# 06 Arranging and cataloguing digital and hybrid archives

## Introduction

Cataloguing marks an important stage in the lifecycle of a digital object: it is the point at which final appraisal decisions are made (see Chapter 04 *Appraisal and disposal*); records which are to be subject to continued closures and access restrictions are identified; the digital and paper elements of a hybrid archive are intellectually united; and rich contextual information is provided by the archivist, enabling the researcher to make sense of the collection and assess its relevance to their field of research. Descriptive cataloguing is likely to take place a considerable amount of time after ingest to the repository because of the lengthy closures necessitated by copyright and other legislation (see Chapter 09 *Legal issues*). However, where archives contain only a small proportion of digital material which is subject to closure, cataloguing may take place at an earlier point, and it is important to record at least some descriptive metadata soon after an archive has been accessioned.

Paradigm worked with exemplar collections of politicians' hybrid personal archives in an effort to discover what arrangement and description challenges these new materials might raise. This chapter provides an introduction to the cataloguing standards employed by the project team - principally Encoded Archival Description (EAD) and the *General International Standard Archival Description*, 2nd edition (ISAD(G) 2) – and explores how these standards might be applied to digital and hybrid personal archives. Some basic models and templates for arrangement and cataloguing are proposed, but Paradigm did not attempt to catalogue any exemplar collection in its entirety and it is therefore likely that there are more practical issues that will only be uncovered in light of more detailed experience.

## Standards for archival description

The use of standards by the archive profession arose from the increasing use of computerised systems in the 1980s. Automation demands consistency, and consistency demands standards. Archivists, unlike their library colleagues, felt for many years that the 'uniqueness' of their holdings exempted them from standardisation. Large-scale automation projects challenged this notion. It was soon recognised that inconsistency in archival descriptive practices was hampering both in-house and collaborative initiatives. The widespread adoption of standards would enable the virtual reunification of archival fonds, which had been scattered between archives, or even between countries, in union lists and other finding aids. Standardised data would also facilitate metadata sharing with colleagues in the closely related museum and library worlds.

Standards detail how to structure archival descriptions as well as prescribing preferred forms to be used when indexing (see p. 172) place, person and corporate names. The key standard governing the content of archival description is ISAD(G), the *General International Standard Archival Description* (see p. 170).

### Project standards

Paradigm used the in-house cataloguing manuals produced by the Bodleian Library and the John Rylands University Library as a starting point for thinking about descriptive practices for hybrid archives. Both institutions support the use of international standards to ensure the production of high quality digitial finding aids, which can be cross searched using standardised access points for subjects, places and names. Paradigm adopted the following descriptive cataloguing standards:

- ISAD(G)2: *General Standards Archival Description* 2nd Edition.
- EAD: Encoded Archival Description.
- LCSH: Library of Congress Subject Headings.
- NCA Rules: National Council on archives, *Rules for the Construction of Personal, Place and Corporate Names* (1997).

## ISAD(G) 2

ISAD(G) was formally adopted by the International Council on Archives' Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999 and published in 2000. The standard provides general guidance for the preparation of archival descriptions. It can be used in conjunction with existing national standards. The purpose of the archival description is to identify and explain the context and content of the archival material in order to promote its accessibility. This is achieved by creating accurate and appropriate representations and by organising them in accordance with predetermined models. The rules may be applied irrespective of the form or medium of the archival material.

### Multi-level description

ISAD(G) advocates hierarchical multi-level description and recommends four descriptive principles to achieve this:

- **Describe from the general to the specific**: at the highest level the archivist should provide information which pertains to the archive as a whole; descriptions at lower levels should give information specific to that level. The end-result should be a hierarchical description which represents the relationships within the archive.
- **The information provided should be relevant to the level of description**: the archivist should provide information which relates to the level of description. For example, a biographical note about the creator of a series of letters received by an individual would be included in the description of that series, not at the highest level or at a lower one.
- **Descriptions should be linked**: the purpose of linking descriptions, or expressing the level of description, is to allow users to determine the context of an item within the archive as a whole.
- **Information should not be repeated**: by providing information relevant to subordinate levels of description at the highest possible level the archivist can avoid redundant description.

### Elements of description

In addition to these overarching principles, ISAD(G) comprises 7 areas of descriptive information:

1. **Identity statement area,** where essential information is conveyed to identify the unit of description.
2. **Context area,** where information is conveyed about the origin and custody of the unit of description.
3. **Content and structure area**, where information is conveyed about the subject matter and arrangement of the unit of description.
4. **Conditions of access and use area**, where information is conveyed about the availability of the unit of description.
5. **Allied materials area**, where information is conveyed about materials having an important relationship to the unit of description.
6. **Notes area**, where specialised information and information that cannot be accommodated in any of the other areas may be conveyed.

7. **Description control area**, where information is conveyed about how, when and by whom the archival description was prepared.

The key structural fields of an archival description proscribed by ISAD(G):

**Identity statement area**
3.1.1 Reference code(s)
3.1.2 Title
3.1.3 Date(s)
3.1.4 Level of Description
3.1.5 Extent and medium of the unit of description (quantity, bulk or size)

**Content and structure area**
3.3.1 Scope and content
3.3.2 Appraisal, destruction and scheduling information
3.3.3 Accruals
3.3.4 System of Arrangement Conditions of access and use area
3.4.1 Conditions governing access
3.4.2 Conditions governing reproduction
3.4.3 Languages/scripts of material
3.4.4 Physical characteristics and technical requirements
3.4.5 Finding aids Allied materials area
3.5.1 Existence and location of originals
3.5.2 Existence and location of copies
3.5.3 Related unites of description
3.5.4 Publication note

**Context area**
3.2.1 Name of Creator(s)
3.2.2 Administrative/Biographical history
3.2.3 Archival History
3.2.4 Immediate sources of acquisition or transfer

**Notes area**
3.6.1 Note

**Description control area**
3.7.1 Archivist's note
3.7.2 Rules or conventions
3.7.3 Date(s) of descriptions

For a full description of each of these elements illustrated with worked examples see Appendix B of the ISAD(G)2 standard.[1]

### Access points and ISAD(G)

There are provisions in ISAD(G) for 'access points' for creators and subjects, and it was recognised that the content of these access points needed to be controlled if they were to function effectively. The use of standard terms for name and subject indexing facilitates consistency in the exchange of data between repositories, and in the retrieval of data by remote, online users. In accordance with practice at the Bodleian and the Rylands, Paradigm used the NCA rules for creating indexing terms for names and places, and the LCSH index for subject terms (see p. 172).

## EAD

EAD is a widely adopted standard for encoding archival finding aids and is used by both the Bodleian and Rylands libraries to catalogue their archival holdings. It provides a means for archivists to structure finding aids using technology that is independent of proprietary hardware and software platforms and which can be delivered via the Internet. The structure of EAD is modelled upon ISAD(G) (see p. 170). It therefore reflects the hierarchical, multi-level nature of archives and archival descriptions.

EAD was originally devised between 1993 and 1995 at the University of California, Berkeley, but is now maintained by the US Library of Congress in partnership with the Society of American Archivists. EAD was initially based on the Standard Generalized Markup Language (SGML), but with the subsequent development of Extensible Markup Language (XML) it has been made fully XML compliant. It is widely used both in the United States and in the United Kingdom.

1 International Council on Archives, Committee on Descriptive Standards, *General International Standard Archival Description*, Second Edition (2000). URL: <http://www.ica.org/biblio.php?pdocid=1>

The EAD standard is represented as an XML document type definition (DTD) or XML Schema. It can be obtained free of charge from the EAD official website, which also provides background information on EAD, an overview of its structure and guidelines for its implementation.[1]

### Overview of stucture

EAD contains three high-level elements: <eadheader>, <frontmatter> and <archdesc>. The <eadheader> is a wrapper element which contains information about the finding aid itself, not about the archival materials described in it. <frontmatter> is used to present a title page and prefatory text. The actual archival description is contained in <archdesc> which is thus the heart of EAD. All the descriptive elements are nested within <archdesc>. There are thirteen primary descriptive elements; further subordinate elements are nested within most of them.

The Descriptive Identification <did> element serves as a wrapper for the essential information needed by researchers to determine whether the materials described are relevant to their line of enquiry. It contains the basic building blocks for any level of description, from fonds-level down to item and piece: title, reference code, date of creation, name of creator, physical extent and (recommended at fonds-level only) an abstract of contents.

Following the <did> are elements which contain more detailed information on the archive materials. EAD does not prescribe the order in which these elements should appear, but it is common practice to follow the order of ISAD(G) areas and elements (see p. 171).

#### Example EAD files

Example EAD files are available in the online version of the Workbook.

The Manchester Men's League for Women's Suffrage Archive, John Rylands University Library of Manchester - mml.sgm.

A legacy finding aid converted using the RLG-Apex service: Papers of Clement Richard Attlee, 1st Earl Attlee, 1924-57, Department of Special Collections and Western Manuscripts, Bodleian Library - attlee.xml.

Archives Hub collection level description for the Papers of Clement Richard Attlee, 1st Earl Attlee, 1924-57 - attlee.txt.

## ✧Indexing and authority files

## Common indexing standards for archival description

There are provisions in ISAD(G) for 'access points' for creators and subjects. It was recognised that these access points needed to be controlled if they were to function effectively. The use of standard terms for name and subject indexing facilitates consistency in the exchange of data between repositories, and in the retrieval of data by remote, online users.

### Library of Congress Subject Headings (LSCH)

The Library of Congress Subject Headings thesaurus was originally designed as a controlled vocabulary for representing the subject and form of the books and serials in the Library of Congress collection. Its purpose was to provide subject access points to the bibliographic records contained in the Library of Congress catalogues. In recent years, it has been widely adopted by the UK archival community. The LCSH terms are useful as they can incorporate multiple subdivisions. For example, subject terms can be further divided by geographic locality and chronological period:
Architecture--England--Oxfordshire --17th century

1 Network Development and MARC Standards Office, *Encoded Archival Description (EAD) Version 2002 website.* URL: <http://www.loc.gov/ead/>

There are some drawbacks to the LCSH thesaurus. The headings are sometimes criticised as representing an American-centric worldview, which makes indexing some topics difficult. There are also differences between Anglo-Amercian and UK English spelling conventions. Despite this, the LCSH thesaurus is acknowledged as the most comprehensive subject thesaurus currently available. It also has the advantage of being maintained and updated regularly by the Library Congress.

