The National Archives

Digital Cataloguing Practices

at The National Archives

March 2017

Digital Cataloguing Practices at The National Archives

Executive Summary

This position paper outlines the evolution and current state of cataloguing practices for digital records at The National Archives in the United Kingdom.

The disruptive impact of digitised and born-digital records has shaken ways of working as well as ways of interpreting the international standard for archival description ISAD(G). Change is not new, however, as over the last seventeen years of digital catalogues archivists have moved forward with new technologies, new ways of balancing access and privacy, the impact of user experience, social media, content sensitivity and the challenge of digital transfer.

> ‘Digital technologies are creating a paradigm shift in the archival sphere: posing challenges, but also throwing open the doors to greater access and a world of new opportunities’.[1]

This paper introduces some new concepts and considers what a second generation of born-digital records may look like. We share our understanding of born-digital, digital surrogates and digitised records while concepts such as probabilistic, contextual and temporally aware description are also defined. Although fairly new for us, we are conscious that these ‘new concepts’ are in fact interdisciplinary propositions that have been around for some time, and that colleagues around the globe are having similar thoughts and ideas.

The National Archives is aware that the UK archive sector and some international archives may look at our digital practice as a benchmark. A word of caution: our approach is not being shared in this document in order to become a set of best practice cataloguing guidelines for digital records. This position paper summarises our current practice and aims to open the conversation as widely as possible, recognising that both technology and archival practice will continue to evolve.

This paper has been written by co-creation. I would like to thank the many colleagues from different areas of expertise who have contributed their thoughts, memories and suggestions.

Jone Garmendia, Head of Cataloguing, 31 March 2017

[1] Archives Unlocked: Delivering the Vision, Introduction by Jeff James, Chief Executive and Keeper, The National Archives.

Table of Contents

1. Seventeen years of digital catalogues

PROCAT (the Public Record Office Catalogue) was launched at Kew on 9 October 2000 and on the public website in March 2001, offering an online version of the paper lists, PRO Guide and some related finding aids. As a result of the move from paper to an online catalogue in 2000, the following established practices were changed.

- We moved from paper-based browsing only to online search retrieval.
- We converted a range of finding aids (PRO Guide, Introductory Notes, series lists and some supplementary finding aids) to a single catalogue dataset, describing whole series of records in one place.
- Context was provided by linking and browsing within an archival hierarchy largely based on the General International Standard Archival Description, ISAD(G).

In November 2001, The National Archives released a separate website to deliver digital copies of some of our most popular records. It was called PRO-Online and later rebranded as DocumentsOnline.

PROCAT was renamed '**the Catalogue**' in June 2004 when a rebranded website for The National Archives was launched, after the Public Record Office and the Historical Manuscripts Commission merged to form The National Archives. The cataloguing back office system 'PROCAT Editorial' was not changed.

In 2008 we decided to stop maintaining a **paper catalogue**. This change was implemented in consultation with our Public Services Development department. To facilitate the transition and mitigate impact on our regular readers, we continued printing new accessions, amendments, newly open closed descriptions and Freedom of Information releases until 1 April 2011, when all printing of catalogue paper lists ceased.

Also in 2008, the Catalogue Manager reported to the International Council on Archives (ICA) Standards Sub-Committee in Kuala Lumpur on how our cataloguing practice was moving away from the international descriptive standard ISAD(G) on three fronts.

- The 'non-repetition of information' rule was not being respected (to facilitate retrieval).
- The Scope and Content field was being expanded to allow structured data tags (e.g. surname, forename, place, occupation).
- Access area of description and web delivery options were being developed well beyond ISAD(G).

The committee's verbal response was to go ahead with our descriptive practices and to expand the standard with in-house schemas, but there was no appetite to review or update a standard that was serving the vast majority of the international archives sector well.

In the meantime, between 2005 and 2008, we attempted to deliver a system to enable the seamless ingest of digital records from government departments to a digital archive and a new presentation interface with the **Seamless Flow** programme. A pilot service for born-digital records, **Electronic Records Online** (**ERO**), was released in 2008. At the time, document level descriptions for digital records were migrated into the separate ERO and cross-referred in the Catalogue. It was felt that we did not have the functionality to cater for hybrid record series (series containing both digital and paper records); in such cases the records were split into two separate series – one analogue and one digital.

However, the Seamless Flow model and ERO service were not able to handle high volumes of digital objects; therefore, in 2011, The National Archives started the development of the **Digital Records Infrastructure** (**DRI**) system as a robust alternative to preserve and provide access to digital records at scale.

In late 2009, a cataloguing project had developed a methodology to catalogue websites providing links to the **UK Government Web Archive**. Websites did not lend themselves to file level description under ISAD(G), so the approach was to catalogue them only at series level, including a direct link to the UK Government Web Archive where individual dated instances (snapshots) were then presented. Specific website instances and embedded files were not referenced or catalogued individually.

Cataloguing at series level means that a whole website is described generally, as a group of records. This approach can also work for other special collections or future digital accumulations. This method is scalable and also respects the independent development of the UK Government Web Archive while serving three key purposes:

- the inclusion of websites within our official inventory of public records based on provenance
- a reasonable degree of intellectual control and academic rigour
- the provision of high level findability across formats: a keyword search retrieves website entries together with entries for analogue, digital records and other data sources.

