# Norfolk Records Committee

| | |
|---|---|
| **Report title:** | **Metadata Migration and Management** |
| **Date of meeting:** | **2 November 2018** |
| **Responsible Officer:** | **Steve Miller – Assistant Director, Culture and Heritage** |

### Strategic impact

The Norfolk Record Office archive management system is a key asset which holds information describing the archives within the Collection, including their provenance and their location. This information consists of around 840,000 records and is the essential gateway to the Collection for both the public and staff.

## Executive summary

The Norfolk Record Office archive management system is a key asset because it holds the information describing the archives in the Collection, their provenance and their location. This information (referred to here as metadata) consists of around 840,000 records and is the essential gateway to the Collection for both the public and staff. Its loss would be catastrophic – a notional risk value of greater than £8 million has been assigned to these records.

This metadata has been built up over many decades and since the late 1990s has been entered into an archive management system, CALM. This system has now been in use for 20 years and enabled us to move from paper to electronic delivery of catalogue information. The Norfolk Record Office now needs to plan its metadata requirements for the next 20 years.

This report examines options which will deliver the following benefits:

- Information contained in archive collections will become far more discoverable and, therefore, more widely used.
- The risk of losing irreplaceable metadata will be negated.
- Archive discovery will improve along with advances in data technology and the semantic web
- Descriptive metadata will be accessible as a research resource in itself
- New audiences will engage with heritage
- Increased income streams to support the service

**Recommendations:**

That the Records Committee:

- **Approves Option C as detailed in this report and supports the submission of a bid for capital funding from NCC. If this is unsuccessful, then it approves the use of reserves to fund the project.**

# 1. Proposal (or options)

INTRODUCTION

1.1. The importance of metadata to archives cannot be understated. Metadata is the essential element that forges the link between collections and the user. The better the metadata, the better the interaction, the research and the evidence. Additionally, good metadata underpins good archival management and service provision.

1.2. The Norfolk Record Office holds this key asset in an archive management system called CALM. The information in the system describes the archives held by the record office, their provenance and their location. This information consists of around 840,000 records and is the essential gateway to the Collection for both the public and staff. Its loss would be catastrophic – a notional risk value of greater than £8 million has been assigned to these records.

Background

1.3 Archivists have always generated metadata, i.e. information about the Collections they hold. Accession registers gathered information on provenance and custody, catalogues provided descriptions of records and banks of card indexes offered a means of navigating these tens of thousands of pages of paper. Much of this basic methodology still holds true. Provenance is still an essential component of archive management. Archives still need to be catalogued before the information they contain can be used by researchers. Access points are still needed to enable places, people, organizations, subjects and functions to be identified.

In the late 20th century, IT provided an opportunity to make this metadata more accessible. Finding aids were entered into searchable databases and web catalogues were launched meaning that users could find more out about what a service held without visiting or writing to staff.

1.4 At the Norfolk Record Office, as in most other archives, a similar path was followed. Twenty years ago, the NRO invested in an archive database system, CALM. By 2016, this system contained all of the Record Office's catalogues; in all more than 840,000 descriptions. This has greatly improved access to information about the Collection, but the job is far from complete. The Record Office now needs to plan its metadata development for the next 20 years.

1.5 The digital age presents some significant challenges in making sure that this metadata remains accessible. However, it also offers numerous exciting opportunities for opening up archives in ways undreamt of in the past, bringing the benefits of using archives, which remain a hugely under-exploited resource, to new and diverse audiences.

CASE FOR CHANGE

The Sustainability Challenge

1.6 As well as providing the means of access to the Collection, the metadata held by the NRO has significant value of two kinds:

i. Evidential Value: Archives are records that require contextual interpretation.

Information on who created them, who held them, how they relate to other records etc., has enormous significance for how they are discovered, interpreted and, ultimately, how much they can be trusted as a reliable source of information.

ii. Financial Value: Estimates show that the cost of recreating a single record in the NRO's catalogue (where this is possible) would be around £10. This means that the records in the catalogue have a total value in excess of £8 million.

1.7 Clearly, ensuring the long-term accessibility and reuse of this metadata is of paramount importance. This data has to be kept and managed for at least as long as the records to which it refers; in the case of archives, this is forever.