**Constructing LCSH headings**
For advice on constructing LCSH headings please refer to the *Constructing Library of Congress Subject Headings* section in the online version of the Workbook.[1]

**Examples of LCSH headings**

| Place name: | Subject: | Genre: |
|---|---|---|
| Oxfordshire (England) | Economics--History--17th century | Diaries--19th century |

## National Council of Archives *Rules for the Construction of Personal, Place and Corporate Names*, 1997 (NCA Rules)

The NCA Rules were created in response to a recommendation from The IT Standards Working Party of the National Council on Archives, which, in March 1991, recommended that the national repositories and the Historical Manuscript Commission should 'examine the desirability of standardising name authority controls at national level'. The NCA Rules were developed in consultation with the authors of the *International Standard Archival Authority Record for Corporate Bodies, Persons, and Families ISAAR (CPF)*, 1st edition Beijing, 1996, 2nd edition Canberra, 2004 (see p. 174). This co-operation ensured that the UK National standard could adequately express the idiosyncrasies of British naming conventions whilst conforming to international naming standards.

After much research and consultation, the NCA Rules were published in 1997. The Rules represent a shift away from in-house conventions to nationally agreed standards. The Rules assist cataloguers in forming names for persons, places and corporate bodies which are unique and readily identifiable by users. By ensuring consistency in the structure of proper names (person, family, place and corporate) the Rules facilitate the exchange of data between repositories and greatly enhance data retrieval; for this reason NCA Rules are used by major archival gateway services such as the Archives Hub,[2] AIM25[3] and Genesis: Women's History Sources in the British Isles.[4]

**Constructing names with NCA Rules**
For advice on constructing names using the NCA Rules please refer to the NCA Rules section in the online version of the Workbook.[5]

**Examples of names constructed using the NCA Rules**

| Personal name: | Family name: |
|---|---|
| Bodley \| Sir Thomas \| (1545-1613) \| Knight \| Diplomat and Scholar | Leigh family \| Adlestrop |
| **Corporate name:** | **Place name:** |
| Manchester Men's League for Women's Suffrage | Grimsby\| Lincolnshire\| TA 2709 |

## National Name Authority Files (NNAF)

A name authority record comprises the recognised, authorised or prescribed form of a name, usually supported by sufficient information and sources to ensure reliable recognition and use of such a name. The Historical Manuscript Commission (now part of The National Archives) showed a great deal of interest in creating a database of National Name Authority Files and launched a pilot study assisted by several institutions which created an extensive NNAF database. The evolving NNAF is likely to comply with the ISAAR(CPF) standard and be encoded using the Encoded Archival Context (EAC) standard.[6]

1 Online Paradigm Workbook. URL: <http://www.paradigm.ac.uk/workbook/cataloguing/lcsh.html>
2 Archives Hub, *Archives Hub website.* URL: <http://www.archiveshub.ac.uk/>
3 AIM25, *AIM25 website.* URL: <http://www.aim25.ac.uk/>
4 Genesis, *Genesis website.* URL: <http://www.genesis.ac.uk/>
5 Online Paradigm Workbook. URL: <http://www.paradigm.ac.uk/workbook/cataloguing/ncarules.html>
6 Peter Gillman, *National Name Authority File: Report to the National Council on Archives* (British Library Board, 1998). URL: <http://www.ncaonline.org.uk/materials/nationalnameauthorityfile.pdf>

## ISAAR (CPF)

ISAAR (CPF) provides guidance on creating archival authority records for corporate bodies, persons and families associated with the creation and maintenance of archives. The description of individuals, families and organisations that create records is a key component of the preservation of the documentary evidence of human activity. Identifying record creating entities; recording the names or designations used by and for them; and describing their essential functions, activities and characteristics, and the dates and places they were active is an essential component of the management of archival records. Creator description facilitates both access to and interpretation of records.

Archival authority records may be used to describe these entities; to control the creation and use of access points within an archival description; to document relationships between the entities and the records created by them; and to link to other resources about, or by, them. For an example of an ISAAR (CPF) record see the online version of the Workbook.[1]

*International Standard Archival Authority Record for Corporate Bodies, Persons and Families: ISAAR (CPF)*, 1st edition, Beijing, 1996, 2nd edition, Canberra, 2004 can be downloaded from the website of the International Council on Archives.[2]

## EAC

Encoded Archival Context (EAC) is an ongoing initiative within the international archival community to design and implement a prototype standard based on XML for encoding descriptions of record creators. The proposed metadata standard is intended to comply with ISAAR (CPF) (see above) and complement other standards governing name authority control for personal and corporate entities. The primary developers of this prototype standard are members of the international archival community.

The XML DTD and the Tag Library documentation for EAC was developed in co-operation and with support from the LEAF project (Linking and Exploring Authority Files),[3] which ran from March 2001 to March 2004. The EAC website has further information, an EAC schema and example EAC documents, which are displayed firstly in XML and secondly as they might be viewed by the public.[4]

# A proposal for intellectual access to hybrid archives

Paradigm proposes that multi-level EAD catalogues continue to be a primary interface for discovering and navigating archives, setting the material in context and outlining its structural relationships. The means by which the user will actually access the digital archives themselves will depend on the rights associated with the material.

As shown in the Paradigm lifecycle diagram (see p. 2), a two-stage opening process is envisaged. Archives will first be made accessible in a controlled environment, where the various rights (including privacy and copyright) of donors and third parties can be appropriately managed. Once copyright (and therefore other rights which expire much sooner) has expired, the archive can consider publishing born-digital archives to an online repository where they may be accessed directly by researchers. If the institution is able to provide a secure repository with terminals in a reading room setting, then the researcher should be able to move via a link from the relevant EAD catalogue de-

1 Online Paradigm Workbook. URL: <http://www.paradigm.ac.uk/workbook/cataloguing/isaarcpf.html>
2 International Council on Archives, *ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons, and Families, Second edition*, 2nd Edition (August 2004). URL: <http://www.ica.org/biblio.php?pdocid=144>
3 LEAF Consortium, *LEAF website.* URL: <http://www.crxnet.com/leaf/>
4 Ad Hoc EAC Working Group, *Encoded Archival Context (EAC) website.* URL: <http://www.iath.virginia.edu/eac/>

scription to a specific digital archival object; if not, no link will be present and readers will probably order digital materials for delivery to special stand-alone computers in the reading room. For objects published to an online digital repository, links from the EAD catalogue to born-digital archives will undoubtedly be present, and researchers may also be able to browse and search the born-digital archives independently of the EAD catalogue by interacting with the repository interface(s).

The purpose of the access repository is to make 'presentation versions' of digital archives available to researchers in accessible formats. These presentation versions of the digital archives will be wrapped with the appropriate level of metadata (descriptive, structural and administrative) to make them self-describing. It is likely that the user will encounter this metadata in the form of 'cover sheets' or 'splash pages' for individual digital objects. Such self-describing digital objects essentially form the Dissemination Information Package (DIP) (see p. 189), as defined by the OAIS Model (see p. 3). The DIP differs from the Archival Information Package (AIP) in that it is designed to fulfil the access function rather than the preservation function; whilst the researcher is likely to want a degree of technical information (e.g. an indication of the creator's software environment, the past and present states of the digital object, and an assurance that it is authentic), they will not require the kind of detailed technical information stored as part of the AIP.

The proportion of descriptive metadata is also likely to be higher in the DIP. It is important that the DIP is independently understandable because, whilst the EAD catalogue will supply the final, comprehensive level of descriptive metadata, it is unlikely that digital archivists will be able to produce detailed item-level EAD catalogues for every archive in their care. Additionally, when researchers are able to browse and search born-digital archives via an online repository without mediation through an EAD catalogue, the metadata supplied must enable the researcher to evaluate the material. Whatever the level of detail, it is also important to remember that the archivist will be drawing on the AIP when compiling an EAD catalogue, and that much of the information required for the access function must be supplied at ingest and fully recorded by the archivist compiling the AIP; otherwise it could be lost forever.

## EAD templates for a personal archive

To date the use of EAD for cataloguing born-digital archives is not at an advanced stage of development, so a body of practical experience to draw on is lacking. Paradigm therefore began by thinking about the overall arrangement of the project's exemplar archives and how each hierarchical level would map to a component level (c01, c02, etc) in EAD.

Next, the issue of a collection level description for each archive was addressed: conclusions about which EAD elements to include at this level were made as well as recommendations for the content of these elements. There is not scope at this stage to develop cataloguing templates for every record format encountered in Paradigm's testbed archives and the work was therefore limited to tackling three major areas:

- Most of the exemplar archives acquired from politicians working with Paradigm included folders based on subjects, projects, or record types (e.g. minutes, financial material), and these contain a range of file formats: word-processed documents predominate, but there are also image formats, plain text documents, Portable Document Format (PDF) and spreadsheets. Some recommendations on EAD cataloguing at folder and item level for this 'generic' material are included here.
- Email forms another major component of these archives; as it is unique to the digital environment and poses its own particular challenges, and as the Paradigm Academic Advisory Board identified it as being of particular historical value, the project has suggested some approaches to cataloguing personal email directories – at email directory, folder and item level.

- Paradigm took some snapshots of politicians' websites and weblogs during the 2005 election period. These too were identified as a particularly valuable resource by the Advisory Board, and as they provide various new challenges for archivists, Paradigm also looked at options for cataloguing and providing access to archived websites.

## Further considerations

The design process for an EAD profile should give consideration to the search and display mechanisms that the EAD metadata will be expected to support, and to whether the EAD is written purely for a local implementation or should interoperate with network services such as the Archives Hub. These issues are not specifically addressed here, Paradigm being more concerned with how the content of the catalogue is impacted by the addition of born-digital material to the archive.

## Suggested templates for the arrangement of a personal archive

Chapter 03 *Working with record creators* outlines the kind of material typically encountered in a politician's digital archive. However, every archive is unique, as is every archivist, so there is often more than one equally valid approach to arranging a digital or hybrid personal archive.

The personal archives of some politicians accessioned by Paradigm fell into obvious groupings based on the division between their Westminster and constituency offices, each of which were staffed by different people; each office therefore represents a sub-fonds as defined by ISAD(G): 'a subdivision…containing a body of related records corresponding to administrative subdivisions in the originating agency…or…geographical…functional…or similar groupings of the material'.

In other cases, the project only accessioned constituency papers because the records produced by the politician in their Westminster capacity fall within the Public Records Act (see p. 247). Here, the principal division may be between different record types (e.g. email correspondence) or different members of constituency office staff; for instance, in one example a principal member of staff was responsible for the MP's diary, some constituency casework files, monitoring local issues of interest, maintaining the main email directory and filing system; a second member of staff was responsible for logging casework and general administrative work; and another maintained the politician's website.