Where datasets were concerned, government datasets were held externally in the **National Digital Archive of Datasets** (**NDAD**) until October 2010, when a cataloguing project created collection level information for their inclusion in the online catalogue and Discovery. Following this, the deliverable dataset files were presented through DocumentsOnline (The National Archives' first service to deliver digitised records). This approach allowed us to cease a contract with an external provider and bring NDAD data in house. Although these datasets are listed and available through our catalogue, we would like to develop new approaches to enhance their presentation and usability in the future.

From 2011, the Catalogue became the first component part of our new **Discovery** service: it brought into one presentation system data from the Catalogue, DocumentsOnline and the Digital Records Infrastructure (DRI). Metadata and catalogue entries from other archives services were ingested later – namely data from the National Register of Archives, Access to Archives (A2A) and the Manorial Documents Register. The old catalogue was finally switched off at the end of April 2013. The National Archives' own name authority data for corporate bodies and personal names was migrated to Discovery in September 2015.

At the time of the PROCAT launch in 2000, the online catalogue had 8 million document level descriptions. Discovery now holds over 23 million information assets for records held by The National Archives.

2. Digital records

In 2014-15, during the Digital Transfer Project, we made changes to several cataloguing practices around referencing, arrangement and presentation of digital records. When making these decisions, internal experts from across a number of fields considered the different nature of several types of digital records. With this in mind, new cataloguing practices started to evolve for each category of digital record.

Digital records and their metadata are now ingested into our digital preservation system (DRI), which holds three types of digital material within what can be defined as the **first generation of digital content**.

Born-digital records: these are records created digitally in the day-to-day business of an organisation, such as word processed documents, PDFs, emails, image files, videos and so on. Our approach to digital preservation has been defined as ‘Parsimonious Preservation’: preserving the original as received and creating access copies in a more accessible format where necessary. A new way of referencing, arranging and presenting born-digital records was formulated during 2014-16.

Digital surrogates: records created by converting analogue material such as paper, microfilm or microfiche to digital images. The paper record remains in the custody of the archive as the original public record. Sometimes microfilm may have been accessioned as the public record but it is generally a surrogate. Digital surrogates are analogue in essence. Hence traditional referencing and ISAD(G), as qualified within The National Archives Cataloguing Standards, apply to these records.

Digitised records: these are the result of analogue material being digitised to a high standard, with provenance metadata captured separately and embedded in each file, in order to become the accessioned public record. The digital version becomes the official record for permanent preservation in lieu of the original analogue source, which would generally be destroyed after five years or deposited elsewhere. Again, these records are analogue in essence, hence traditional referencing and ISAD(G) as qualified within National Archives Cataloguing Standards apply.

Born-digital and digitised records share a new existential challenge: record properties around veracity, accountability, authenticity and integrity –their ‘recordness’ in one word– are not found in the digital object itself, but in the metadata that accompanies it, which becomes inextricably bound with it. Metadata therefore becomes part of the record.

Second generation of born-digital content

It is not difficult to anticipate that an untamed second generation of born-digital content is already accumulating within government departments (and other creators of public records). This scenario has been referred to as the ‘Digital Wild West’ or the ‘Digital Heap’ where the identification of unique, authentic records becomes blurred and rules around expected record-keeping behaviours no longer apply. **Second generation born-digital content** possesses at least one of the following characteristics:

- amorphous accumulation
- lack of clear creator(s) or hybrid creation including non-government parties
- not fixed in time (the end date of the record may be uncertain or set in the future as re-use may bring the record back into activity)
- unreliable provenance
- erratic or broken link between government function and digital accumulation
- suspected duplication
- dubious authenticity
- include corrupted and embedded files
- include objects composed of multiple file formats
- stored in unstructured shared drives.
- the only apparent common characteristic or shared property of the material may be its need to be preserved or 'archived', 'got rid of' or shared together at one point in time.

Examples of second generation born-digital records would include shared drives without meaningful folder and file names or unstructured email servers created in environments where basic record-keeping and information management practices have not been applied. It may well be very difficult to render these accumulations within a traditional online catalogue.

3. The National Archives' Digital Strategy

Our Digital Strategy addresses the challenge of digital records as well as our *Archives Inspire*[2] goal to become a digital archive by instinct and design. This can be done by:

- embracing the 'disruptive digital archive' and the digital transformation of the physical archive
- acknowledging that digital records disrupt archival practice
- strengthening our digital capability and culture.

Cataloguing and descriptive practices have already been disrupted by the first generation of born-digital records. The strategy considers the challenge to develop and adopt an entirely new approach to record description for the second generation of born-digital content, moving away from the international standard ISAD(G) for file level metadata. Probabilistic, contextual and temporally aware description is considered as the way forward.

[2] Archives Inspire 2015-2019 is our four year strategy to think and organise ourselves differently in order to meet the needs of our audiences and face our biggest challenge: digital.

