 Corporate names

Corporate name entries would have the following main components

- Organisation title
See preferred form, from variant
See *also* predecessor, successor
Epithet
Operational dates
Place of business or operation
Notes
{Collections}

7.1.3 Place names

Place names would be recorded with the following structure:

- Place
See preferred form, from variant, including bi-lingual preference
See *also* predecessor, successor
Epithet (e.g. OS grid reference, County name)
Notes
{Collections}

7.2 Active file

An active file would contain all of the above attributes. The active component is the {Collections} field. This would ideally be a hypertext 'hot' link to another resource outside the NNAF file. The link would be likely to be implemented in Web technology. A user would see a list of the collections holding material related to the name in question, and would be able to switch directly to the site(s) concerned.

This does not mean that pointers should be made only to sites that are accessible over the Web: three possibilities are available.

1. the link would lead straight in to a Web-mounted finding aid at the site;
2. the link would lead to a passive directory-like Web site to get access information;
3. there would be no link, only the indication that a relevant collection was available.

The nature of each type of link could be shown by the Web conventions of highlighted text, bold text and plain text respectively.

Such a form of a NNAF would be a major justification for its creation since it would introduce a complete new class of use, and one which is likely to be of interest to archivists and users equally. The 'hot' links need not be introduced from the start, but could be brought in as sites became available.

The active file approach presupposes the use of Web technology to deliver the NNAF.

8. MEANS OF DELIVERY AND ACCESS

There is a variety of means by which a NNAF can be delivered to users. Each has distinct advantages and disadvantages, and each is suited to a particular set of intentions that may underlie the development of such a national finding aid. There is no obvious hierarchy (i.e. from best to worst) that could be produced to grade these options. The following table lists the principal considerations of each. In the nature of such things, some of the considerations 'for' might be seen as being equally 'against' in certain circumstances; and *vice versa*. The considerations have therefore been given, as far as it is possible, from the point of view of the producer of a NNAF. Records are assumed to be relatively simple and to be of around 250 characters each, plus the contents of 'notes' fields.

	Considerations for	Considerations against
Index cards	<ul style="list-style-type: none"> a) Simple technology. b) Cards can be computer-produced, ready sorted for filing. c) 24-hour by 7-day availability. 	<ul style="list-style-type: none"> a) Very high maintenance overheads. b) High production costs. c) Very high storage overheads. d) High distribution costs. e) Access cannot be provided to remote enquirers. f) Low functionality. g) Dynamic content changes cannot easily be shown.
Printed hard-copy	<ul style="list-style-type: none"> a) Simple technology. b) Copy can be computer generated. c) Can be spin-off (alternative form) from a database system. d) 24-hour by 7-day availability. 	<ul style="list-style-type: none"> a) Low functionality. b) High production costs. c) High maintenance overheads. d) High distribution costs. e) Updating while maintaining single-sequence integrity is difficult. f) Dynamic content changes cannot be easily shown.
Floppy disk	<ul style="list-style-type: none"> a) Readily-accessible technology with a PC. b) Disk contents can be added to local databases. c) Relatively low production and distribution costs. d) Data can be loaded locally in to a variety of software environments. e) Simple method of delivering updates for local merging with previous data. 	<ul style="list-style-type: none"> a) Capacity. Low data volume per disk means distribution on a very large number of disks. b) Users will probably have to load the data in to their own retrieval systems, with variable functionality.
CD-ROM	<ul style="list-style-type: none"> a) Costs of mastering and production are low. b) Costs of distribution are low. c) The data are not particularly volatile (i.e. most additions to the database will be of new 	<ul style="list-style-type: none"> a) Capacity. One CD-ROM might be expected to hold between 1.5 and 2 million records, plus the indexes and simple retrieval software. This could mean around 15 - 20 CD-ROMS in the full set