Two possible templates for arrangement might look as follows:

Example 1:

| Structure | ISAD(G) level | EAD encoding |
|---|---|---|
| Politician | Fonds | collection level |
| Member of politician's staff | sub-fonds | <c01 level="subgroup"> |
| Folder (e.g. subject folder held on staff member's c: drive) | series | <c02 level="series"> |
| Sub-folder | subseries | <c03 level= "subseries"> |
| Individual files | item | <c04 level="item"> |

Example 2:

| Structure | ISAD(G) level | EAD encoding |
|---|---|---|
| Politician | Fonds | collection level |
| Constituency office | sub-fonds | <c01 level="subgroup"> |
| Email directory/subject folder | series | <c02 level="series"> |
| Email folder/subject sub-folder | subseries | <c03 level= "subseries"> |
| Individual file/email | item | <c04 level="item"> |

These models of arrangement may differ for other politicians and would differ again if the kind of personal archives generated by an individual outside of their official capacity were taken into account.

## Paradigm exemplars for arrangement

### Example A

This exemplar collection consists of two accessions.

**Accession 1:** this accession was made from the personal computer of a personal assistant to the politician based at the constituency office. It includes subject folders and sub-folders, arranged along the following lines:

- Press releases
- Europe
- Election leaflets
- Mailings
- Diary back-up
- My Pictures
- Letters

**Accession 2:** This accession was made from the office computer of a politician's Westminster-based personal assistant. The records were kept in two areas of the PC: the 'My Documents' folder and C:/. There was no apparent logic to this split. In order to retain the original directory, the archivist accessioned the two areas as two separate series. The accession also included select folders of email, arranged by subject.

The original order of the accession was along these lines:

- My Documents
  - Briefings
  - Diary
  - Library research papers
  - Ministerial responses
  - Press release
  - Website
- C drive
  - Answers to written questions
  - General constituents' reports
    - Press releases
  - Parliamentary questions
  - Speeches
    - Council of Europe
    - Debates
    - Oral questions
    - Statements
- Email folders
- Various subject files

These types of records are readily comparable to their paper equivalents and their arrangement seems very straightforward. Where a good record keeping system has been established by the creator, archives are much more logical and accessible, even if filed only to the level of folder. Archivists are also assisted in their arrangement decisions by the capacity of computers to automatically order material in a hierarchical structure, and often to impose alphabetical or other arrangements.

## Suggested arrangements for Example A

Paradigm identified two possible arrangements for this example, based on the in-house cataloguing guidelines of the Bodleian and the Rylands.

1. The Bodleian approach resulted in this arrangement:

Papers of politician (fonds)
Westminster papers (series)
- Subject folders (subseries)
- Email correspondence (subseries)
- Paper correspondence (subseries)
- Diary (subseries)
- Speeches (subseries)

Constituency papers (series)
- Subject folders (subseries)
- Email correspondence (subseries)
- Paper correspondence (subseries)
- Diary (subseries)

**In EAD:**

**<c01 level = "series">**Westminster papers
- **<c02 level = "subseries">**Subject folder A
- **<c02 level = "subseries">**Subject folder B
  - **<c03 level = "item">**Document foo
- **<c02 level = "subseries">**Email Correspondence
  - **<c03 level = "otherlevel" otherlevel="subsubseries">**Email subject folder A
    - **<c04 level = "item">**Email
      - **<c05 level = "piece">**Attachment or enclosure
- **<c02 level = "subseries">**Paper Correspondence
  - **<c03 level = "otherlevel" otherlevel="subsubseries">**Subject folder A
    - **<c04 level = "item">**Letter
      - **<c05 level = "piece">**Enclosure
- **<c02 level = "subseries">**Diary
- **<c02 level = "subseries">**Speeches

**<c01 level = "series">**Constituency papers
- **<c02 level = "subseries">**Subject folder C
- **<c02 level = "subseries">**Subject folder D
- **<c02 level = "subseries">**Email correspondence
  - **<c03 level = "sub-subseries">**Email subject folder A
    - **<c04 level = "item">**Email
      - **<c05 level = "piece">**Attachment
- **<c02 level = "subseries">**Diary

2. The JRUL approach resulted in this arrangement:

Papers of politician (fonds)
Westminster papers (subfonds)
- Subject folders (series)
- Paper correspondence (series)
- Email correspondence (series)
- Diary (series)
- Speeches (series)

Constituency papers (subfonds)
- Subject folders (series)
- Paper correspondence (series)
- Email correspondence (series)
- Diary (series)

**In EAD:**
**<c01 level = "subfonds">**Westminster papers
**<c02 level = "series">**Subject folder A
**<c02 level = "series">**Subject folder B
**<c03 level = "item">**Document foo
**<c02 level = "series">**Email correspondence
**<c03 level = "subseries">**Email folder A
**<c04 level = "item">**Email or letter
**<c05 level = "piece">**Attachment or enclosure
**<c02 level = "series">**Paper correspondence
**<c02 level = "series">**Diary
**<c02 level = "series">**Speeches
**<c01 level = "subfonds">**Constituency papers
**<c02 level = "series">**Subject folder C
**<c02 level = "series">**Subject folder D
**<c02 level = "series">**Email correspondence
**<c03 level = "subseries">**Email subject folder A
**<c04 level = "item">**Email
**<c05 level = "piece">**Attachment
**<c02 level = "subseries">**Paper correspondence
**<c02 level = "subseries">**Diary

### Example B

This example contains three accessions of digital archives from the London office of a politician, accessioned from the personal computers of two assistants. The folders are arranged by brief and then into subject folders. Some work would be required to arrange the three accessions into one structure, but it looks like it would be possible to harmonise the three as they are arranged similarly.

**Suggested arrangement for Example B**
Papers of politician
Papers relating to brief 'X'
Subject folders

**In EAD:**
**<c01 level = "series">**Papers relating to Brief X
**<c02 level = "subseries">**Subject folder A
**<c02 level = "subseries">**Subject folder B
**<c03 level = "item">**Document foo
**<c02 level = "subseries">**Miscellaneous
**<c03 level = "item">**Document
**<c01 level = "series">**Papers relating to Brief Y
**<c02 level = "subseries">**Subject folder C
**<c02 level = "subseries">**Subject folder D
**<c03 level = "file">**(10 "file"s)
**<c02 level = "subseries">**Miscellaneous
**<c03 level = "file">**(2 "file"s)

### Hybrid archives

In either of these examples, paper and digital records would be combined and arranged according to context and content, rather than format. Email and paper correspondence would be arranged in separate series because interleaving the two would offer little benefit for considerable effort.

## Deciding on the level of description required

A collection-level description should be compiled for a digital or hybrid archive as early as possible, both for reasons of appraisal (see Chapter 04 *Appraisal and disposal*) and in order to draw together the various sections of an archive and set them in context; some degree of administrative metadata should also be recorded in the collection-level description, so that the archive can be

managed appropriately with regard to rights and data protection. It may be that archivsts opt to produce collection level descriptions for researchers before deciding on final arrangement and moving down to lower levels of description; this will allow a minimum level of intellectual access and enable researchers to identify material of interest, even if they cannot link directly to it at this stage.

In an ideal world, full item-level descriptions of all archival holdings would be produced; and in the world of hybrid archives it would be useful to drill down to a low level, so that researchers can fully understand an archive and its various components, as well as the kind of formats and record types it contains.

Paradigm's Academic Advisory Board believed that full descriptions will be of most use to researchers. Similarly research carried out by the Archives Hub in 2005 suggests that the two main priorities of their users (primarily information professionals, students and academic researchers) are the addition of further online catalogues to the Hub and having access to fuller, item-level descriptions.

## Arguments in favour of cataloguing digital archives to item level in EAD

- To deliver what users want in an online catalogue.
- To mediate between the researcher and the digital object, providing rich contextual information and ensuring that users fully understand the archival item.
- To distinguish clearly between the hard-copy and digital elements of a hybrid archive, and to draw out the relationships between these elements more fully.
- To enable direct links to be made from the catalogue entry to the digital object it describes, something which was also highlighted as important by the Academic Advisory Board and by users of online archive catalogues.
- Often researchers want to see what an archive contains by browsing the catalogue before having direct access to original material. If archivists only catalogue to folder level, and the link to born-digital material occurs at that level, readers will not benefit from having access to a full catalogue description of the component items.
- Producing item level descriptions provides fuller intellectual control over the archive, which will help to ensure that the archive is administered in a way which protects the interests of those who are represented in it.
- Basic item-level metadata can be extracted automatically, making cataloguing at this level less labour-intensive.

## Arguments against cataloguing digital archives to item level in EAD

- Each digital object will be wrapped in its own METS file, which should contain sufficient automatically extracted metadata to render the object self-describing and independently understandable. Users should therefore be able to view item-level digital objects directly from a higher level in the catalogue and still have access to an adequate level of descriptive metadata about individual items.
- Even in the world of traditional archives, institutions often lack funding and staff time to devote to compiling full item-level descriptions of all their collections. Given that the quantity of material in a digital archive is likely to be much greater than that in a traditional archive, item-level cataloguing is less likely to be feasible.
- In many cases, the historical value of individual records within an archive will not be high enough to warrant item-level description. This may be even more relevant in the digital environment where archives (even after appraisal) are likely to be much larger.
- Full text search technologies will allow researchers to search directly on the content of certain categories of textual digital archival objects; this means that catalogue information can afford to be less detailed.

- Cataloguing the digital component of a hybrid archive to item level but not the hard copy element may lead to an imbalance between the two components, with the hard copy material becoming 'hidden'.

## Linking from the catalogue to digital objects: at what level should this be done?

The ability to link directly from an EAD catalogue to the digital archive material it describes was identified as desirable by the Paradigm Academic Advisory Board. The survey of the Archives Hub users suggested that linking directly to born-digital records was a slightly lower priority for them than having access to item-level descriptions, but it was still considered of some importance. The 2004 Linking Arms survey co-ordinated by the National Council on Archives (which reflects the needs of a slightly different audience, including a high proportion of family and local historians) found that 55% of those who responded expressed a wish to access and/or download content of documents (such as digital images), rather than just catalogue descriptions.

It will only be possible for remote users to link directly to born-digital records in those cases where the digital material is either already in the public domain or copyright permission has been obtained from the rights holder, and any other legislative restrictions have expired. It may also be possible to provide direct access from catalogue to digital archive on a standalone PC in a searchroom if copyright is the only restriction which applies (see Chapter 09 *Legal issues* for more detail on IPR and access restrictions, p. 252).

EAD allows for linking directly from catalogue entries to born-digital archive material held outside the catalogue, by means of two elements or tags: <dao> (Digital Archival Object) and <daogrp> (Digital Archival Object Group). To date, these tags have mostly been used to link to digitised images of the material described in the catalogue, but they are also explicitly recommended (in the Research Libraries Group (RLG) *Best Practice Guidelines for Archival Description*) for creating links to born-digital objects if they are the focus of the finding aid.

Linking to an individual digital object from an item level description should be relatively straightforward, and where item-level descriptions have been created, it is recommended that the link be made from that level. RLG recommend using <daogrp> rather than <dao>, both for consistency throughout a finding aid and because it facilitates linking to more than one digital representation. However, it seems most appropriate to use <dao> if the repository intends to use METS files for disseminating objects; the link could be made to a METS file, which contains links to different representations of the item and further metadata about it. This would reduce the need to update the EAD catalogue as files are migrated to new formats.