Probabilistic description is about acknowledging in a transparent manner that data is imperfect and embracing uncertainty. We are considering the introduction of confidence ratings in our future metadata for born-digital and other records. This confidence rating might be a combination of computational and curatorial scores. Purely numerical approaches for handling uncertainty might miss human data knowledge gathered around the selection, transfer and ingest processes. For example: there is speculation that a confidence rating might be a combination of a computational score (derived from file match signatures, or checksums) and a human confidence score (a value representing the accuracy of record dates or other metadata)[3].

Contextual description covers several areas:

- the record-keeping system and existing arrangement (or lack of)
- the administrative history of the body or person creating the records
- the creating body or person
- the administrative function
- the custodial and archival history surrounding the record, series or accumulation.

For the second generation of born-digital content, metadata around arrangement and creation might be extracted by computational methods (e.g. for records migrated between successive systems or storage media). On the other hand, metadata around the administrative history, function and custodial matters surrounding the business use of the record may lend itself more to curation.

A new *Records in Contexts standard (RiC-CM), a conceptual model for archival description*, was issued for consultation in September 2016, just as this position paper was being written. We welcome the development of a comprehensive standard designed to integrate the four current descriptive standards[4] together with an ontology to define descriptive entities using linked open data techniques. This piece of work could deliver an improved archival standard and also allow the archival community to begin to move beyond ISAD(G). Whether or not RiC-CM will be able to deliver what digital archives need to operate at scale is still in the air. We contributed detailed feedback on RiC-CM by the December 2016 deadline.

Temporally aware description

The records life cycle model identified boundaries between current and historical records. This intellectual model (in common use) is not helpful in the new digital environment as born-digital records are not necessarily fixed in time. The Records Continuum premise that ‘records are in a state of becoming’[5] offers a much more compelling model to handle and interpret born-digital records. Temporal variation is becoming a new reality for both metadata and digital objects. Different date structures

[3] We have already experimented with linked data and confidence ratings through the Traces Through Time project, for example a name match in ADM 273/18/97 is presented as a ‘strong match’.

[4] The existing standards are: ISAD (G), ISAAR-CPF, ISDIAH and ISDF.

[5] McKemmish, S. Archival Science (2001) 1: 333. doi:10.1007/BF02438901

are already appearing, which has a disruptive impact on current metadata schemas. In addition, official born-digital records already in the archive might in the future be re-used and transformed into new entities. This is completely new territory for us. How we capture, describe and present temporal variation and how we render its different components into a usable interface for the general public are yet to be formulated.

4. Digital Cataloguing Practices

This section summarises our digital cataloguing practices around a number of key elements of description. It includes the six mandatory elements of description in ISAD(G), other ISAD(G) elements that are being used to describe digital records, and other elements not in ISAD(G) that have been added to our catalogue data model as the international standard did not fully serve our business, descriptive and presentation needs.

4.1 Referencing

The purpose of a reference is to uniquely identify the record and to provide a link to the metadata that represents it. **Digital surrogates** and **digitised records** are allocated TNA citable references in the same ‘classic’ style as paper records. However, our classic referencing style did not scale up to cater for born-digital collections, as these have complex folder structures well beyond our paper catalogue hierarchy of up to seven levels of description.

References for **born-digital records** are automatically generated on ingest, using the RFC 4648 Base32 alphabet[6] as the basis of the DRI referencing functionality. We have removed the vowels and the Y (a semi-vowel) from this alphabet in order to avoid the creation of random offensive words as part of our references. We have also removed the Z character, which is used for special purposes. This makes our referencing encoding schema a Base 25 alphabet.

References for digital records have four parts: a creating department code (for example, LEV), a series number (e.g. 2), a forward slash followed by a base 25 encoded character which is a short string of automatically generated letters and or numbers (e.g. CCWS) and a final slash followed by the letter Z. The Z indicates that the reference is automatically generated for a born-digital asset. In this example the reference is: LEV 2/CCWS/Z.

With paper records, references reflect the number of levels within which the record exists. As for paper records, references for digital provide unique identifiers but, unlike paper record references, they do not reflect the number of folders they exist within or the hierarchical relationships between digital records.

When there is more than one version of the same record, the Z can be followed by another slash and a number. For example: a fully closed digital record has a reference ending in /Z. If a redacted (open) version of this closed record were to be created, then

[6] Wikipedia article on the Base 32 alphabet.

the new version would have a reference ending in Z/1. If we had a second redaction, making available more information, the new reference would end in Z/2. Each version or manifestation (including those that might be created by migration to a different file format for preservation reasons) would be assigned a new reference in the same way. When the fully closed record becomes fully open (after the closure period has expired), the content of the closed record is made available without changing its reference. The record and its metadata have not changed; only its closure status has changed. This will also facilitate the identification of the complete record and avoid legacy closed references that can be misleading.

Hybrid series (containing both paper and born-digital files) are no longer split at The National Archives. When we have a hybrid series, paper files display classic TNA references, whereas digital files display the automatically generated references ending in Z. We believe that this helps users to become familiar with references ending in Z and distinguish digital from paper files more easily. A paper reference within the same series would look like LEV 2/1/235.

Former department and former TNA references may also exist for digital records (e.g. the original identifier in an Electronic Records Management System or ERMS). When they exist they should be stored and presented to the user.