Central to the issue of metadata sustainability is interoperability. The data itself has to be in a format where it becomes system agnostic so that over the years it can be easily moved between systems. This rests on adherence to international standards such as those developed by the International Council on Archives. In this way, sustainability of the data can be insured whilst also avoiding tie-ins to a single software provider with its associated risks and costs.

1.8 There is also a significant need to make sure that the NRO is able to provide the information on its Collection in a way which meets growing user expectations. As well as quantity and quality of data the NRO needs to move towards an authority file, linked data approach to its catalogues. Traditionally, archivists have catalogued records and then created indexes to help locate catalogues. To meet the needs of future users, archivists need to take a more structured approach in effect creating catalogues featuring an assembly of controlled linked data terms.

1.9 ***Currently, not all of the NRO's collection metadata is fully compliant with these international standards.***

New Opportunities

1.10 **'**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.'

(*Archives Unlocked: Delivering the Vision, Introduction by Jeff James, Chief Executive and Keeper, The National Archives, 2017*)

1.11 Development of the semantic web, and what is popularly termed AI, offer huge opportunities for making archive metadata accessible. Unlike the World Wide Web which links pages containing information, the semantic web links information in a way which supports the extraction of meaning. For the first time, we need to take into account that future users of our metadata might not be human. And the NRO needs to ensure that its metadata is structured in such a way that it can better exploit this opportunity.

1.12 One of the negative sides of retro-converting catalogues has been that of an increased reliance on free text searching. As the number of catalogue entries increase so do the number of search results. This can be alleviated by more careful use of faceted searching, where choices can be made about the type of entity returned (e.g. being able to choose personal name, corporate name, place name etc. to filter the search results) or, more powerfully, by linking authority records to external sources that help identify the referent and link data around the world.

1.13 ***Currently the NRO has only limited data searchable in this way. The largest category, records linked to place names, contains 227,647 links; however, almost all of these relate to probate records.***

1.14 Currently, the NRO catalogue has around 840,000 entries describing somewhere around 10 million documents. To authoritatively answer the question "Does the Collection **not** contain any information on x?" would require the catalogue to contain many millions, probably billions, of records at transactional level (i.e. names, subjects, places detailed in each records).

1.15 Obviously, this is not something that is ever likely to be achieved, but the greater the quantity of this information in the catalogue the greater the opportunities for access. The NRO need to embrace new ways of working so that it can accumulate this metadata over the long term, something it is already working on thorough volunteer projects, crowdsourcing in King's Lynn and its first collaborative PhD.

1.16 ***However, the NRO's ability to carry out the tasks of both enhancing its data and increasing its quantity is restricted by the software it uses. To make the best of these opportunities it needs to be able to move data more easily, to access back end databases to run queries and updates and to use new tools as they become available.***

GOALS, OBJECTIVE AND BENEFITS

1.17 The NRO's goals are to:

- Ensure the sustainability of its collections metadata
- Provide excellent discoverability of its collection by
    - enhancing existing metadata
    - increasing the amount of metadata
    - enabling reuse of metadata

1.18 This will deliver the following benefits:

- Information contained in archive collections will become far more discoverable and, therefore, more widely used.
- The risk of losing irreplaceable metadata will be negated.
- Archive discovery will improve along with advances in data technology and the semantic web
- Descriptive metadata will be accessible as a research resource in itself
- New audiences will engage with heritage

- Increased income streams to support the service

1.19 It will achieve this by delivering the following objectives:

Objective 1: Provide functional online tools to **enhance discovery**

- Improved interface to the online catalogue
- Access to digitized images and data files via catalogue

Objective 2: Improve the **quality** of its existing metadata

- Links to authority files established
- Ability to run queries on raw data
- Provide functionality for sharing data between systems so that unified catalogues can be created

Objective 3: Increase the **amount** of metadata relating to its holdings

- Enables creation of metadata through new programmes including crowdsourcing, volunteer project, and collaborative PhDs.
- Import of data from projects in a variety of formats including EAD, EAC and CSV
- Deploy entity recognition and extraction software on current descriptions to create new authority records or add data to existing ones

Objective 4: Ensure that its metadata is held in **system agnostic formats** which comply with relevant international standards and which can be moved between systems

- Data is compliant with archive and other related standards (ISAD(G), ISDIAH, ISAAR, ISDF, ISO-8601)
- Data can be exported in reusable formats (e.g., CSV, EAD, EAC and XML etc.)
- Data is stored in such a way that it is accessible using software different to that which was used to create it

## 2. Evidence

OPTIONS APPRAISAL

2.1. Three options have been considered and subjected to qualitative and financial appraisal. These are:

2.2. OPTION A: Do Nothing

This option assumes that the NRO would continue to use the existing CALM database, adding new data as collections were catalogued and, where possible adding new data from projects and programmes.