The <dao> tag is probably most appropriately placed as the last sub-element within the <did> tag. The <did> is a required wrapper element that bundles together elements identifying core information about the described item, such as reference code, title, date and physical description of the object. If <dao> is placed in this context, the archivist will not have to supply the additional <daodesc> element, which is designed to contain information about the contents, usage or source of the digital archival object when the information in other tags is insufficient to identify the object.

<dao> must contain various attributes in order to: connect the finding aid to the location of the digital object (or DIP); determine how this link is made by the user; and determine how the DIP will be displayed, e.g.:

```
<dao linktype="simple" href="http://shuttle.paradigm.ac.uk:8085/fedora/get/demo:100/
actuate="onrequest" show="new" />
```

In this example, the linktype is simple (because only a single digital object is concerned), the link to the specified location of digital object is made at the request of the user rather than automatically,

and the DIP will be displayed in a new window.

Whilst linking at item level is relatively straightforward, in many instances we may be unable to produce full item-level descriptions. Some solutions might be as follows:

1. Provide a single link, using <dao>, from folder level. Users will be taken to the DIP for that folder; obviously a folder is not a document which can be viewed, so the DIP will only contain metadata about the folder. This will include structural metadata describing the relationships between this and other digital objects; users will therefore be able to identify all constituent files and navigate freely around them (all of which will have a level of descriptive and other metadata to aid understanding).
2. Produce item-level catalogue entries with only a bare minimum of information (possibly automatically extracted) in each (reference code, title and date), with a <dao> link to the digital object.
3. Catalogue to series level only, but embed a list of component digital items in the <scopecontent> at that level and make use of the <daogrp> tag to provide links to these items. <daogrp> is generally used to link to multiple digital representations of the same thing (e.g. a number of digitised images of the item described, each at a different resolution; or images of different pages from the same digitised manuscript); however, there seems no reason why it cannot be employed to link to a sequence of related items which make up a record series. The <daogrp> tag acts as a wrapper for a number of <daoloc> elements, pointing out the location of each object linked to; <daoloc> may contain a <daodesc> element to supply a brief label for the digital object.

Using option 3, a co2 or co3 series (folder) level might look something like this:

```
<scopecontent>
    <p>This folder contains various press releases, as follows:</p>
    <daogrp linktype="extended">
        <daoloc linktype="locator" href="http://shuttle.paradigm.ac.uk:8085/fedora/get/
        demo:1/PDF">
        <daodesc>ABC/1/1 File: "Press release, 1 March 2005"</daodesc></daoloc>
        <daoloc linktype="locator" href="http://shuttle.paradigm.ac.uk:8085/fedora/get/
        demo:2/PDF">
        <daodesc>ABC/1/2 File: "Press release, 4 April 2005"</daodesc></daoloc>
    </daogrp>
</scopecontent>
```

Paradigm has no direct experience of linking digital archive material to an EAD catalogue in this way; ultimately the use of other attributes within <daoloc> should be determined locally, and guidance will be forthcoming in the light of practical experience in this area.

Of course, it would also be useful for users to link from a DIP to its associated catalogue description. This should be possible via the <dmdSec> section of the digital object's METS (see p. 117) document; here the <mdRef> tag allows users to link to external descriptive metadata (in this case an EAD catalogue). By encoding a unique ID within the relevant <c> tag in the EAD catalogue (e.g. <c03 id="abc23" level="item">), the XPTR attribute in the METS <mdRef> can be used in conjunction with the ID value to identify and link to the relevant section of the EAD catalogue. This will be most effective where the decision is made to catalogue fully to item level in EAD, or when taking approach 2 outlined above. However, it will also be possible to link from the METS document for a single digital object to a collection- or series-level EAD description which includes the digital object as a component item.

## Suggested EAD elements required at Fonds level

The importance of the collection-level description is indicated elsewhere in this chapter (see p. 179). Here, we focus on those elements Paradigm considers essential at fonds level when cata-

loguing a digital or hybrid archive. The <eadheader> and <frontmatter> elements are omitted here because they relate to metadata about the finding aid itself and the encoding of prefatory text and title page, so will be repository-specific. The focus is therefore on the information included in the <archdesc>.

## Elements in the <did>

**ID of the Unit <unitid>** A unique reference code forming the basis of the shelfmark for each component level description. This should consist of: a country code based on ISO 3166 *Codes for the Representation of Names of Countries*; a repository code in accordance with the national repository code standard or other unique location identifier; and a specific local reference code or shelfmark.

**Title of the Unit <unittitle>** A title for the archive. Traditionally for an individual, the term 'Papers of [name]' is used; in the digital environment this should perhaps be replaced by 'Archive of [name]'.

**Date of the Unit <unitdate>** Covering dates for the whole archive. Where the information is obtainable (via the preservation metadata stored as part of the AIP), these should be span dates recording the initial creation of the earliest item in the archive, to the last modified date of the latest item in the archive. At collection level this will be a span of years rather than anything more specific. This will be supplied for researchers in traditional form. The 'normalise' attribute will be used to record the date in accordance with ISO 8601 *Representation of Dates and Times*. Embedding this information as an attribute facilitates information retrieval queries based on dates, and both EAD and PREMIS recommend using ISO 8601, e.g.:

**<unitdate type="inclusive" normal="1975/2006">**1975-2006**</unitdate>**

**Origination <origination>** The name(s) of the individual(s) responsible for the creation, accumulation, or assembly of the archive prior to its accession. Traditionally this is limited to the principal creator. However, in the case of politicians, a number of office staff might also be involved in generating a substantial proportion of the archive, so multiple names may need to be recorded here. Names should be supplied in accordance with the NCA Rules (see p. 173) to facilitate information retrieval; a register of authority files should be maintained and the subelements <persname> and <corpname> should be used to encode the names created.

**Physical Description <physdesc><extent>** For digital materials, Paradigm's Academic Advisory Board suggested that researchers would be interested in both 'intellectual' extent (i.e. the number of series, folders and files) of the archive, as well as the size in megabytes, with the former taking priority over the latter. Note that the preservation metadata record is likely to record file sizes in bytes; the EAD catalogue should use megabytes, as a more understandable indication of size for the researcher. For a hybrid archive, information should also be supplied here about the extent of the hard copy material, with an indication of the relative proportions of digital to hard copy in the archive as a whole.

**<materialspec>** This element is used to record data unique to a particular class or form of material which is not assigned to any other element of description. This is probably the most appropriate element in which to record information about the creator's original file formats. At fonds level this should comprise a broad overview, e.g.

**<materialspec>**The bulk of Politician A's digital archive is comprised of files in Microsoft Word for Windows Document 6.0/95 or 97-2002 (.doc). Image formats include both JPEG File Interchange Format 1.01 (.jpg) and Tagged Image File Format (.tiff). Other file formats include Binary Interchange File Format (Microsoft Excel) spreadsheets (.xls) and Portable Document Format 1.4 (.pdf). **</materialspec>**

It is suggested that <materialspec> is used in preference to <genreform> for supplying free-text information on file formats; if <genreform> is used at all, it should only be used in the controlled access section.

**Language of the Material <langmaterial>** Record the language(s) of the material in the archive. If only one language is represented in the material, this information need only be supplied at collection level. If an archive is predominantly in one language but with a proportion of material in another, this should be noted at collection level; <langmaterial> can then be used at lower levels to indicate where the second language is represented.

**Digital Archival Object <dao>** If desired, a link can be made to the METS document which represents the 'collection' in the digital repository and points to the children folders and files of the archive.

## Other elements

**Physical Characteristics and Technical Requirements <phystech>** Used to describe: physical conditions and characteristics which affect the storage, preservation or use of the archive, including the physical composition or hardware and software requirements for the preservation of and access to records held in electronic formats. The archivist may need to record information here which relates to the hard copy component of a hybrid archive, but for the digital component these issues are managed for readers by the digital repository and fully documented as part of a digital object's preservation metadata. Detailed information relating to the form of the digital material therefore need not to be included in the EAD catalogue, though a chronology of the creator's hardware and software environments, which could be referred to from lower level descriptions may be useful. Information about the repository's preservation policy (which will evolve) should not be included, but researchers might be referred to such information via a URL held in this element.

**Biography or History <bioghist>** Paradigm's Academic Advisory Board emphasised the usefulness of biographical information about an archive's creator(s), and this contextual information has always been considered important by archivists. Where the principal creator (i.e. the politician) has an entry in a standard published source (e.g. the *Dictionary of National Biography*, *Who's Who* or *Who Was Who*), researchers can be referred to this. The archivist should, however, provide additional information gleaned from the archive itself, or emphasise particular activities or achievements where these are well-represented in the archive. If creating a biographical account for a politician from scratch, it might include references to: dates; education and early life; political career, including positions held, any ministerial roles, groups or committees involved with, constituency represented etc. The archivist should also supply information about the administrative structures, staffing, and functions of the politician's Westminster and constituency offices where relevant.

**Custodial History <custodhist>** Depending on which approach to collection development is taken (see Chapter 02 *Collection development*), the chain of custody of a digital or hybrid archive may be a complex one, especially in the case of politicians, whose various staff members also generate records. Digital provenance will be recorded fully in the preservation metadata associated with the digital objects making up an archive; at collection level this should be summarised, with an emphasis on the record creators and pre-accession provenance, rather than focusing in detail on authenticity checks or migrations; custodial information about the paper and digital components of a hybrid archive should also be pulled together here.

**Acquisition Information <acquinfo>** This element is used to record the immediate source of acquisition of the archive (the politician) and the terms under which it is held by the repository. It may be helpful to provide more detailed information here about the number of accessions; the Paradigm Academic Advisory Board felt it was important for users to have information about accession dates.

**Scope and Content <scopecontent>** An overview of the content of the archive, including reference to: significant individuals, organisations, events and activities represented; range of the material, in terms of geography, subject, timespan; record types and their creators; record keeping practices at politicians' offices; and an indication of research potential. This may include comments on what isn't included; e.g. in one of Paradigm's exemplar archives there were no presentation slides or texts of speeches because the politician concerned tends to speak off the cuff, from brief factual bullet points.

**Appraisal Information <appraisal>** This element should record all appraisal decisions and actions, and the rationale on which they are based. See Chapter 04 *Appraisal and disposal* for a detailed discussion of appraisal in relation to digital and hybrid personal archives.

**Arrangement <arrangement>** This element is used to record information on how the archivist has arranged the material. See the section on templates for arrangement (p. 176) for examples of how a digital or hybrid archive might be arranged.

**Preferred Citation <prefercite>** In the context of a hybrid or digital archive, it is important that researchers know how to cite the material in their work. Citation should be based on shelfmark rather than on any digital identifier generated by the repository software.