As mentioned at the beginning of this section, the purpose of a reference is to uniquely identify the digital record and to provide a link to the metadata that represents it.

4.2 Levels of description

All deliverable born-digital records and their metadata (including redacted versions) are ingested at piece level [7]in a flat structure (to improve findability). DRI (and Discovery) are capable of handling item and sub item levels but these are only being used for digital surrogates and digitised records when required.

The master data set for contextual collection level metadata is managed at department, division and series level using the back office for our analogue catalogue (PROCAT Editorial). DRI only holds the series name for registration purposes.

Subseries and subsubseries levels are not used for digital records.[8] Digital folders are not presented as digital assets in their own right for the following reasons.

- Subseries and subsubseries are not relevant and are largely the result of the conversion of paper lists into online catalogues.
- The conversion of deep nested folder structures into subseries, subsubseries, folder, subfolder, subsubfolder, and so on, would undermine findability and the overall browsing experience.

[7] File level in the General International Standard Archival Description, ISAD(G).

[8] There is one slight exception. For some hybrid series, subseries have been used to separate the digital and paper portions.

Our approach is to treat multilevel folder structures as part of the metadata for each digital record, presenting the parent folder names in the Arrangement field for each digital asset (see 4.4.).

The following definition of the seven levels of description for analogue records at The National Archives may be useful for reference:

Department (Fonds)[9]: A government department, agency or body that creates the records. Formerly known as 'lettercode' and earlier as 'group'. For example: Ministry of Defence.

Division (Subfonds): Administrative section of a department, where this exists. For example: Records of the Defence Chiefs of Staff.

Series: Main grouping of records of the same provenance, with a common function or subject. Formerly known as class. For example: Registered Files prior to 1964.

Subseries: A grouping of related records within a series with a common function or subject. Formerly known as header. For example: Manpower.

Subsubseries: A grouping of related records within a subseries with a common function or subject. Formerly known as subheader. For example: National Service.

Piece (file): This is generally the deliverable unit for both paper and born-digital. It can be a box, a volume, a file, a bundle, a roll or a digital deliverable object. For example: Intake of entrants.

Item: A subdivision of a piece, a smaller unit of description for a part of a piece. Items vary considerably in nature: a file in a box, a letter in a file, a name in a register, a part of a larger file, a docket within a volume, etc. It may exist for paper, digitised records and digital surrogates but not for born-digital. For example: Closed extracts 2 pages.

4.3 Extent and medium (labelled Physical description in Discovery)

At The National Archives this data is only mandatory at the three highest levels of description: at department, division and series level. For paper records, we do not record the number of pages at piece and item level; this is for a matter of resource and consistency with the legacy of existing material.

Information about the number of digital records is provided both at series and at piece level. In addition, open digital records display information on the approximate download size of the presentation copy. In the future we would like to make available additional technical metadata around original size and format.

4.4 Arrangement

The purpose of the arrangement element in cataloguing is to provide information on the internal structure and arrangement of the unit of description. This includes the physical or logical ordering or filing sequence of the records, or how they have been treated by the creator or the archive repository. In some cases the original arrangement may have been disturbed, altered or even lost. The lack or loss of arrangement should also be recorded as part of the arrangement information.

[9] The level terms in brackets are the ISAD(G) equivalent terms.

When born-digital records have been kept in file plans or other filing structures, the names of the parent folders are recorded under the arrangement field for each record, at piece level. This arrangement information displays the name of each folder which could be clickable in the future, offering a provenance trail and enabling contextual viewing of all records originally kept under the same folder. This approach enhances discoverability.

The arrangement information is introduced by a standard form of words that can be seen, for example, in RW 33/XH/Z:

> This born digital record was arranged under the following file structure: RW 33 >> Preservation >> Archival conservation >> Binders >> Limp vellum conservation >> Talk on Conservation of limp parchment books 2002-03-14

Our Digital Strategy highlights the importance of contextual description and contextual understanding. By bringing contextual information into the metadata at piece (file) level, we are strengthening the contextual links between digital records and also facilitating information retrieval, as many digital titles can seem meaningless without context. The arrangement element of description (which is fully indexed for search purposes) has become mandatory for born-digital.

Providing context is an essential part of how the digital archive provides value in the future. We remain open to new ideas or ontologies that may allow us to enhance the contextual information surrounding digital records.

4.5 Title

The purpose of this element in ISAD(G) is to name the unit of description, providing either a formal or supplied title.

At The National Archives, the file names of born-digital records are catalogued as title information. Titles are stored and presented as created; they are not curated by staff or corrected if a member of the public reports a spelling mistake or suggests an enhancement. The reason for this is two-fold:

- We cannot change the original file names as this would constitute tampering with the original record. At the point of ingest into the DRI the filename is present as a separate metadata element and must match the actual file name for validation purposes.
- The original file title needs therefore to be respected, preserved and presented as part of the original metadata because, as mentioned in section 2 of this document, the metadata of born-digital records becomes part of the record itself.