### 2.3. OPTION B: Do Minimum

This option assumes that the NRO would continue to use its existing system for the time being, upgrading the system as required, and moving to the new web module for its online catalogue. Currently only selected staff have access to the system and this option assumes that additional licences would be purchased.

This option is divided into two sub-options for the financial appraisal. The second (Option B2) includes the costs of data migration in Years 14 and 15.

### 2.4. OPTION C: Migrate data to new system

This option assumes that work is completed on metadata to ensure it is standards complaint and can be exported out of the existing system and imported into a new system, Atom (Access to Memory).

Atom is an open source archive management system originally developed under the auspices of the International Council on Archives (ICA), a UNESCO (the United Nations Educational, Scientific and Cultural Organization). Users include:

- UNESCO Archives
- United Nations Archives and Records Management Section
- The World Bank Group Library and Archives of Development
- International Monetary Fund Archives
- NATO Archives
- The National Library of Wales
- The Borthwick Institute, University of York

The system is standards based and allows import and export of data in a variety of formats. It also integrates with Archivematica (another system developed under the auspices of the ICA) as part of a digital preservation workflow.

It provides access to either authenticated users (i.e. those with log in and password) and general users via a web browser. Authenticated users have differing rights within the system (e.g. read, write, edit, access to certain areas only) whilst general users can only view published information.

Good examples of Atom in use can be found at:

- https://borthcat.york.ac.uk/
- https://catalogue.millsarchive.org/

## QUALITATIVE APPRAISAL

This section looks at each of the options and analyses them against the project objectives.

2.5. ***OPTION A: Do Nothing***

Objective 1: Provide functional online tools to **enhance discovery**

2.6. Existing metadata would remain as it is now. This provides for some discoverability, but will provide no improvements, which means that discoverability will fall even further behind growing user expectations, leading to a decrease in use of the collection. The online catalogue would not be supported and may stop functioning on upgraded servers.

2.7. Objective 2: Improve the **quality** of its existing metadata

2.8. Limited opportunities would be available to enhance existing data. Authority files would remain self-contained within the existing system that would not link to any external sources or provide facetted searching.

There would be no opportunities to access the data outside of the system.

2.9. Objective 3: Increase the **amount of metadata** relating to its holdings

2.10. Some information can be imported into the database which would allow a variety of projects to continue. For example, the volunteer cataloguing of building regulation plans and marriage licences using spreadsheets and then importing the data.

2.11. Objective 4: Ensure that its metadata is held in **system agnostic formats** which comply with relevant international standards and which can be moved between systems

2.12. This option does not address the risk to sustainability of metadata. In the financial appraisal section of this report, two Do Nothing options are presented, one with no costs for data cleansing and ensuring standards compliance and one where this cost is delayed until years 14 and 15. In both options the metadata would remain in same format beyond the next decade.

2.13. ***OPTION B: Do Minimum***

2.14. Objective 1: Provide functional online tools to **enhance discovery**

2.15. Within the next 12 months an upgrade to the new version of the back-end system will be required. This would require the purchases of CALM VIEW which would represent a significant improvement in the user experience when searching the catalogue.

2.16. Objective 2: Improve the **quality** of its existing metadata

2.17. There would be no wholesale improvement in the quality of metadata. Structures of data would remain the same and no external links would be created.

2.18. Objective 3: Increase the **amount of metadata** relating to its holdings

2.19. As with Option A, this would enable some work new work to increase the amount of metadata. Upgrading the system may also provide new opportunities for this. However, the NRO would not be able to get direct access to back-end data and would be dependent on a single supplier.