**Conditions Governing Access <accessrestrict>** This element is used to record conditions affecting access to the archive material by users. This is very important in a digital environment, where a range of legislation can affect whether or not records may be opened to the public. See Chapter 09 *Legal issues* for a detailed discussion of relevant legislation.

In the case of the Paradigm testbed material, exemplar archives were closed to readers under agreement with the depositors. Where a collection is closed, the closure period and reasons should be stated here. However, in the case of digital and hybrid personal archives more generally, it may be decided over time to open parts of a collection before the full copyright duration has expired under certain conditions; paper elements of the archive may also be made available. Feasibly, then, a single hybrid archive may contain material subject to a range of different restrictions, e.g.:

- Digital material which is still in copyright: made available in secure search room to registered readers, provided they have signed a condition of use form; probably made available in read-only formats and on a specially configured PC.
- Digital material which has no copyright restrictions, or for which permission has been obtained (e.g. websites): made available online, provided readers click on a condition of use form before accessing the material.
- Digital material closed to researchers under copyright, *DPA* or other legislation.
- Hard copy material with no *DPA* restrictions open to registered readers in secure search room.
- Hard copy material subject to closures under the *DPA* or by agreement with depositor.

This has the potential to confuse researchers, and any access restrictions should be clearly outlined at collection level.

The METS documents for an individual digital object may also hold IPR metadata specific to conditions governing access.

**Conditions Governing Use <userestrict>** This element relates to the material in an archive which is listed as open to researchers (whether online or in a searchroom) in <accessrestrict>. It describes restrictions on a researcher's reuse of the information for the purposes of quotation, publication or reproduction. These may be imposed by the repository, by the donor or depositor, or by national and international statutes.

It should identify the principal copyright holder in the material (usually the donor or depositor). Given that there will be multiple copyright holders in any one archive, a general statement should also be provided advising researchers that it is their responsibility to seek the copyright holder's permission before the material can be reproduced or published. If the digital repository has an institutional 'take down policy' it should also be referred to here.

A statement should be included about the need for users to sign a copyright declaration form before viewing or reusing certain categories of material. See Chapter 09 *Legal issues* (p. 258) for an explanation of fair and lawful use and related issues.

The METS documents for an individual digital object may also hold IPR metadata specific to conditions governing use.

**Other Finding Aid <otherfindaid>** This may be a useful element in which to explain that each digital object in the archive has an associated METS document in the digital repository which provides a limited quantity of descriptive metadata along with other relevant metadata.

**Controlled Access Headings <controlaccess>** Researchers may undertake full-text searches of catalogue records quite easily (e.g. by using the Find command in their web browser). However, searches carried out like this are indiscriminate and operate at all levels of description, meaning a searcher may be overwhelmed with hits, many of which are irrelevant.

<controlaccess> facilitates searching by acting as a wrapper for key access points. Entries should be authority-controlled to ensure that standardised and authoritative versions of the terms are used.

There are ten possible <controlaccess> subelements. Paradigm's Academic Advisory Board suggested that researchers would want to browse on terms like subject, place and creator, and Paradigm therefore recommends using only the following:

- Personal name <persname>.
- Family name <famname>.
- Corporate Name <corpname>.
- Geographic Name <geogname>.
- Genre/Physical Characteristics <genreform> [to provide information on particular genres represented in the archive, such as 'speeches', 'diaries' or 'correspondence'].
- Subject <subject>.

Each element used should include a 'source' attribute to indicate the source of a controlled vocabulary term or the rules that were used to formulate it. See above on *Indexing and Authority Files* (p. 172) for information about index terms, rules and thesauri.

## Suggested EAD elements required at c03 folder level (generic)

The following elements are recommended for describing digital material at this level. In many cases, it is likely that this is the lowest level to which it is practical to catalogue.

### Elements in the <did>

**<unitid>** The shelfmark for the folder, e.g. TEST/1/1/1

**<unittitle>** This element should contain a supplied term indicating the type of object being described, along with the original title allocated to the object by the creator (the latter will have been recorded in the PREMIS element originalName). It is useful to give the original title in inverted commas to indicate quotation, e.g. Folder: "Speeches". The same applies for sub-folders; e.g. a subfolder within "Speeches" might be given the title Sub-folder: "Oral questions".

**<unitdate>** This should record a span date, usually at month level, covering the earliest date of creation for a document within the folder, to the latest last modified date of any of the documents. The 'normal' attribute will be set to record a date in accordance with ISO 8601.

**<physdesc><extent>** Supply the number of sub-folders and/or the total number of items (i.e. files). Also include size in MB.

**<materialspec>** Use to indicate the creator's original file formats represented in the folder.

**<dao>** If desired, a link can be made to the METS document which represents the 'folder' in the digital repository and points to the children sub-folders and files of the folder.

**Example <did> for folder level:**

```
<did>
    <unitid>TEST/1/1/1</unitid>
    <unittitle>Folder: "Speeches"</unittitle>
    <unitdate normal="2002-06/2005-09">Jun 2002-Sep 2005</unitdate>
    <physdesc>
        <extent>26 files; 1.8 Megabytes</extent>
    </physdesc>
    <materialspec>The folder is comprised of 26 files originally created in the following formats: 20 files in Microsoft Word for Windows 97-2003 (.doc); 3 files in Portable Document Format 1.4 (.pdf); and 3 Binary Interchanage File Format (BIFF) 7 (Microsoft Excel 95) (.xls).
    </materialspec>
</did>
```

### Other elements

**<phystech>** This element could be used to refer researchers to the creator's relevant hardware and software environments as listed in the chronology of environments at fonds level. It should also be noted that material which originated elsewhere and was simply stored in a folder by the principal creator may have been generated in an altogether different environment.

**<scope and content>** An overview of the intellectual content of the material in the folder. Archivists could get a feel for the scope and content of different folders by using indexing technologies or automatic keyphrase extraction tools like Kea[1] Or Data Fountains[2] to pick out key words and phrases which are frequently used; this would help them to summarise the intellectual content of a folder without having to read all of its contents.

If not using <dao> to link to representations of the folder in the digital repository, and if not cataloguing to item level in EAD and providing links at that level, it may be appropriate to include a list of the files included in the folder here, using <daogrp> and <daoloc> to link to the digital objects via their repository METS documents.

**<arrangement>** Use to explain the arrangement of the component sub-folders or files and its rationale. Arrangement will usually replicate the order in which the records are found; this may reflect automatic arrangement by the computer, or decisions made by the creator on the arrangement of their files (by date, format, etc.). If any changes to arrangement have been made by the archivist they should be recorded here.

**<accessrestrict>** Provide an overview of any access restrictions (e.g. closures under the *Data Protection Act (DPA)*) which apply to the component items.

**<controlaccess>** A representation of the types of material held in the series could be given using the <genreform> element and a controlled vocabulary. The Dublin Core Metadata Initiative (DCMI) types vocabulary is one example, but the distinction between digital and traditional formats in this vocabulary is not clear (e.g. StillImage could apply to an analogue or digital image).[3]

## Suggested EAD elements required at c04 item level (generic)

It is unlikely that repositories will be able to catalogue any but the most important material in a digital or hybrid archive at item level. In the event that item-level cataloguing is undertaken, the following EAD elements are recommended:

1 The New Zealand Digital Library, *Keyphrase Extraction Algorithm (KEA) website.* URL: <http://www.nzdl.org/kea/>

2 Data Fountains Project, *Data Fountains website*. URL: <http://datafountains.ucr.edu/>

3 Dublin Core Metadata Initiative, 'DCMI Type Vocabulary' (August 2006), *Dublin Core website.* URL: <http://dublincore.org/documents/dcmi-type-vocabulary/>

## Elements in the <did>

**<unitid>** The shelfmark for the file, eg TEST/1/1/2/3

**<unittitle>** Indicate record type and also supply creator's original title in inverted commas, e.g. File: "Campaign speech, May 2005".

**<unitdate>** In the case of an individual file, researchers are likely to be interested in knowing the date on which the file was first created, as well as the date on which it was last modified (if this information can be ascertained from the preservation metadata). This would be useful, for example, in the case of spreadsheets where new information is added over time; and speeches or papers, where the length of time a document took to write and revise before reaching the final version might be interesting. These can be recorded as span dates, with the normalise attribute set to record the ISO 8601-conformant format. At this level, a day (and even, in some cases, a time) might be useful and standards for formatting dates and times should be agreed on.

**<origination>** Record the name of the document's creator. If more than one person has modified a document, ideally multiple names should be recorded here, although the limitations of current technology may prevent the extraction of this information. NCA Rules should be used.

**<physdesc><extent>** Record the size of the file in KB/MB.

**<materialspec>** Free-text description of the creator's original file format.

**<dao>** Use to link to the digital object in its METS wrapper.

**Example <did> for file level:**

```
<did>
    <unitid>TEST/1/1/2/3</unitid>
    <unittitle>File: "Campaign speech, May 2005"</unittitle>
    <unitdate normal="2005-04-13/2005-05-01">13 Apr-1 May 2005</unitdate>
    <physdesc>
        <extent>74.5 kilobytes</extent>
    </physdesc>
    <materialspec>Original file format: Microsoft Word for Windows 97-2003 (.doc).
    </materialspec>
    <dao linktype="simple" href="http://shuttle.paradigm.ac.uk:8085/fedora/get/demo:1/PDF"
    actuate="onrequest" show="new" role="application/msword" />
</did>
```

## Other elements

**<phystech>** Use to indicate whether the original file was encrypted or password protected.

**<scopecontent>** Give a very brief indication of content, and if the document exists in a number of different drafts or versions, indicate which version this file represents.

**<accessrestrict>** Indicate any restrictions on access which apply to the file, e.g. closed; accessible to readers who have signed a condition of use form; available in read-only format, etc.

**<relatedmaterial>** Refer to any other version of the document (e.g. an earlier draft) which is included in the archive, citing its shelfmark. There should be a repository-level policy on the method and semantics of identifying versions; such a vocabulary could be based on the findings of the RIVER project, e.g. DigitalCopy, DigitalVariant, DigitalEdition.[1]

1 Salley Rumsey et al., *Scoping Study on Repository Version Identification (RIVER)*, Final Report, v.05, Draft Final Report (March 2006). URL: <http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf>

## EAD and its relationship to other metadata about a digital object

There will inevitably be some degree of overlap between the detailed descriptive metadata provided for researchers in the EAD catalogue and the metadata wrapped with a digital object in its associated METS document, which will also contain a level of descriptive metadata. It is therefore important to ensure that where duplication occurs, the metadata provided is consistent and some workflow for keeping the two metadata instances in sync will need to be developed. In some cases the preservation metadata will be recorded in the form of a code or machine-readable format; where this data is supplied in the EAD finding aid, it should be in an understandable form for researchers, with machine-readable format recorded using attributes if necessary.