	material) and this suits the CD-ROM as a publishing medium. d) CD-ROMS can be networked on larger installations.	b) The majority of potential users do not have CD-ROM drives: those that do are unlikely to have the multi-platter juke-boxes required to access large numbers of discs.
Local database	a) Database records can be provided on floppy disk, CD-ROM, or by downloading. b) Existing database software can be used.	a) Each repository will have to fund its own computer software and storage requirement. b) Support is difficult with no uniformity of software package.
Online database	a) Updating is limited to the master database alone. b) All users have access to the same level of data. c) No local storage requirements. d) Results can be downloaded for local processing. e) Common enquiry structure simplifies training and support. f) Access can be direct without the need for users to go through a service provider.	a) Minimum requirement is computer and telecommunications technology to gain any access. b) Online databases are more suited to complex retrieval needs.
Web access	a) Any Web browser can be used for access. b) Data can be made active, with 'hot' links from entries to other sites offering data or related information. c) Database resource can be delivered alongside related (e.g. directory-type) information.	a) Minimum requirement is high-level computer and telecommunications technology plus <i>Windows</i> to gain any access. b) Access must be through an Internet Service Provider (ISP). c) Web browsers do not offer particularly sophisticated search facilities.

9. CONSTRUCTION AND MAINTENANCE OF A NNAF

There are no direct equivalents to a NNAF that can be used to indicate the likely scales of size, growth rates, maintenance overheads, distribution and, hence, cost. At best the examples of a few related instances can be used to illustrate the possible magnitude of the operation.

9.1 Examples

9.1.1 The Post Office

The Post Office maintain the Postcode Address File (PAF). This is managed and run by a small department in the Post Office, located at Portsmouth. The PAF links the three elements of town, street and street number to the fourth element of the postcode. Postcodes code for between 15 and 20 separate road numbers each. The level of detail of the PAF is generally to the postal town (e.g. Portsmouth), but with further 'dependent localities' such as local area names where some further discrimination is required. Thus a long road passing through two or more distinct areas would be divided by these dependent localities. The facility is also used to discriminate between different roads with the same name in the same area.

The PAF contains details of 26 million delivery points, and around 1.6 million separate postcodes. The data from it are available in a variety of forms - printed, disk or tape, labels etc.

It would be possible to purchase the town/dependent locality data for loading in to a NNAF. This would probably answer for many purposes relating to present names, although it has not been established whether the PAF has the right level of detail. Obviously the PAF does not cover historical names, or historical versions of present ones.

Place names corresponding to the civil parishes used by ROs are not available from the PAF.

9.1.2 Government of Canada

Extremely useful information from Canada on the maintenance of the Canadian Geographical Names Data Base (CGNDB) came to light during the research for this project. While not directly comparable to the UK environment, this does illustrate the extent of maintenance work required of a major national resource.

The CGNDB contains about 500 000 records, maintained on an *Oracle* data base. Each record may contain some 30 or so fields of attribute information, in addition to the correct spelling of the name. As well, there are satellite database files containing such information as titles for the National Topographic System (NTS) map sheets (approx. 15,000); other names data to be included on maps; data on Canadian World War II casualties, etc., and paper copy graphic materials (perhaps another 30,000 digital records and correspondence files for the 15 000 NTS sheet files mentioned above).

To look after the national committee, acquisition of data from the provinces, maintaining and developing the national database, maintaining a Web site, providing information to the public, providing names for mapping, equipment upgrades, salaries, building and administrative overheads, supporting upper echelons of management etc. takes something in the order of C\$1.2 million a year.

Each province and territory is the gathering ground for the names that end up on the national database, and their costs are not included above - anything from 0.5 of a person to 26 people (Quebec) on staff. It should be added too that much work still remains to be done in the collection and recording of data in Canada, with large areas in the north where much is still in oral tradition, and a dynamic situation elsewhere in the country for a variety of reasons (e.g. for administrative regrouping; use of Aboriginal names, naming of smaller features, etc.) This is also a contrast with a relatively static situation in the UK.

Names data are used as an authority file (for example, as a framework for those who by law must file for land use permits). The department involved acquires updates on a yearly basis, and so their only cost is data purchase (a few hundred dollars), as the maintenance is done in any case.

9.2 Data capture

There are, as in other areas for consideration, several options available. Broadly these divide in to where, how and by whom data capture and conversion might be done; and where the resulting system might be hosted.