2.20. Objective 4: Ensure that its metadata is held in **system agnostic formats** which comply with relevant international standards and which can be moved between systems

2.21. This option does not address the risk to sustainability of metadata. In the financial appraisal section of this report, two Do Minimum options are presented, Option 2A with no costs for data cleansing and ensuring standards compliance and OPTION 2B with the cost delayed until years 14 and 15. In both these options the metadata would remain in same format until at least 2032.

2.22. ***OPTION C: Prepare Metadata and Migrate to New System***

2.23. Objective 1: Provide functional online tools to **enhance discovery**

2.24. Whilst the system outlined above would not provide all of the functionality that might be required over the next decade and a half, it represents a significant improvement over what is currently available. The system is being actively developed and, because it is open source, there is the possibility of the user community driving developments.

2.25. Objective 2: Improve the **quality** of its existing metadata

2.26. This option would provide the resources to standardize data and introduce more authority file based metadata creation linked to external sources.

2.27. Objective 3: Increase the **amount of metadata** relating to its holdings

2.28. This option would provide some of the flexibility needed to innovate in the creation of metadata. It would allow access to the back end database where user customised queries and updates could be run.

2.29. Objective 4: Ensure that its metadata is held in **system agnostic formats** which comply with relevant international standards and which can be moved between systems

2.30. The metadata would also be held in an easily exportable format which could be shared with other systems.

## 3. Financial Implications

### 3.1. Financial Options Appraisal

The financial appraisal looks at costs over 15 years to which it applies a Net Present Value of 3.5%. In effect, this means that £100 spent in one year's time has a value today of £96.50, in two years of £93.

For Option A (Do Nothing) the risk associated with not being able to migrate metadata easily still applies. To show the impact of this Option 2 has been subdivided into Option B1, where no work is undertaken to prepare metadata for migration, and

Option B2, where the issue is dealt with in Years 14 and 15.

In all cases, any increase in income from improved metadata availability has not been included.

**OPTION A: Do Nothing Option**

| | |
|---|---|
| Annual: | £6000 |
| | |
| Total Cost: | £90,000 |
| **Net Present Value (NPV)** | **£67,950.00** |

**OPTION B1: Do Minimum Option**

| | |
|---|---|
| Year 1:CALM Upgrade cost | £800 |
| Year 1:CALM View costs | £4,000 |
| Year 1: Four additional CALM licences | £3,200 |
| Annual: CALM support costs | £6,800 |
| | |
| Total Cost: | £110,000 |
| **NPV** | **£85,010** |

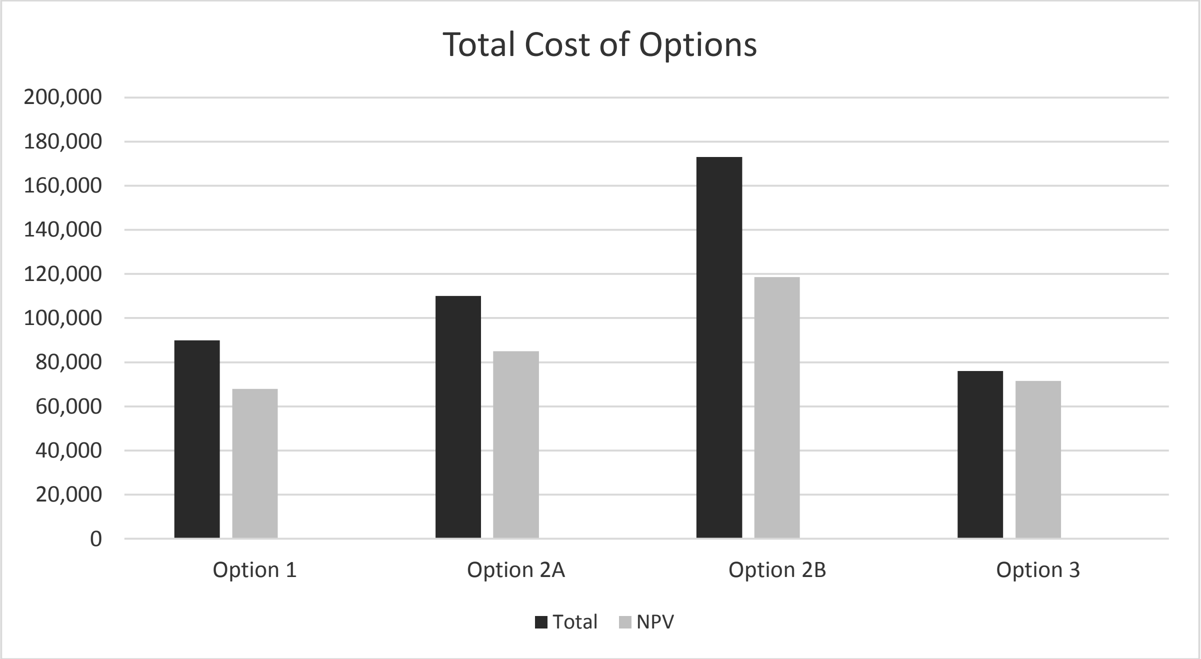
**OPTION B2: Do Minimum with Delayed Metadata Migration Costs**