As item-level EAD catalogues are unlikely to be created, the digital object and its metadata (the DIP) should be as self-describing as possible, so if researchers link from a higher level in the EAD catalogue (e.g. from series description to folder) they will be able to understand component objects.

**Paradigm suggests that the following information is recorded as part of the DIP:**
*Descriptive metadata:*
MODS for description in the digital repository:
MODS is used at Oxford in other digital library settings, such as the Oxford Digital Library. By providing a MODS record for individual digital archives users will be able to cross-search local repositories.

- **<identifier>** (maps to EAD unitid element).
- **<titleInfo><title>** (maps to EAD title element).
- **<name role="creator">** (maps to EAD origination element).
- **<originInfo><dateCreated>** (maps to EAD date element).
- **<abstract>** (maps to EAD scopecontent element).

Dublin Core metadata for OAI-PMH harvesters:

- **<identifier>** (maps to EAD unitid element).
- **<title>** (maps to EAD title element).
- **<creator>** (maps to EAD origination element).
- **<date(s) of creation>**, using ISO 8601 (maps to EAD date element).
- **<description>** (brief account of content; maps to EAD scopecontent element).
- **<relation>** (maps to EAD relatedmaterial element); provide information about different versions of the document which exist in the same archive using the archive's formal identification system.

*Administrative metadata:*
*a) Rights metadata*
Rights metadata relating to access and use: recorded using METSRights (see p. 145)

- **<rightsDeclaration>** A statement describing the Intellectual Property Rights associated with the item and a brief outline of any restrictions on use.
- **<rightsHolder><rightsHolderName>** Identity of the rights holder; 'type' of rights holder can also be specified.
- **<rightsHolder><rightsHolderContact>** Provide initial contact details for the current rights holder or their representative (in most cases researchers will be advised to apply to repository staff as rightsholders' personal contact details cannot be published without permission).

*b) Technical/digital provenance metadata*
Information drawn from PREMIS:

- **<objectIdentifier>** Persistent identifier.
- **<originalName>** The creator's original name for the file.
- **<size>** File size in bytes.
- **<fixity>** The repository should provide a fixity checksum for each instance of the object, so that the user can verify that it is authentic. A link to information about the repository's authenticity mechanisms could also be provided here.
- **<format>** The current format (which may or may not be the same as the creating format).
- **<inhibitors>** Details of any inhibitors or passwords imposed by creator.
- **<creatingApplication>** Creating hardware and software; and date created by application (this should be a span date consistent with EAD record, recording date first created and date last modified).
- **<relationship>** Details of digital provenance; e.g. a pointer to an object from which the current object is derived.

c) Structural metadata
Structural metadata: drawn from METS structMap or generated by the digital repository, recording information about the object's relationships with other digital objects in the archive, expressed using Resource Description Framework (RDF). See Chapter 05 *Administrative and preservation metadata* for more information about how digital objects and intellectual entities are defined in PREMIS.

# Arranging and cataloguing emails

## How are emails different to paper correspondence?

- Incoming mail is often organised in a very structured way, but outgoing mail is frequently unsorted and simply stored in a single sent mail folder.
- Ongoing exchanges of correspondence mean that any one message can also contain an extensive record of previous emails. Whilst this information is useful in documenting an extended correspondence, it can be confusing for anyone who was not involved in the exchange: e.g. correspondents might include a received email in their reply and insert comments at various points rather than sending their responses as a single block of text.
- Whilst most organisations keep copies of outgoing hard-copy correspondence, the same is not necessarily true of individuals. When using email, both outgoing and incoming mail are automatically retained. Sent mail can often be preserved in record strings as well as stored in the sent folder.
- Identities (email addresses) are likely to be less fixed than correspondence addresses.
- Dating is much more precise: sent, received and other dates are recorded, often down to the minute and second.
- It is more likely with email than hard copy that messages will be sent to multiple recipients.

- Sheer quantity of material: in many ways its speed has meant that the email message has replaced the telephone call as an informal mode of contact; this means that far more messages are likely to be preserved in an email directory than in a hard copy filing system.
- This informal status means that much more information about an individual and their varied networks is preserved; many people use a single email directory for both business and personal contacts.
- Emails are likely to have a wider range of attachments than the kind of enclosures included in hard copy mail. These attachments form an integral part of the electronic message and often pose separate preservation challenges of their own.
- It is possible to retrieve blind copies from sent mails, whereas it is very difficult to identify blind or illicit copies in paper correspondence.
- During the course of their work an individual is likely to join many distribution lists and listserves; they are more likely to retain these kinds of group mailings than hard copy circulars, and these can tell us something of their interests and spheres of activity.
- Individuals usually maintain a contacts list or address book as an integral part of their email directory; this can give us further important information about their networks.

## Templates for arrangement

### How does an email directory relate to component levels in EAD?

Some possible models may look as follows:

Example 1:

| Structure | ISAD(G) level | EAD encoding |
|---|---|---|
| Politician | Fonds | collection level |
| *Politician's assistant | sub-fonds | <c01 level="subgroup"> |
| Email archive | directory & diary series | <c02 level="series"> |
| Email folder | subseries | <c03 level= "subseries"> |
| Email | item | <c04 level="item"> |
| Email Attachment | piece | <co5 level="piece"> |

*This information might be recorded when the archivist describes the provenance of a collection, rather than recorded as a subgroup

Example 2:

| Structure | ISAD(G) level | EAD encoding |
|---|---|---|
| Politician | Fonds | Collection level |
| Westminster office | Sub-fonds | <c01 level="subgroup"> |
| Email directory | Series | <c02 level="series"> |
| Mail/contacts/diary | Subseries | <c03 level= "subseries"> |
| Email folder | Subsubseries | <c04 level="sub-subseries"> |
| Individual email | Item | <c05 level="item"> |
| Message & attachment | Separate pieces | <c06 level="piece"> |

Example 3:

| Structure | ISAD(G) level | EAD encoding |
|---|---|---|
| Politician | Fonds | Collection level |
| Constituency office | Sub-fonds | <c01 level="subgroup"> |
| Correspondence | Series | <c02 level="series"> |
| Email/paper correspondence | Separate subseries | <c03 level= "subseries"> |
| Email folder/paper file | Subsubseries | <c04 level="sub-subseries"> |
| Individual email | Item | <c05 level="item"> |
| Message & attachment | Separate pieces | <c06 level="piece"> |

### Paradigm exemplars

The creation of a full catalogue for the personal archive and the need to cite individual items makes it necessary that the repository choose an order for structuring the catalogue by which shelfmarks may be assigned to individual items. These decisions are not always clear-cut and can vary from case to case.

**Example A**

An email directory with a large number of emails, held in over 80 folders. There is also one large, unstructured, sent mail folder.

The directory should be left in its original arrangement of 80 folders and sent mail folder. Within folders, the archivist should retain a chronological order to facilitate the connection between incoming and outgoing emails. The archivist might consider picking out individual emails from particularly significant individuals to catalogue at item level.

**Example B**

An email directory with over 200,000 emails divided between three folders: inbox, archive and sent-mail.

In this example, a version of the directory in its archival form (i.e. unstructured, reflecting the creator's practice) might be retained; using different METS structural maps of the directory the repository could provide alternative views of the data – sorted by correspondent, date or subject. Alternatively, a version of the directory could be provided for the researcher to do their own sorting and retrieval if software to facilitate this is available. Another approach may be to abstract major correspondents and arrange the remaining correspondence by date.

The archivist may encounter some emails which have obviously been filed in the wrong folder. These could be moved to the correct folder if the size of the directory makes this reasonable to attempt. This would, however, necessitate a close analysis of content and context: an individual may have had their reasons for filing a message in a particular folder, even if this is not instantly obvious.

The majority of emails accessioned by Paradigm have been stored in an email account of some kind. There are, however, some messages which have been saved as email files (usually in html, plain text or .eml formats) into a general office subject folder. For example, some offices routinely delete all messages, with messages of special significance being saved with related records outside of the email directory. In these instances, it is recommended that the messages be retained with the related grouping of records outside of the e-mail directory and catalogued as part of that record series.

## At what level should emails be catalogued in EAD?

As the above examples indicate, there are many possibilities for the intellectual arrangement of emails within a personal archive. Similarly, there are many options for cataloguing and access. When making these decisions, the wide range of uses to which researchers might put such material in future should be borne in mind. To date, very few email archives are publicly available, so there is no tradition of past practice to base decisions on. One notable exception to this is the Enron da-

taset:[1] this dataset comprises some 0.5 million email messages, generated by around 150 users, and was obtained by investigators during the Enron accounting fraud scandal in 2001. The Federal Energy Regulatory Commission was charged with investigating the company – which involved reviewing emails along with other data; the email dataset was subsequently made available on the Web. Unfortunately its archival integrity has been lost as the result of a 'cleaning' process (which included the removal of attachments, the deletion of some messages and conversion of email addresses). However, it has already been used as a research resource and the work undertaken offers an insight into how researchers might make use of archival email directories for purposes beyond the more obvious biographical or historical research; for instance, a social-network analysis of the data has been carried out; it has also been used as source material for a number of email visualisation experiments, natural language processing investigations and the subject of research examining methods for the automatic categorisation of email into folders (something which might be useful for the management of email in future).

As this research indicates, there is a good argument for giving researchers unmediated access to an archival email directory, which they would encounter in the same way as the creator would have entered and viewed it. Experiencing email in the same way as it was used is, as Maureen Pennock[2] points out, an important historical and social experience, although changes in tacit knowledge may make this difficult for us to provide and for users to navigate in 100 years time. Enabling researchers to display or reorder the data in multiple ways (using the METS structural map) could also facilitate different types of research.

Paradigm believes that some degree of mediation between researcher and archival email directory is necessary for a full understanding of the material. The EAD catalogue and the metadata included in the DIP will also contain additional information which would not necessarily be obvious simply from viewing the email directly. Paradigm recommends creating a high-level description (at series or sub-series level, depending on the overall arrangement of the archive) for the entire directory.

## Suggested elements for use at email directory (series or sub-series) level

### Elements in the <did>

**<unitid>** Shelfmark
**<unittitle>** Email directory
**<unitdate>** Covering dates by year; based on the earliest message sent or received and the last sent or received; normalise to ISO 8601. This information will be readily available if header details can be automatically extracted from component emails.
**<origination>** Controlled access entry identifying the individual(s) who maintained the directory
**<physdesc><extent>** Record the number of folders, and total number of emails and attachments. Also supply size in MB. This information will be available via automatically extracted metadata.
**<materialspec>** Record the file formats represented (e.g. Microsoft Outlook Personal Folders (.pst); also an overview of formats represented as attachments.
**<dao>** If desired, a link can be made to the METS document which represents the 'email directory' in the digital repository and points to the children folders and emails in the directory.