There are two separate problems embedded in the consideration of data capture. The first is the conversion to a common record standard and compliance with NCA *Rules* of data already in machine-readable form. The second is the capture and conversion in machine-readable form of data at present in hard-copy only. Examples of the routes available for data capture are shown below.

Hard-copy formats

Present form of data	Possible conversion routes
a) Non-fielded narrative text such as lists, descriptions and other finding aids where the individual entries have no implicit structure readily recognised by machine scanning other than breaks between records or entries.	Route 1: Scan or digitise text, recognising breaks between records. On screen manually tag data elements, or transfer them to a structured format such as a database. Route 2: Mark-up all records to indicate the data elements, then re-key them.
b) Fielded entries such as typescript indexes, index cards and other finding aids where some implicit structure exists that could be recognised by machine scanning	Route 1: Scan or digitise records using OCR techniques that can recognise the implicit structure, such as the difference between columns of names, and columns of associated reference numbers; or the placing of names, dates etc. on index cards. Route 2: Re-key records where the print quality of the material is too variable, or the format is inconsistent.

Machine-readable formats

c) 'Flat file' (e.g. WP) formats that are machine equivalents of a) above, with no implicit structure within records.

Route 1: Examine records by eye, inserting standard markers to indicate the different data elements, or physically moving elements in to a standard sequence. Process the file to extract the marked-up elements for transfer to a standard database format. It may be necessary to create ASCII files from the WP ones first, to make the file text accessible in a form that can be uploaded to a database.

Route 2: Print the records, and then re-key from the print-out in the appropriate form and sequence.

d) Fielded records where different types of data are consistently held as separate items. This includes WP files where the different data elements are marked by the use of brackets, emboldening, italics and other recognisable markers.

Route 1: Use software format conversion utilities to extract data elements from text, or to pick them up in the correct sequence, and upload to a database format.

The above routes are concerned only with the physical capture and transfer of data elements from various formats in to a database format. These routes do not address the need for intellectual effort (logical conversion) in editing or cleaning the data, converting data to a standard such as the NCA *Rules* or ensuring conformity with authority files.

9.2.1 Data capture centrally or locally?

Centralised data capture and conversion offers the opportunity to enforce common standards of format and content. However it dissociates the process from the collections which have given rise to the entries. This raises the distinct possibility of meaning being lost in the conversion through lack of contact with the source material. Central data capture and conversion loads all of the costs on to one authority.

Local data capture and conversion places the task adjacent to the source material, increasing the likelihood that the logical conversion will be done in a proper relationship to it. The major drawback is that the task will represent, for most repositories, a whole new and additional area of work for which they may not have the time or the staff and machine resources. In some cases repositories are already engaged in, or are contemplating the conversion of finding aids to computer systems. They may therefore be able to produce fielded, consistent data with little additional effort.

It is important to understand that many repositories do not already have the full range of name indexes as proposed for the NNAF not because they have no resources to build them, but because they are not needed. Name indexes (such as place names) have been created where there are particular needs for just that type of index. Creation of entries for a NNAF may therefore be an additional workload for many archivists. The NNAF must be able to offer value in return for the effort.

On balance the most workable arrangement would seem to be:

1. Specify physical (format layout) and logical (content standard) requirements centrally. Specify the nature of the machine-readable record to be generated. Invite local conformity among participants with these.
2. Encourage local data conversion and capture. The SCAN model proposed, of providing equipment, software and training in order to encourage participation, is a good one to follow.
3. Monitor the quality of data supplied at the centre, and apply appropriate corrective action.
4. Offer back local data sets, extracted from the NNAF as a whole, as part of a package that provides benefits to both parties.

Experience in other areas (for instance the local collection of training opportunities information for the creation of the TAP databases by TECs and LECs) has shown that the biggest problem is in getting conformity with the logical standards, and in monitoring and applying overall quality control.

9.2.2 Editing and reformatting

The work could be carried out by the HMC, or contracted out to an organisation such as the Information Services Department of a University, or to a commercial organisation that has experience in the electronic publishing environment.