This is one of the areas where our cataloguing practice has evolved. During the second half of 2015-16 we moved away from the practice of always storing and presenting descriptive information for pieces (and items) under the ISAD(G) scope and content field (‘Description’ in Discovery). Initially, file names as extracted by DROID[10] were

[10] DROID is our file format identification tool.

mapped under ‘Description’, following the paper descriptive model. Many of these descriptions contained spelling mistakes or were meaningless, which presented an archival and retrieval challenge.

The change in the choice of cataloguing field from Description to Title allows us to show clearly that born-digital records may not have curated descriptions, only given titles (file names). Just as we do not contemplate the idea of changing the title of a published report on the folder of a paper file, we should not contemplate the idea of changing the title (filename) of a born-digital record in its metadata.

The Title element of description has become mandatory for born-digital and can be supplemented with other information under Scope and content.

4.6 Scope and Content (labelled ‘Description’ in Discovery)

The purpose of this element in our standards is to describe the content and scope of the unit of description so that users can judge the potential relevance of the record.

For born-digital records the nature of the file names under Title limits the ability of end-users to find information. Therefore for born-digital records the Scope and content field may be used to store and display **supplementary narrative descriptions** that may be provided by government departments or curated at The National Archives to facilitate information retrieval in certain cases. ‘Scope and content’ is optional, not mandatory, for born-digital records.

Due to the nature and public profile of records generated by Commissions of Inquiry Inquests or similar public bodies, high volumes of spelling mistakes or errors in titles may tarnish the authoritativeness of our catalogue and make the records difficult to find. In certain circumstances, supplementary descriptions might be provided with a digital archivist carrying out a sanity check, or rectifying glaring mistakes that might attract criticism. This recognizes the fact that once metadata has been transferred, stored and presented as part of Discovery, The National Archives becomes the data owner.

The existence of email attachments within a record should be noted under scope and content at series level and where possible at the file level.

4.7 Dates

Digital records generally offer several metadata elements with dates. DRI can preserve all date metadata as transferred but archivists need to decide what date(s) should be presented to our users as the record creation ‘Date’ in our catalogue. In the future we would like to publish additional digital dates.

Ensuring that we provide accurate creation dates for digital records is not straightforward. Different Electronic Records Management Systems (ERMS) store many automatically-generated dates. Record creation or last modified dates can be overwritten when migrating digital records between systems, and again when exporting or preparing data for transfer to The National Archives[11]. If this were to happen, files might be showing a later date (e.g. the transfer date) and as a consequence the record opening date[12] would be later than legally required.

Government departments are being asked to use software (Teracopy) to minimise the risk of altering the last modified date when transferring files. Ideally the last modified date as automatically extracted and ingested in DRI is chosen for the 'Date' field on Discovery. A decision is currently taken for each digital transfer, however, in order to identify the date that best represents the start and/or end date of the record. As a last resort, estimated dates – based on events, for example – and dates derived from the parent series are used. These are presented in square brackets and an explanation provided in the metadata at series level. The estimated date approach also helps members of the public and record advisers who have already encountered difficulties trying to explain some digital dates that seemed to defy common sense.

In the future we would like to introduce a confidence rating in our catalogue, flagging the level of uncertainty around date and other metadata. Probabilistic date information would need to be computed and not curated for each digital object.

4.8 Physical Condition (Physical characteristics and technical requirements in ISAD(G))

The purpose of this element is to provide information about any important physical characteristics or technical requirements that may affect or limit the use of the record.

In the digital context information about corrupt and partially corrupt files should be recorded under this field. There may be other appropriate uses in the future.

The following form of words is used to describe corrupt digital files using plain English:

> Damaged digital file. Partially missing content
>
> or
>
> Damaged digital file. No content

When we are able to recover some of the corrupted file content a repaired version of the file will be made available. Referencing and hyperlinks for the corrupt and repaired files will follow the style of referencing and links for redacted records and their closed counterparts.

Physical condition information should not be confused with information regarding records not available for presentation or legal reasons, which is handled under Restrictions on Use at The National Archives.

[11] TNA's advice for government departments is to use Teracopy or Secure File Transfer Protocol to preserve time stamps.

[12] Record Opening Date is a TNA descriptive property that does not exist in ISAD(G).

4.9 Note

At The National Archives the Note element combines the two Notes in ISAD(G): Note and Archivist's Note. It has been used to record, for example:

- specialised or other important information not accommodated elsewhere
- sources consulted during the cataloguing process or information derived from external sources.
- acknowledgements of cataloguing funding or the donation of catalogue data by researchers.
- date derivation.

A new use of the Note field has arisen for digital records. When appropriate, we are recording at series level how digital material has been described and whether there are uncertain or potentially inaccurate elements of description.

For example, from ASI 2 Al-Sweady Inquiry Evidence:

> 'The catalogue descriptions used in this series are as they were created and used by Inquiry staff. It should therefore be noted that they are the original file descriptions and have not been subsequently curated as part of their transfer to The National Archives.'

Another example, from WA 11 Welsh Language Board, Welsh Language Policy:

> 'The date for digital records in this series is the 'Last Modified' date as generated by the Electronic Document and Records Management system which held and managed these records.'