### Other elements

**<phystech>** Record the creator's email environments here (e.g. webmail account and Mozilla Thunderbird email client).
**<bioghist>** A note on the relationship between correspondent and recipient could be useful at this level.
**<scopecontent>** An overview of the principal topics, events, individuals and dates covered in the directory's content. Make a particular note of folders over a certain size, or messages from prominent people, to indicate the major issues of interest to the owner of the email directory; consider tagging significant individuals and subjects as controlled access terms to facilitate information

1 CALO Project, *Enron Email Dataset website*. URL: <http://www.cs.cmu.edu/~enron/>
2 Maureen Pennock, 'Curating E-mails: A life-cycle approach to the management and preservation of e-mail messages', *DCC Curation Manual Instalment* (July 2006). URL: <http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails/curating-e-mails.pdf>

retrieval. If the repository intends researchers to be able to link directly to folder-level digital objects from this level, then a list of folders could be included here, making use of the linking <daogrp> and <daoloc> elements, as at fonds level.
**<appraisal>** Describe any appraisal (see p. Chapter 04 *Appraisal and disposal*) decisions and actions specific to the email component of the archive.
**<arrangement>** Explain the rationale behind the current arrangement, including an indication of original order and any arrangement carried out by the archivist. If using multiple METS structural maps to present various alternative arrangements, this should be indicated here, using standard vocabulary to identify different versions (DigitalVariant, DigitalEdition, etc).
**<accessrestrict>** and **<userestrict>** may be necessary if there are closures or other access and use restrictions specific to the email directory which are not covered at fonds level.
**<controlaccess>** At this level it would be useful to tag key individuals and subjects represented in the directory to facilitate information retrieval.

## Suggested elements for use at c03 or c04 (email folder/subseries) level

### Elements in the <did>

**<unitid>** The shelfmark for the folder.
**<unittitle>** Title should reflect the type of object described, along with the creator's original title
**<unitdate>** Ideally this should record a span date, by month, covering the earliest and the latest dates of sending or receipt by the creator. Use traditional date format and normalise to ISO 8601.
**<physdesc><extent>** Supply the number of: email messages; the number of these which also include attachments (and total number of attachments); and the size of the folder in MB. This information should be available via automatically extracted metadata
**<materialspec>** Use this element to record the file formats of any attachments; the email client will have been recorded at a higher level.
**<dao>** Use to link to the digital folder and its associated metadata. As described under generic folder level, attributes will indicate the type of link, how it will be activated and how it will appear.

**Example <did> for folder level:**

```
<did>
    <unitid>TEST/1/2/1</unitid>
    <unittitle>Folder: "Neighbourhood renewal"</unittitle>
    <unitdate normal="2004-06/2006-05">Jun 2004-May 2006</unitdate>
    <physdesc>
        <extent>120 email messages, of which 22 have attachments (28 attachments in
        total); 26MB.</extent>
    </physdesc>
    <materialspec>13 attachment files in Microsoft Word for Windows 97-2003 (.doc); 12 attachment files
    in Portable Document Format 1.4 (.pdf); and 3 attached image files in JPEG File Interchange Format
    1.02 (.jpeg)</materialspec>
    <dao linktype="simple" href="http://shuttle.paradigm.ac.uk:8085/fedora/get/demo:1/PDF"
    actuate="onrequest" show="new" ></dao>
</did>
```

### Other elements

**<scope and content>** An overview of the intellectual content of the folder, which will usually be based around a specific subject/project or correspondent. Include some specific detail about what the emails refer to; pick out significant individuals represented; also places; subjects, events, activities and dates. An indication of research value can also be supplied. When the folder is very large and unsorted a higher level of detail may be required here. This may be facilitated by the use of indexing tools to pick out frequently used words.
**<arrangement>** A note on the arrangement of the component messages; this will usually be based on original (chronological) order. If the repository supplies one version in original order and one that users can manipulate/reorder, this should be explained here too.

## Suggested elements for use at c04 (item) level

None but the most significant individual emails are likely to be catalogued at item level. However, it may be possible to produce brief catalogue descriptions using metadata that is simple to extract automatically using an extraction tool; most of the email header information should be easy to extract and is likely to be held in the DIPs for individual emails. If repositories intend to provide EAD item-level descriptions for email, the following elements are recommended:

### Elements in the <did>

**<unitid>** Shelfmark
**<unittitle>** Subject line of the email. If the sender fails to supply a subject line, the archivist (depending on the bulk of material concerned) may supply an appropriate title in square brackets (this convention should be made clear to users); in reality the field may simply be left blank, reflecting the original. Another habit of many email correspondents is – rather than searching for and then retyping the email address – to find the first available message from the desired correspondent and click on reply without adjusting the title. Again this leads to misleading subject headings. In reality, archivists are unlikely to have the time to record information like this and it will be left to the researcher to discover.
Two **<unitdate>** elements will be required:

- **<unitdate>** Date and time sent; the date format should be granular – to the minute level – to reflect the pace of email exchange.
- **<unitdate>** Date and time received: [ditto]

**<origination>** The sender of the email, with their email address.
**<physdesc><extent>** This is likely to be the size in bytes, as extracted automatically. If cataloguing manually, the archivist should also include the 'intellectual' extent, e.g. 3 pieces: email message and two attachments.
**<materialspec>** Use to indicate the file format of any attachments. It is probably unnecessary to include any information about the email message itself here.
**<dao>** If desired, a link can be made to the METS document which represents the individual email in the digital repository.

### Other elements

**<scopecontent>**
Much of the information that may be automatically extracted will not map easily to a specific EAD element. Therefore the <scopecontent> tag may need to be used to record such metadata as:

- Recipient, with email address.
- Other primary addresses listed in the 'To' field of the email, with email addresses.
- If a listserv posting, name and address of listserv.
- Names and email addresses of those copied into the message (recorded in the cc and bcc field of the email) .
- Message priority: record only when a message is flagged in some way (e.g. urgent).
- Whether an automatic signature was attached.
- Any encryption information.
- List of attachments.
- Link to email and attachment by means of either <dao> if to single message, or <daogrp> and <daoloc> for an email with attachments. If possible use <daodesc> in each case to indicate name and format of attachment.

Paradigm recommends repeated text (in the form of message strings) should be retained as part of the record; this text can contain important contextual links and relationships. However, the catalogue entry for the email should only reflect the latest transaction.

## Practical issues

### Ascertaining identities of correspondents

Email archives often include emails from individuals, which are not signed and where the correspondent information amounts to an email address only. In these circumstances, some cross-referencing with other emails, searching for email addresses, or contact with the depositor may be necessary.

It may be useful for Encoded Archival Context/National Name Authority Files (NNAF) to include address histories for individuals. These addresses could be physical (perhaps work and home) and virtual (email and web addresses); such information would have to be managed appropriately according to the *Data Protection Act*.

# Arranging and cataloguing websites

Between 5 April and 9 May 2005 (the General Election period), Paradigm made regular snapshots of the websites maintained by select politicians (from whom permission was sought). Both HTTrack and Adobe 7.0 Professional software were used.

Many politicians (and individuals in other fields) maintain their own websites, and these sites therefore form an integral part of an individual's personal archive. The approach to collection development (see Chapter 02 *Collection development*) taken in any one case will determine how many snapshots or versions of the website are held by the digital repository, and the period these cover.

To date, much of the work on archiving websites has been undertaken by libraries; this means that whilst Dublin Core is being employed in some cases, catalogues for the archived websites often take the form of MARC records stored in bibliographic databases and library catalogues. Websites have therefore largely been viewed as publications, and the principal creator as the publisher. There are also moves in a number of countries towards making websites subject to Legal Deposit. In the UK, the *Legal Deposit Libraries Act* 2003 extends legal deposit to non-print forms and currently comprises enabling legislation. The British Library is encouraging the deposit of online material under the *Voluntary Code of Practice* 2000; an independent Legal Deposit Advisory Panel has also been established by Government to oversee the various stages of secondary legislation which are likely to lead to Regulations by format (one of these formats being websites).

Should websites become subject to legal deposit, archivists may take the decision to treat an individual's website as one of their publications and to exclude it from their digital archive unless there is good reason for its inclusion. Until then, it is still important for archivists to deal with the personal websites of their donors and depositors, and to capture different versions of these mutable records over time.

Rather than taking these website snapshots themselves, archivists may decide to submit details of the relevant sites to an organisation like the UK Web Archiving Consortium (UKWAC)[1] which would undertake web harvesting and archiving on their behalf. The EAD catalogue could then link to these snapshots, whilst making it clear that the snapshots were held in a different repository to the rest of an individual's digital archive.

If undertaking this work in-house, a succession of snapshots of an individual's website should probably be placed at series level in an EAD catalogue. The subfonds above it will either reflect the office where the site is maintained, or the individual (usually a member of politicians' staff) responsible for maintaining the site. Each snapshot (whether of a homepage or entire site) should be treated as an item. Each item will therefore be the equivalent of what the Australian web archiving project PANDORA refers to as a date stamped 'edition'.

1 UK Web Archiving Consortium, *UK Web Archiving Consortium website.* URL: <http://www.webarchive.org.uk/>

Cataloguing at series level should pull all the essential information about a single website together. Item level cataloguing will be minimal and focus on controlled access index terms to facilitate searching and browsing. The functionality of websites, and the fact that they are or have been in the public domain, may mean that users are permitted to navigate sites freely, although before doing this they should be made aware (via the EAD catalogue and DIP metadata) of IPR issues, of the fact that they are viewing an archived website, and of metadata regarding creator, dates and technical issues (like loss of functionality).

## Suggested elements for use at c02/3 (series) level

### Elements in the <did>

**<unitid>** Shelfmark for the series (individual snapshots will be demarcated by splitters)
**<unittitle>** Title supplied by the archivist, e.g. 'Series of website snapshots'
**<unitdate>** Span date: from date the website was created to the date of the last snapshot. Normalise date attribute in accordance with ISO 8601.
**<physdesc><extent>** Supply the number of items, i.e. individual snapshots. Supply the overall size of the series in MB.
**<materialspec>** Overview of file formats represented, e.g. static html, css, javascript, JPEG images, etc.
**<dao>** If desired, a link can be made to the METS document which represents the series of wensite snapshots in the digital repository.