Scanning or digitising text is probably best left to a specialist commercial organisation (such as SAZTEC) that has the intelligent OCR/digitising software and machinery required to carry out most of the tasks identified in the tabulation in Section 9.3.1 above. This activity would produce a database load file in the appropriate format.

9.3 System management

The management of the system covers the policy and administrative aspects of: terms and conditions for data collection; enquiry and access; update and maintenance; legal issues of data ownership and use.

If a NNAF is built as an active file (Section 7.2) then it will, in time, take on many of the functions of the NRA. It will act as the central UK collecting organisation for pointers to collections. It appears logical therefore that the NRA should take the lead in developing, implementing and delivering the NNAF.

9.4 Host site

Two different areas of database maintenance have to be considered: manipulating, reformatting and indexing the data; and making the database available as a searchable resource. These two do not need to take place on the same machine, or be carried out by the same organisations. Indeed, there are very good technical and operational reasons why they should be separated: separation acts as a 'firewall' to prevent the unauthorised or unintended amendment of entries; and it prevents conflicts of service between the requirements for maintenance, and requirements for access. Further, it allows a database system suited to maintenance and update to be used for that part of the operation, and one best suited to support enquiry and research to be used for that purpose.

9.4.1 Database access

The database can be delivered as either an online system, or as a Web site. The first instance will probably require that the database be hosted by a specialist bureau that

can provide the 7-day by 24-hour access expected, and handle the system maintenance.

A Web site could be hosted at a site such as the HMC's own. Equally it could be placed with a University for delivery over JANET, or placed with a commercial third party.

JANET It is questionable how much longer the JANET network will continue to be free at the point of use, which would considerably alter the pricing. Given that the large institutions (which are on JANET) are likely to be net providers of data; and the smaller repositories (which are not able to access JANET) are likely to be net users of the data, this route appears to offer little operationally. The network is also heavily loaded, with priority being given to intra-university data traffic.

University IS Departments It is also questionable whether University IS Departments would have the capacity to mount and run such a resource; or be able to make a long-term commitment to its continued support. Data volumes indicate that an IS department would have to make what amounts to a commercial charge for such a significant application.

Commercial third party The volume considerations alone indicate that it might be better to locate the database with a specialist host such as Head Software International, where it would be one of a number of Web databases made accessible to either closed or open user groups. A company such as this could also offer the 6gb - 8gb disk capacity likely to be needed, with all data conversion and 'crunching' required, and bespoke browser software suited to the application.

An indicative cost for this sort of operation would be: Set-up on a Windows NT server at an Information Service Provider (ISP) £5,000 Annual running costs of a dedicated machine (annual) £15,000 Annual update costs (say monthly updating) £7,200

Placing the database with a third party would leave the database administrators free to manage data quality and the administrative and managerial problems associated with data conversion, leaving the technical considerations of mounting and delivering the database for routine access by users.

The most suitable database structure is likely to be a relational one, to manage the cross-linking of data elements such as place names used in epithets. A relational DBMS (RDBMS) would also be of great value when managing collection location information, since this can be considered as dynamic - the manual effort of tracing all similar locations when a collection is moved, or the repository changes its name, is not trivial and an RDBMS removes the need for it entirely.

9.5 Data protection and intellectual property

These two aspects of database construction and use must be considered: the operation cannot legally be undertaken unless they are resolved.

9.5.1 Data protection

If the NNAF contains references to living persons then the Data Protection Act and EU Data Protection Directive have effect. In any event references to living persons will have to be cleared with the individuals concerned, who have an absolute right of correction or amendment. It will have to be established whether this is sought by the contributing institutions or by the organisation that assumes the management of the NNAF.

9.5.2 Intellectual property

The EU Database Directive and the UK Copyright and Rights in Databases Regulations 1997 will have to be observed in respect of the ownership of the database; and the use and exploitation of the records (these are separate legal issues). There will have to be a contractual relationship between the data providers and the database managers.

10 CONCLUSIONS AND RECOMMENDATIONS

The conclusions and recommendations noted here have not been weighted or prioritised in any way. This would not be possible, given their interdependence.