| | |
|---|---|
| Year 1:CALM Upgrade cost £800 | £800 |
| Year 1:CALM View costs £4000 | £4,000 |
| Year 1: Four additional CALM licences £3200 | £3,200 |
| Annual: CALM support costs | £6800 |
| Year 14: £62K | £62,000 |
| Year 15: £21K | £21,000 |
| | |
| Total Cost: | £188,000 |
| **NPV:** | **£126,785** |

**OPTION C: Prepare Data and Migrate to New System**

| | |
|---|---|
| Year 1: | £42,000 (plus £15K external funding) |
| Year 2: | £21,000 |
| Annual | £1000 |
| | |
| Total Cost: | £76,000 |
| **NPV** | **£71,625** |

3.2. **Comparison of Total Cost and NPV**



**Conclusion**

Clearly, Option 1 (Do Nothing) is untenable as it would result in the NRO catalogue no longer being accessible online and subject the service's metadata to risk.

Whilst Option 2A and 2B would go some way to addressing this risk, they do not offer all the benefits of Option 3.

In terms of the financial appraisal, Option 3 offers considerable savings over Options 2A and 2B.

Option 3 is recommended as the best way to proceed.

## 4. Issues, risks and innovation

4.1. Issues

In October 2018 a 15 month traineeship will start at the NRO as part of the Bridging the Digital Gap project financed by the Heritage Lottery fund and delivered by the National Archives. The trainee will be focussing on metadata creation, storage and management and will be available to support the project.

4.2. Innovation

This project is a significant step in realigning how the Norfolk Record Office makes information on its holdings discoverable. It represents a step change in how this data is assembled and how it can be linked to other resources.

## Officer Contact

If you have any questions about matters contained in this paper or want to see copies of any assessments, e.g. equality impact assessment, please get in touch with:

**Officer name :** **Gary Tuson** **Tel No. :** **01603 222599**

**Email address :** **gary.tuson@norfolk.gov.uk**

If you need this report in large print, audio, braille, alternative format or in a different language please contact 0344 800 8020 or 0344 800 8011 (textphone) and we will do our best to help.

## APPENDIX: Glossary

**AUTHORITY FILE**
A list of controlled terms covering such topics as places, corporate names, personal names, genres (e.g. will, diary, letter), and subjects

**AUTHORITY RECORD**
The authorized form of name combined with other information elements that identify and describe the named entity and may also point to other related authority records.

**EAC**
Encode Archival Context. An XML schema for sharing and moving archive authority files

**EAD**
Encoded Archival Description. An XML schema for sharing and moving archive catalogues

**GENRE**
A class of authority record which covers document type e.g. tithe map, diary, letter.

**ISAAR(CPF)**
International Standard for Archival Authority Records for Corporate Bodies, Families and Persons.

**ISAD(G)**
The International Standard of Archive Description (General). The standard followed by the NRO and most UK archives for the creation of hierarchical catalogues.

**ISDF**
International Standard for Describing Functions of corporate bodies associated with the creation and maintenance of archives.

**ISDIAH**
International Standard for Describing Archive Institutions with Holdings.

**ISO-8601**
An internationally accepted format for storing and displaying Date and Time.

**LOCATION**
Physical location of the documents to which the metadata refers

**NAME**
A class of authority record dealing with personal, family and corporate names

**PLACE**
A class of authority records relating to geographical location

**SUBJECT**
A class of authority record dealing with concepts and events.