Names of staff are not credited or displayed on our catalogue, as our cataloguing output as staff at The National Archives is Crown Copyright. The Copyright Officer (in January 2007 and in February 2013) advised that it is universal government practice not to acknowledge the author of corporate texts, except for instance in the case of a chairperson of a body producing a report and the author of work published as an editor. This advice was confirmed in August 2016. Therefore we do not publish the name of the cataloguer(s) under Note. ISAD(G) suggests that the Archivist's Note is used to explain how the description was prepared and by whom but our policy and practice differ from the 1999 standard.

A strong audit trail capability should allow archives to control the creation, quality assurance, release and update of metadata by cataloguers without having to publish their personal names online. Publishing these names online carries the risk (and burden) of having to handle potential requests to take down personal information.

Finally, the Note field should not be used for new digital metadata that does not have a natural match within the existing descriptive standard. Additional metadata required to describe born-digital records should be added to our schemas, to ISAD(G) or to a new standard that may replace it.

4.10 Custodial History (Archival history in ISAD(G))

Custodial history describes where and how records have been held from creation to transfer to the archive, giving those details of changes of ownership and/or custody that may be significant in terms of authority, integrity and interpretation. For born-digital records, we envisage new uses for this element of descriptions.

Colleagues contributing to the 'Best Guess Guidelines for Cataloguing Born-digital Material' drafted during the UKAD workshop[13] in London in March 2016 reflected that

> 'the archival history of born digital material does not stop at any point, but is a continuous process of preservation actions, such as integrity checking and possibly migration to newer formats or media, carried out both before and after the material is accessioned into the archive'.

At The National Archives, preservation and other technical metadata is kept in our Digital Records Infrastructure, although this is not currently published in Discovery.

In June 2016 we encountered our first piece of practical evidence regarding the continuous nature of the custodial history of digital records. This arose during a mixed transfer of hybrid and digitised material from one creating body. In the only case of this sort so far, a government department took on loan some paper files that had been transferred, catalogued and made available in Discovery some years earlier. Once back at the department, these original papers were digitised, re-used and then misplaced. They reappeared as digitised records in a challenging new digital transfer. After considering a number of imperfect solutions for this exceptional case we decided to proceed in the following manner.

- Accept the images and the automatically generated new references as part of a new record within a digitised accession.
- Copy the scope and content information from the original paper files in the supplementary description field in DRI.
- Keep the new digital filename under title (standard practice).
- Create cross-references between the paper and digitised entries.
- Create custodial history metadata for the paper and digitised entries.
- Keep the dates as published for the original paper records to ensure that the record opening dates meet statutory requirements (rather than using the dates when the scans were created, for example).
- Add an explanatory Note: 'Formerly missing whilst on loan to a government department. Digital copy located in 2016. '

[13] The UK Archives Discovery workshop was led by Jenny Bunn. The outcome of the workshop was published at www.archives.org.uk/about/community/groups/viewbulletin/58-contribute-to-the-best-guess-guidelines-for-cataloguing-born-digital-material.html?groupid=50

This ensured that as much information was made available for the newly digitised records and that there was a clear link between the paper originals and the new digital assets.

4.11 Conditions governing access

We have never treated this ISAD(G) element of description as a single entity at the piece (file) and item levels. To replace conditions governing access (since 2001) we developed four new metadata properties which enabled us to:

- Provide more granular information about conditions governing access.
- Manage our data to meet our statutory requirements around closed and retained records, the Public Records Act, Data Protection Act and Freedom of Information legislation. This enables crucial search functionality for staff and members of the public.

The four properties applied at piece (file) level are briefly described below in sections 4.12 to 4.15.

The National Archives also has a separate system called System for Access Regulation (SAR) which holds details of closed records (both analogue and digital). Only certain open information from this internal system is presented (where relevant) in our catalogue, for example:

- Legal closure exemption under Freedom of Information (FoI)
- FoI decision date (when the Advisory Council approved the closure application)

Therefore we have several very specific metadata properties instead of conditions governing access.

4.12 Closure Status[14]

This is a mandatory property, which indicates whether the record and its description are open or closed. When records are open, descriptions must also be open. Closed records may have open or closed descriptions. This applies to both paper and digital.

When a digital file has been redacted the whole original file becomes a closed record with a unique reference, but an open redacted version is made available with its own reference. For example: JA 418/CBK/Z and JA 418/CBK/Z/1.

4.13 Closure Type[15]

This property identifies the specific conditions that affect access to UK public records (paper and digital). Examples include:

[14] Closure Status is a National Archives descriptive property that does not exist in ISAD(G).

[15] Closure Type is a National Archives descriptive property that does not exist in ISAD(G).

- Open on Transfer
- Normal Closure before FoI Act
- Retained by department under Section 3.4 [Public Records Act]
- Closed until
- Closed whilst Access is Reviewed
- Reclosed in

The last two types were introduced in 2011 as a consequence of the implementation of our Re-Closure Policy.

Closure Type and Closure Code are displayed together in Discovery under the label ‘Access Conditions’. For example: ‘Closed until 2020’.

4.14 Closure Code[16]

This is a number which works in conjunction with the closure type. It indicates the number of years for which a record is closed (e.g. 0, 30, 80, 100) or the calendar year until which the record is closed or retained (e.g. 2020, 2035).