10.1 Conclusions

1. There is not enough support for a simple passive reference NNAF to justify its creation.
2. There is good support for a NNAF that contains active location information, able to point to sites that carry the source material and, if possible, link users directly to their finding aids.
3. Web access appears to be the most likely technology to deliver the functionality required, and the most likely to be able to cope with the data volumes envisaged.
4. The majority of repositories below national level are unlikely to be able to supply much data in machine-readable form, but are likely to be the main users of a NNAF.
5. Most finding aids have been developed in response to strictly local needs. A considerable amount of data conversion effort will be required to convert formats (physical conversion) and data elements (logical conversion).
6. Most repositories have finding aids that are not readily converted to machine-readable form, implying a large volume of material to be converted by labour-intensive means.
7. The technical resources available to many repositories do not at present allow for Web access by the majority of them. However the cascade effect of hardware and software upgrading should ensure that by the time a NNAF is a reality, most repositories should have the basic means of access to access Web sites.
8. Data protection and intellectual property implications are considerable, and require as much planning as the technical aspects of creating a NNAF.

10.2 Recommendations

1. An active NNAF should be created, mounted on a Web site, and offering 'hot' link access to the Web sites of other repositories.
2. The level of detail in the NNAF should be that indicated in Section 7.1. Personal names for inclusion should be chosen on the basis of prominent people, selected to criteria to be established by agreement with participating organisations. Place names should be chosen to the level of townland, township or civil parish. Corporate names should be chosen to criteria functionally equivalent to those for prominent persons (i.e. of more than purely local significance).
3. Established standards for record format and description should be used.

4. The NNAF database should be managed centrally to ensure data consistency through monitoring and corrective procedures. The collection and conversion of data should be carried out locally where possible.
5. Conversion of certain types of data (for example typescript) may most efficiently and cost-effectively be carried out under centralised direction, by a commercial conversion organisation.
6. The NNAF should be hosted by a commercial organisation that is also able to handle the data formatting and database uploading required, and provide an assured level of Web access.
7. Some incentives may be required in order to encourage repositories to collaborate with the NNAF, especially where this activity is right outside their normal planned workload.
8. The form of management of the NNAF, and the operational location of it, require careful consideration. The NRA should take the lead in developing, implementing and delivering the NNAF.
9. Work is required to establish the cost/benefits of the various regimes under which the NNAF might be funded and run.

11. BIBLIOGRAPHY

The following documents were made available for consultation during the project:

ICA: *Dictionary of archival terminology - English version*

British Film Institute: *Guidelines for managing an cataloguing written archives relating to the moving image*

Cygnor Archifau Cymru / Archives Council Wales: *A Welsh archives network*

Report to the NCA IT standards working party from the Name Authority project:
Standardisation of name authority controls

NCA: *Rules for the construction of personal, place and corporate names*

East Sussex County Council Libraries and Records Committee Member's Seminar

Nov 96: *Archives in your livingroom? The past, present and future of new technology at East Sussex Record Office*

BL: *Authority Control : an automated authority file at the British Library*

BL Manuscripts Collections: *Automated cataloguing: a manual*

APPENDIX: SUMMARY DATA FROM RESPONSES TO QUESTIONNAIRE

The following comments and conclusions have been extracted from the survey results. Machine names, publication titles etc. are mostly recorded as supplied by respondents.

Q1: Status of archival services

6 bodies listed themselves as autonomous archival institutions. Of the remainder, 111 names of parent institution were supplied.

Q2: Funding of repositories

There were 121 sources of funding listed. In a number of cases repositories have funding from more than one source, so there is a small amount of overlap in the numbers: Higher Education sector - 29; Local Authority - 71; National - 14; Independent - 15. All 8 of the 'Other' category are in fact independent in some sense.

Q3: Summaries of holdings

These summaries were often extensive, providing a flavour of each collection. They helped the analysts to understand the context of many answers, but could not in themselves be analysed or easily represented. The original responses will be made available for further analysis.

Q4: Contact names and addresses

These details have been used by the analysts to develop the list of visits and interviews. Most respondents provided contact details.