### Other elements

**<phystech>** Possibly give an indication of the software used by the site's author(s). This should probably also include an explanation of any loss of functionality and links to external sites in the archived version of the site, as well as different look-and-feel from the original.
**<scope and content>** General overview, including: history of the site; its author and any other individuals involved in supplying content, where the information is available; when it was first created; its general structure and content (e.g. the site includes homepage, biography of MP, latest news, constituency reports, etc.); an indication of its research potential; frequency of snapshots. Also record snapshot depth: whether the entire website is captured in every snapshot, or whether snapshots are limited to the homepage.
**<appraisal>** This might be used to record the rationale behind the frequency of snapshots, e.g. to record that where identical snapshots (i.e. shots made on different dates where no new material was added during the intervening period) have occurred, the duplicates have been ommitted from the web archive.
**<arrangement>** A note on the arrangement: generally a chronological series of dated snapshots.
**<userestrict>** Use to outline the copyright status of the material. This is particularly important in relation to websites which are classed as published material. The copyright holder should be identified and fair dealing provisions outlined (see p. 258). Also state that the repository has sought permission from the copyright holder to make the site available to researchers. It may be a good idea to have a copyright disclaimer too (in case anyone represented in the site objects to their inclusion in the archive).
**<controlaccess>** Use subject indexing at a very general level here and identify significant individuals associated with the website. Much more detailed indexing should take place at the item level.

## Suggested elements for use at c04 (item) level

### Elements in the <did>

**<unitid>** The shelfmark.
**<unittitle>** Use a term to indicate the element of the site preserved, e.g. Website or Homepage, and the title that appears at the top of the home page in inverted commas; e.g. Website: "Rt Hon Politician MP, Member of Parliament for xxx".
**<unitdate>** Record the date on which the snapshot was taken. Dates of original creation or pub-

lication of the site can be recorded at series level. Ensure that this is normalised to ISO 8601 to facilitate information retrieval.
**<origination>** Use this element to record the name of the site's author. If there are multiple authors, list the principal author first; use controlled access terms based on the NCA Rules.
**<physdesc><extent>** May be used to record the number of pages represented by the capture and size in MB.
**<dao>** Use to link to the digital object in its METS wrapper (DIP).

### Other elements

**<scope and content>** Possibly use this element to record the URL of the original site and the address of the archived location for citation purposes.
**<controlaccess>** This is probably the most important element of descriptive metadata for websites: researchers are likely to want to search or browse on subjects or individuals and then have direct access to the websites themselves. Include extensive subject access terms.

### DIP metadata about websites

In addition to the information recorded in the EAD catalogue and the recommended generic DIP metadata (see p. 189), the DIP metadata for websites should also include information on the snapshot software used by the digital repository.

## Useful resources

### ISAD(G) 2

International Council on Archives, Committee on Descriptive Standards, 'History of the International Council on Archives Committee on Descriptive Standards', *International Council on Archives, Committee on Descriptive Standards website*.
URL: <http://www.icacds.org.uk/eng/history.htm>

International Council on Archives, Committee on Descriptive Standards, *General International Standard Archival Description*, Second Edition (2000).
URL: <http://www.ica.org/biblio.php?pdocid=1>

### Library of Congress Standards

The Library of Congress, 'Cataloguing and Acquisitions', *The Library of Congress website.*
URL: <http://www.loc.gov/aba/>

Network Development and MARCS Standard Office, *Metadata Encoding & Transmission Standard (METS) website.*
URL: <http://www.loc.gov/standards/mets/>
*Note: An XML schema which provides a framework for encoding descriptive, administrative, and structural metadata. METS is being developed by the Digital Library Federation (DLF) and is maintained by the Library of Congress. METS has been adopted by an increasing number of digital repositories .*

Network Development and MARCS Standard Office, *Metadata Object Description Schema (MODS) website.*
URL: <http://www.loc.gov/standards/mods/>
*An XML schema of bibliographic elements for describing objects.*

## Rules for personal, corporate and family names

Anglo-American Cataloguing Rules (AACR2), *Anglo-American Cataloguing Rules (AACR2)*, Second Edition (2002; Revison 2005 Update).
URL: <http://www.aacr2.org/>
*Note: AACR2 is a bibliographic standard widely used in libraries. Useful for personal, corporate, and place names.*

International Council on Archives, Committee on Descriptive Standards, *International Standard Archival Authority Record for Corporate Bodies, Persons and Families*, Second Edition (2003).
URL: <http://www.ica.org/biblio.php?pdocid=144>

National Council on Archives, *Rules for the Construction of Personal, Place and Corporate Names* (1997).
URL: <http://www.ncaonline.org.uk/materials/namingrules.pdf>

## Rules for geographic names

Getty Vocabulary Program, 'Thesaurus of Geographic Names', *The Getty website.*
URL: <http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html>
*Note: The Thesaurus of Geographic Names is also known as the Getty Thesaurus or TGN, and is useful as a source for the current, English-language form of place names. The TGN is useful for searching for information on UK counties prior to the 1974 boundary changes, and their current names. It is best to use TGN's preferred form for making the choice of name, and as a basis for AACR2 or NCA rules index terms, rather than using TGN directly.*

## Authority files

Encoded Archival Context Working Group, *Encoded Archival Context (EAC) website.*
URL: <http://www.iath.virginia.edu/eac/>
*Note: EAC is a formal method of 'encoding descriptions of persons, corporate bodies, and families responsible for the creation of records and other resources, where such descriptions provide context for understanding and interpreting the records and resources'.*

Gillman, Peter, *National Name Authority File: Report to the National Council on Archives* (British Library Board, 1998).
URL: <http://www.ncaonline.org.uk/materials/nationalnameauthorityfile.pdf>

Linking and Exploring Authority Files (LEAF), *LEAF website.*
URL: <http://www.crxnet.com/leaf/>
*Note: LEAF was a three year project (2001-2004). It is co-funded within the Information Society Technologies Programme of the Fifth Framework of the European Commission. LEAF developed a model architecture for establishing links between distributed authority records and providing access to them.*

## Archival network projects

Aim25, *Aim25 website.*
URL: <http://www.aim25.ac.uk/>
*Note: Aim25 provides electronic access to collection level descriptions of the archives of over fifty higher education institutions and learned societies within the Greater London area.*

Archives Hub, 'Data Creation', *Archives Hub website.*
URL: <http://www.archiveshub.ac.uk/arch/dc.shtml>
*Note: the Archives Hub provides guidance to archivists submitting archival descriptions to the service. The Hub also provides an online template which renders an ISAD(G) collection level description to EAD 2002. The Hub's website also maintains a register of forthcoming training events[1] relevant to archivists.*

Archives Network Wales, *Archives Network Wales website.*
URL: <http://www.archivesnetworkwales.info/>
*Note: Archives Network Wales allows easy searching of collections held by record offices, universities, museums and libraries in Wales.*

CASBAH Project, 'About the CASBAH Project', *Casbah website.*
URL: <http://www.casbah.ac.uk/about_project.stm>
*Note: the CASBAH Project maintains a website for locating resources relating to Caribbean Studies and the history of Black and Asian people in the United Kingdom.*

The Genesis project, *Genesis website.*
URL: <http://www.genesis.ac.uk/>
*Note: Genesis is a mapping initiative for women's history collections from libraries, archives and museums from around the British Isles.*

The National Archives, 'Access 2 Archives (A2A) programme', *The National Archives website.*
URL: <http://www.nationalarchives.gov.uk/partnerprojects/a2a/standards.htm>
*Note: The National Archives offers useful advice on standards, which was created for repositories wishing to participate in the Access to Archives (A2A) programme. A2A is a national database of catalogues from over 340 repositories in England. In order for the project to operate successfully all contributors need to conform to common professional and technical standards. The National Archives sets out the basic editorial standards to which catalogues sent for inclusion in A2A must conform. These standards interpret the General International Standard Archival Description, ISAD(G) in the context of A2A.*

Scottish Archive Network (SCAN), *Scottish Archive Network website.*
URL: <http://www.scan.org.uk/>
*Note: The Scottish Archive Network (SCAN) provides access to collection level descriptions for historical records held in fifty-two Scottish archives.*

## Descriptive cataloguing and EAD

Archives Hub, 'Encoded Archival Description', *Archives Hub website.*
URL: <http://www.archiveshub.ac.uk/arch/ead.shtml>

Fox, Michael J., *The EAD Cookbook: 2002 Edition* (July 2003)
URL: <www.archivists.org/saagroups/ead/resources/ead2002cookbook/EAD2002cookbook.pdf>

Network Development and MARC Standards Office, *Encoded Archival Description (EAD) Version 2002 website.*

1 Archives Hub, 'More training and events', *Archives Hub website.* URL: <http://www.archiveshub.ac.uk/arch/training2.shtml>

URL: <http://www.loc.gov/ead>
*Includes the DTD, the Tag Library, and Application Guidelines*

RLG EAD Advisory Group, *RLG Best Practice Guidelines for Encoded Archival Description* (August 2002).
URL: <http://www.rlg.org/en/pdfs/bpg.pdf>

## Archiving emails

Boudrez, Filip and Van den Eynde, Sofie, *DAVID: Archiving e-mail*, Version 1.0 (August 2002).
URL: <http://www.expertisecentrumdavid.be/davidproject/teksten/Rapporten/Report4.pdf>

CALO Project, 'Enron Email Dataset', *Carnegie Mellon University website.*
URL: <http://www.cs.cmu.edu/~enron/>

Pennock, Maureen,'Curating E-mails: A life-cycle approach to the management and preservation of e-mail messages', *DCC Curation Manual Instalment* (July 2006).
URL: <http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails/curating-e-mails.pdf>

## Web archiving

Bailey, Steve and Thompson, Dave, 'UKWAC: Building the UK's First Public Web Archive', *D-Lib Magazine*, 12, 1 (January 2006).
URL: <http://www.dlib.org/dlib/january06/thompson/01thompson.html>

Beresford, Philip, 'UKWAC – the first two years', *DPC Forum on Web Archiving* (12 June 2006).
URL: <http://www.dpconline.org/docs/events/060612Beresford.pdf>

Bevan, Paul, 'Archiving the 2005 UK General Election', *DPC Forum on Web Archiving* (12 June 2006).
URL: <http://www.dpconline.org/docs/events/060612Bevan.pdf>

National Library of Australia, 'Web Archiving', *Preserving Access to Digital Information (PADI) web-site.*
URL: <http://www.nla.gov.au/padi/topics/92.html>
*Provides a survey of web archiving initiatives by national libraries around the world.*

National Library of Australia and Partners, *PANDORA: Australia's web archive website.*
URL: <http://pandora.nla.gov.au>

UK Web Archiving Consortium , *UK Web Archiving Consortium website*.
URL: <www.webarchive.org.uk>

## General

Rumsey, Salley et al., *Scoping Study on Repository Version Identification (RIVER), Final Report*, v.05, Draft Final Report (31 March 2006).
URL: <http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf>

## Descriptive metadata extraction

Data Fountains Project, *Data Fountains website*.
URL: <http://datafountains.ucr.edu/>

The New Zealand Digital Library, *Keyphrase Extraction Algorithm (KEA)website*.
URL: <http://www.nzdl.org/Kea/>