4.15 Record Opening date[17]

The date a closed record will be made available depending on its closure type and code. This information has considerable research value so it is retained in the catalogue and continues to be published online after the record becomes open.

4.16 Restrictions governing access (Restrictions on Use in Discovery)

This is a note to indicate restrictions on the use or reproduction of open records, including copyright and re-use if applicable. For example, in a recent born-digital accession of video recordings UKSC 1/CT/Z: ‘This content is made available under the Open Supreme Court Licence’. A hyperlink takes the user to the official licence webpage.

5 Series level description

Series level description is a crucial step in born-digital cataloguing, as it allows us to note distinct information and evolving professional practices relevant to a whole series, accumulation, transfer or collection.

Common information about the arrangement and referencing of a group of records is provided at series level rather than for every digital piece. The arrangement field at series level is also the appropriate place to record that the original arrangement has been disturbed or that we may not able to restore it. The fact that a digital accumulation

[16] Closure code is a National Archives descriptive property that does not exist in ISAD(G).

[17] Record Opening Date is a National Archives descriptive property that does not exist in ISAD(G).

may not appear to possess any order whatsoever should also be documented at series level.

Some examples of new types of information appearing now at series level have already been mentioned earlier, under section 4.9 Note. The examples below refer to other elements of description.

- Scope and content: ‘Over 9,000 emails within this series contain attachments. These attachments are primarily made up of PDF documents, images and various Microsoft Office files, such as Word documents, Excel spreadsheets, Outlook emails and PowerPoint presentations. The attachments have not been treated as separate records but remain within the emails as the digital record’, (from ILB 2).
- Arrangement: ‘Born-digital records are generally arranged and stored differently to paper records. References for born-digital records are automatically generated and display a 'Z' after a forward slash’, (from RW 33).
- Restrictions on use: ‘BT 95/96-2525 require 3 working days’ notice to produce. There is no restriction on the digital files’, (from BT 95, a hybrid series where 96-2525 are paper files).
- Scope and content: ‘This series is a hybrid series that contains paper and digital files (2525 volumes and 2 digital files)’, (from BT 95).

Analysis of series level entries will provide a picture to help us decide which cataloguing practices should develop into our future guidance and policy. We anticipate that, as the number of digital and hybrid series increases, standard means of describing their characteristics will continue to crystallise.

6 Delivery Options

Enabling access to the records is the key purpose for our catalogue. As ISAD(G) did not offer a granular approach to describe and present access and online delivery information, The National Archives has developed its own model over the years. Initially, users were only able to request copies of records and place advance orders online.

At present, options to download digitised and born-digital records include an image viewer and a full download facility that informs the user of the approximate size of the file. A more recent option provides access to the Discovery video player within our catalogue, although videos are streamed from a third party server. Our delivery options also include a range of links to partner websites when a licensed company has digitised our content.

When paper records are closed, the user has the option to submit a request for the record to be opened under Freedom of Information legislation. If a record has been retained by the creating government department (for administrative reasons), the user is informed of this fact and offered a link to visit the department website to submit a request to the record holder.

Specific delivery options also exist to convey in a transparent manner whether a paper record is missing or misplaced.

There is also a varied selection of access delivery metadata around unavailable and unfit paper records, including: records stored off site that require notice to be produced; records on display in the Keeper's Gallery; records on loan; records to be seen under supervision in the invigilation room, etc. These options are not applicable to born-digital records.

7 Cataloguing and the Digital Records Infrastructure (DRI)

The DRI is a modular and extensible set of networked systems that comprise the digital archive of The National Archives. Its primary purpose is the preservation of digital records and at present this does not include an editable archival catalogue. Part of the metadata ingested alongside digital records does, however include cataloguing information.

The DRI's metadata schema is primarily based on Dublin Core but has been extended to include ISAD(G) elements of description. DRI exports these fields, mapping them to the Discovery BIA schema[18] in order for digital records to be accessible in much the same way as our paper records are via Discovery.

Three types of digital records (born-digital, digital surrogates and digitised records) are ingested into the DRI using specific workflows which are customisable per series. Metadata for born-digital records is extracted, added to by the creating government departments and submitted to The National Archives as a CSV file.

Each series of digital records and their accompanying metadata CSV file undergo a series of fixity, validation and file characterisation checks before they are ingested into the digital repository. Two of the tools used in these processes are developed and maintained by The National Archives: DROID, our file format identification tool, and the CSV Validator which uses our CSV schema language.

Describing our digital preservation infrastructure is not the purpose of this document, Therefore the following information will suffice to conclude this section:

- Technical metadata about the digital records is extracted and archived alongside the digital objects, file system, access and rights metadata. This includes information such as the file format and version of all files and – for image files, for example – the height, width, resolution (ppi: pixels per inch), colour space, etc.

- The DRI has a file substitution service which is currently used for converting archival copies of digital files to compressed copies suitable for presentation on Discovery. For example, JPG2000 images are converted to jpegs and MXF broadcast quality videos are converted to MPEG4 for streaming.