Technical matters

Stand-alone PCs

Q5(a): Hardware used

115 stand-alone PCs were listed, covering 45 different manufacturers. The responses are biased towards 386-processor machines (approximately 45%) as the single largest group. There is a scattering of much older equipment at the low end of the range; and a slightly larger group of very current high-specification machines at the top end. The remainder are quite recent 486-generation machines.

Q5(b): Operating systems

The largest group by far is MS-DOS with 71 users. Insufficient numbers were able to report the version of MS-DOS used for any particular conclusion to be drawn. Windows 3.1 and 3.11 account for 45 users between them; Windows 95 for 2; and UNIX for 1.

Q5(c): Memory sizes

5 machines have below 1mb of memory; 21 between 1mb and 5mb; 42 between 8mb and 16mb; and 6 greater than 16mb. This pattern is consistent with the figures for models of PC, and Operating Systems.

Q5(d): Hard drive sizes

43 respondents have hard drives of less than 499mb (half gigabyte) capacity. 11 have drives of between 500mb and 999mb. 16 have drives of capacities greater than 1 gigabyte.

Q5(e): Accessories

CD-ROM drives are fitted to 29 of the machines listed; and 16 have modems.

The implications of Q5(a) to (e) are:

1. Most users are at least two generations behind current software, and for at least 60% of users access to the Internet using the equipment listed is difficult or impossible, other than for simple e-mail.
2. The technical capacity of most equipment is quite low.
3. This pattern is likely to persist with archive bodies continuing to have the bulk of their equipment investment in older, slower and lower-capacity machines (the same trend is evident in Local Authority and University libraries). This could restrict the Web-based delivery of a NNAF unless it is delivered at quite a low technical level.

Q6: Software in use

Word processors

There are 92 users of WP, spread over 12 different packages. The largest groups are WordPerfect (in a wide variety of versions, mostly for MS-DOS); and Word.

Database Management Systems

There are 49 users of DBMS, spread over 18 different packages. There is no obvious single largest group or type.

Electronic Mail

There are 27 users of e-mail, spread over 12 different packages.

Spreadsheets

There are 17 users of spreadsheets, spread over 4 different packages.

Records Management Systems

There are 8 users of RMS, spread over 6 different packages.

Networked PCs and workstations

Q7: Operating Systems

There are 80 network OSs listed, covering 7 different types. The most popular are Novell Netware and various versions of *Windows*.

Q8(a): Network type

There was considerable confusion here between the network type (e.g. LAN, Ethernet) and the OS as recorded in Q7. Little useful information can be gained other than a general impression that most respondents are not familiar with their technical environment.

Q8(b): Network size

86 respondents answered, recording a total of 1,192 workstations. Although the question asked for only the number of workstations in the archive area, some

respondents limited this to the number of terminals dedicated to staff use, whilst others included public-access machines.

48 respondents had between 1 and 9 workstations on their networks; 25 had between 10 and 99 workstations; and 3 had 100+.

Q9: Workstation make and model

69 different machines are listed, covering 31 different manufacturers. Because of the lack of information in Q8(b) above, no conclusion can be drawn as to the actual numbers of any one make and model in use.

Q10: Software on networked PCs

Word processors

There are 82 users of WP on networks, spread over 9 different WP packages.

Database Management Systems

There are 23 users of DBMS, spread 23 different packages.

Electronic mail

There are 56 users of e-mail, spread over 22 different packages.

Spreadsheets

There are 54 users of spreadsheets, spread over 7 different packages.

Specialist software on networks

Q11(a): Records Management Systems

Software packages

There are 22 users of RM software, spread over 14 different packages.

Modules used

11 different module types are listed. Only Accessions and Cataloguing with 3 users each, stand out.

Q11(b): Library Management Systems

Software packages

There are 38 users of LMS software, spread over 20 different packages.

Modules used

16 different modules are used by a total of 49 users. Significant modules in use are Cataloguing and Circulation (6 each), and OPAC with 7.

Q11(c): IR/Text Management software

Software packages

There are 40 users of IR/TM software, spread over 20 different packages.

Modules used

12 different modules are used by a total of 25 users. Significant modules in use are Accessioning (7) and Thesaurus (5).