- We are planning the development of a new, unified catalogue back office system to enable the management of metadata for both our paper and digital catalogues.

[18] The BIA schema is an xml file containing the machine readable metadata for our online catalogue, Discovery.

8 The way ahead

We do not know exactly what lies ahead but our ultimate goal is to become a digital archive by instinct and design. One of our strategic priorities for 2017-18 is to lead a transformation in how digital records are managed at scale, from creation to presentation. Archival arrangement and description (which includes metadata management) are core elements of our digital task.

Our digital strategy acknowledges that:

> ‘There is also much that we still don’t know. To manage that uncertainty we will work in an agile way, continuing to adjust our plans as we learn more. We will refine our priorities as we develop, iterating this digital strategy as we move forward.’[19]

Our descriptive practices, standards and schemas will also evolve within the same spirit as we move forward. Over the last year we have provided detailed feedback to colleagues working on the development of a new archival standard (Records in Contexts – Conceptual Model). This engagement and collaboration beyond national boundaries will continue in the future as we advance digital archival practices and standards.

Cataloguing is defined at The National Archives as the process of arranging, assigning, creating or enhancing descriptive information (data) to make records accessible to users. We are engaged in a user experience research project which aims to understand better how users perceive our catalogue and what gaps there are between user expectations and our Discovery service.

So far we have ingested (and made available through Discovery, our catalogue) 17 digital or hybrid record series with over 220,000 born-digital and digitised records[20]. Our experience with the transfer, preservation and presentation of these records has shaped our practice, although we continue to develop and streamline our process to automate this process fully.

We are currently enhancing our digital infrastructure and handling several digital transfers from government departments as business as usual. In addition to this work we want to widen the types of digital records that we can preserve, meet the changing expectations of users in a digital world and develop our own digital capability, skills and culture.

[19] The National Archives Digital Strategy, as agreed by the Board in January 2017, Executive Summary by John Sheridan, Digital Director.

[20] This does not include the 8.9 million entries in Discovery for digital surrogates (copies of original paper records) or the very large datasets of digital surrogate copies published online by our commercial and academic partners.

Appendix: List of digital and hybrid record series

Series for digital surrogates (8.9 million copies of original paper records) are not included here.

Reference	Title	Series Dates	Type	Files
ADM 362	Admiralty: Royal Navy Registers of Seaman's Services	1925-1939	Digitised	77,931
ADM 363	Admiralty: Royal Navy Seamen's Services Continuous Record (CR) Cards	1925-1939	Digitised	29,062
ASI 2	Records of the Inquiry into allegations of human rights abuse of Iraqi nationals by British troops in the aftermath of the 'battle of Danny Boy' (The Al-Sweady Inquiry): Evidence.	2009-2014	Born-digital	5,361
BT 31	Board of Trade: Companies Registration Office: Files of Dissolved Companies. [Also includes 47,620 paper files.]	1855-1995	Hybrid	1,273
BT 95	Board of Trade: Companies Registration Office: Classified Index to Files of Dissolved Exempt Private Companies. [Also includes 2,525 paper files.]	1856-1990	Hybrid	2
ILB 2	Coroner's Inquests into the London Bombings of 7 July 2005.	2007-2011	Born-digital	33,255
JA 418	Infectious Diseases and Blood Policy: Files Submitted in Evidence to the Penrose Inquiry.	1971-2006	Born-digital	287
LEV 2	Inquiry into the Culture, Practices and Ethics of the Press (The Leveson Inquiry): Transcripts and Evidence.	2011-2012	Born-digital	9,675
LEV 3	Inquiry into the Culture, Practices and Ethics of the Press (The Leveson Inquiry): Judicial Reviews and Administrative Records	2011-2013	Born-digital	100
MINT 20	Royal Mint: Registered Files: Annual Series. [Also includes 4,529 paper files.]	1901-1986	Hybrid	7
MINT 33	Royal Mint: Quinquennial Series 3: Digitised Paper Files.	1972-1976	Digitised	411
MINT 34	Royal Mint: Quinquennial Series 4: Digitised Paper Files.	1977-1981	Digitised	931
MINT 35	Royal Mint: Quinquennial Series 5: Digitised Paper Files.	1982-1986	Digitised	969
MINT 36	Royal Mint: Quinquennial Series 6: Digitised Paper Files.	1987-1991	Digitised	703
RG 101	General Register Office: National Registration: 1939 Register [Details for over 40 million individuals from the 66,612 register booklets are available from a partner website]	1939	Digitised	66,612
RW 33	The National Archives: Records of the Preservation and Digital Preservation Departments.	2006-2008	Born-digital	1,454
UKSC1	Supreme Court: Video Recordings of Court Proceedings.	2009	Born-digital	81
WA 11	Welsh Government: Records created by the Welsh Language Board relating to Welsh language policy. [Also includes 14 paper files.]	1997-2014	Hybrid	113
WA 12	Welsh Government: Records relating to the use of the Welsh Language in Education and Communities. [Also includes 2 paper files.]	2008-2010	Hybrid	76
WA 13	Welsh Government: Welsh Language Policy: Records relating to the use of Welsh Language in the Private and Public Sector.	2009-2014	Hybrid	47