Q12: Thesaurus management

Only 3 users of thesaurus management systems are listed. (These are in addition to the 5 users of thesaurus modules listed under Q11(c) above.)

Q13: Network access

Access over a network to CD-ROM drives is available to 47 users; to the Internet to 53 users; and to external online services to 36 users.

Q14: Electronic access/communication of information

Fax is available to 86 respondents, of whom 4 use fax modules on computers. Floppy disk transfer is apparently available to only 86 (which probably reflects a lack of awareness; or need for this form of transfer). E-mail is available to 72; CD-ROM to 49; Internet connection to 59; and online services to 37.

These figures cannot be compared directly with other figures in this survey, since the question was specifically about the use of these technologies for information access and transfer.

Q15: Hardware and software support

Hardware support is available to 99 computer users of all types; and software support to 93.

Q16: Finding aids

This question was designed to provide some feeling for the likely size of an NNAF - the order of magnitude, if not the actual size in numbers. The question provoked more enquiries and discussion than any other in the survey. Despite this, the results were very useful since large numbers of respondents had no problem understanding the question and providing good answers.

Names and functions of finding aids

176 different types of finding aids are identified. Many of these are highly specific, but a significant number are generic. Of the latter general types, the notable ones are: Accessions Registers (5); Catalogues (14); Handlists (7); Indexes (10); Library catalogues (4); Lists (18); Subject indexes (31). Of particular note are the Corporate names indexes (2); Personal names indexes (36); Place names indexes (28). The PRO Class Lists (15,000) plus non-standard finding aids (400) do not readily fit in to any of the categories noted above.

All of the specific named finding aids are assumed to fall in to one or more of the above categories.

Physical forms and methods of production

43 different types are listed, with 740 applications between them. The most notable are: Card indexes (142); Manuscript (84); Typescript (121). Of all the forms, 275 are largely computer based. The largest single group within this set is the 95 word processor applications.

The implication of these answers is that in the main IT and computer-based systems are largely devoted to automating the means of production of otherwise conventional finding aids. Relatively little effort is being put in to using computer systems in themselves as the finding aid. It is quite likely, for instance, that the high use of word processing is mostly attributable to producing lists on a WP, without any fielding or categorisation of the information elements as would happen with a database application. If this is true, then the extraction of name elements from the files will be extremely labour-intensive.

Sizes of finding aids

This question was only asked in the pilot, and was subsequently dropped as being too confusing. The responses noted below are therefore from just 30 respondents. The question asked was “size of this finding aid as an approximate percentage of all such aids”. No conclusions are drawn here based on replies to this question.

Total of index entries

225 different total sizes were provided. The total sum of entries in all recorded aids is 23,332,615 entries. Only 6 institutions reported sizes greater than 1,000,000; and their number amount to 10,400,00, or just over 40% of the total.

Many respondents could not answer this question; several provided estimates based on entries per sheet grossed up by the estimated number of entries, etc.

Number of index entries added per year

166 different additions-per-year figures were provided. The total sum of all recorded additions is 1,403,972. 28 finding aids were identified as closed, having no further entries added.

Number of index entries deleted per year

140 different deletions-per-year figures were provided. The total sum of all recorded deletions is 2362. 120 finding aids are identified as having no deletions made from them.

Number of index entries amended per year

97 different amendments-per-year figures were provided. The total sum of all recorded amendments is 21,917. 26 finding aids are identified as having no amendments made to them.

Q17: Standards, guidelines and authorities for name control

Area of control	Published works	Documents	In-house lists	In-house rules	External rules
personal names - spelling		41			
personal names - alternatives	57	32	51	57	29
family names - spelling	43	36	42	45	26
family names - alternatives	51	32	46	49	25
place names - spelling	43	34	41	43	23
place names - alternatives	64	29	59	59	22
corporate names - spelling	51	38	47	48	21
corporate names - alternatives	42	31	40	42	25
	35		33	37	23

The notable figures here are that the use of in-house rules outweighs the use of external rules by about 2:1.

Physical forms of standards, authorities and guidelines

Listed below are all of the methods of production reported, together with the numbers of instances of each.

Index cards	176
Database	9
Hard-copy	17
Manuscript	1
Online	9
Print-out	47
Screen-based	23
Sheaf binder	3
Typescript	16

It is notable that of the 301 instances above, at least two-thirds are manually prepared. This poses a considerable problem for any conversion of these authority tools.

Q18: Sharing finding aids of and with other organisations

Use of finding aids of other archives

A total of 73 organisations use the finding aids of other institutions. The total number of finding aids used is in excess of 240.

Supply of finding aids to others

104 organisations supply finding aids to others. The total number of finding aids supplied is in excess of 192.

Supply of entries to union catalogues

9 organisations supply their entries to union catalogues or similar shared resources.

Other organisations similarly supplying aids to union catalogues

21 organisations are known to be also contributing to union catalogues.

Q19: Standards used within the archive

Source of funding/ Standards used	HE sector	Local authority	National	Independent
ISAD(G)	16	12	2	5
ISAAR(CPF)	3	3	1	1
NCA Rules	9	12	3	2
MAD2	11	21	2	6
AACR2	12	5	2	5
Other, local, in-house	APPM2 (2) In-	GLRO (1) In-house (7) Local	Art & Architecture Thesaurus (1)	In-house (2) NAI British

house (6)	Moray District	Local NLW Dept	Library (1)
In-house	RO (1) MAD2	records standards	Social History
ISAD(G)	modified (1)	(1) In-house (5)	Industrial
based (1)	Oxford	PRO Manual of	Classification
NAL	Dictionary of	Records	(1)
British	Christian Names	Administration (1)	
Library (1)	(1) SRO Guide	Thesaurus of	
	to Cataloguing	Geographic	
	(1) SRO Manual	Names (1)	
	of Cataloguing		
	(2)		

Q20: Amount of staff time to maintain finding aids

A total of 78 institutions supplied information. The total reported amount of staff time for maintenance is 11,023.5 hours per month; grossed up to 18,897 days per year - an average of 141 hours per month, or 242 days per year.

Many institutions had no idea of the amount of staff time given over to maintenance, or else could not separate out this activity from other staff activities.

A small number of organisations (including the PRO with 19 people having a significant part to play in list creation and amendment) have a clear idea of the staff consequences of maintenance.

Future developments Respondents were asked to best-guess what developments might take place within the next 2 to 4 years, and to confine the answers to activities not currently undertaken..

Q21: Co-operation

Source of funding/Form of co-operation likely	HE sector	Local authority	National	Independent
Share/contribute to a Web site	17	44	8	8
Contribute to shared or union finding aids	15	25	8	5
Use records from other's systems	11	21	6	6

Q22: Technical developments

14 institutions that do not yet have a computer are likely to install one.

32 institutions are likely to be installing new software to manage finding aids. 13 of these responses concern generic software, such as *Oracle*. The remainder list specific packages.

Web sites are likely to be set up by 39 organisations that do not currently have them. 63 organisations that either have Web sites or are likely to set them up, are also likely to provide public access to finding aids through the sites.

Q23: NNAF

The answers to this question (concerned with the use of a NNAF) contrast strongly with the answers to Question 21(b) which was concerned with contributing entries to a shared or union finding aid. 48 respondents said they would be likely to contribute; while 95 say they are likely to contribute to a NNAF. There is no obvious reason for this almost 100% disparity.

95 institutions would use a NNAF to aid in the cataloguing and listing of their collections; 90 would use it to aid the searching of their collections; and 81 would use it as an authority, and offer modifications and amendments to it.

Q24: Technical delivery of a NNAF

Respondents were asked to indicate various technical means by which they would be able to access a NNAF now; and within 2-4 years. Certain aspects of this question were designed to cross-check the consistency of answers. Some disparities are evident when the answers are compared with the questions relating to, for example, CD-ROM drives installed. The total for stand-alone plus networked CD-ROM access is higher than the number of respondents who say they are able to accept CD-ROMs now.

	Form of access of availability	
	Available now	Available in 2 - 4 years
Floppy disk (diskette)	104	18
CD-ROM	70	35
Online	45	34
Web	60	44
E-mail	72	36
Microform	97	13
Other	none	none