The Design And Implementation Of A Necessary and Sufficient System For The Long-term archival Retention of Digital Documents

|  |  |
| --- | --- |
| **Viv Cothey** | **Claire Collins** |
| *Gloucestershire County Council*  *UK* | *Gloucestershire County Council*  *UK*  *claire.collins@gloucestershire.gov.uk* |
|  |  |

**Abstract – we describe how the design of a digital preservation system suitable for the long-term archival retention of digital documents mimics conventional archival practice with regard to provenance, authenticity and workflow. The design further ensures the evidential nature of digital “documents” in the Archive. Exit plans recognize both the routine expiry of time limited supplier agreements and the effects of a disorderly supplier exit. The design has been successfully implemented and is now being migrated from test to production as a “business as usual” component of the archivist role.**

**Keywords – Archive, Provenance, Authenticity, Fixity**

**Conference Topics – Resilience; Community**

# Introduction

This paper emphasizes that the conventional underpinnings of archival procedures and practices as seen in (English) local authority record offices should apply also when accepting custodial responsibility for digital documents.[[1]](#footnote-1) We describe how the experience of ensuring the continuing evidential quality of physical documents has informed the design and implementation of a digital preservation system by concentrating on provenance and authenticity.

The perceived role and purpose of the Archive has evolved as evidenced by The National Archives (TNA) reporting now to the Department for Digital, Culture, Media and Sport. Although being undoubtedly a major part of the heritage sector the Archive is much more than a library of old stuff. Schellenberg carefully distinguishes between the organization, operation and management of a library and that of an Archive [1]. These distinctions are readily apparent when comparing the technology of library catalogues and their standardized bibliographic records and classification schemes with the hierarchical arrangement of archival descriptions [2] and institutionally devised individual arrangements. A significant area of contrast is that, mostly, a library is a collection of published documents whereas an Archive comprises documents that have not been published. The Archive receives documents from its parent body which evidence the bureaucratic systems and decision making which are the activity of that parent. A further feature is that access to many archival documents is restricted, that is, the documents are “closed”.

Whereas the lack of demand for an item may be a cause of concern to a library, for an Archive it is to be expected. Most archival documents will never be requested. However what is important is that every document has the potential to be accessed, possibly by users yet unborn. The existence or lack thereof of a document in the Archive can have life changing consequences for the individual.

Jenkinson [3] asserts that the archivist has a duty to both preserve its intellectual properties, for example how the document relates to other documents, and to protect the physical document. Taken together these two duties can be summarized as preserving the provenance and the authenticity of the document. We know where and why the document was created and that the document is authentic, that is, it remains to be what it purports to be.

Archival procedures and practice address these duties by maintaining archival “provenance” and the archive “strongroom”. Archival provenance is established and maintained by the descriptive entries in the hierarchical catalogue. Access to documents is strictly controlled; they are stored under lock and key in environmentally controlled vaults [4]. Also, document access is mediated including acclimatization and other conservation procedures as necessary.

A special property of the Archive is the evidential nature of its documents. Investigations into the behavior of institutions or the need to revise previous decisions often rely upon the demonstrable authenticity of the document. The information retained and organized in Archives protects people and has legal force. It is not an exaggeration to say that users trust the integrity of information managed by archivists and rely upon it “to hold government and organizations to account” [5]. In a similar vein, Procter [6] says,

“[Archivists] are often unaware of ... the way in which the characteristics of archives – an ability to provide information and evidence and sustain rights – have provided, and continue to provide, the rationale for their maintenance over time.” [emphasis added]

(p xv)

However simply producing an authentic document may not be enough to access its meaning. Many older documents employ either or both an archaic style or language, and an archaic script. Paleography is the study of the script but even a modern transliteration does not remove the need to understand an archaic usage. Words and phrases change their meaning over time and there are traps for the unwary [7].

The design for a system to support the archival retention of digital documents by Gloucestershire County Council (GCC) is based on the need to preserve provenance and authenticity.

# Previous work

Retaining digital information over the long-term Cothey [8] proposes a system architecture to preserve both provenential and authenticity information. The paper introduces the notion of “authentic preservation” which entails the known survival of a digital document. Authentic preservation requires that a) information, including provenance, must survive, b) surviving information must be authentic, and c) authenticity can be demonstrated.

The ‘Archives First’ consortium’s report Further investigations into digital preservation for local authorities [9] documented relevant digital preservation issues and options. In particular the report proposes a long-term authentic preservation architecture that is based on a sequence of interlinked short-term authentic preservation systems. The report therefore draws attention to the need to manage exit plans to successfully transfer curated documents to successor systems.

# Design

The scope of the design is limited. It is assumed that co-lateral digital information hygiene, such as information security, disaster recovery (DR) and business continuity plans are in place and are regularly tested. The main co-lateral threats manifest suddenly and unexpectedly. In contrast the main threats to a long-term retention system are gradual and expected. An important exception here is supplier failure giving rise to a disorderly exit.

The system is considered to comprise two components. The first is “operational”, that is, it forms part of the day to day operational IT of the Archive. Information is dynamic and frequently modified. Like many local authority record offices the operational IT is managed through an outsourced facilities management contract which has a time limited duration.

The second component is the storage of the digital documents, stored information is static and infrequently accessed. Like the operational IT this store is provided and managed via a time limited outsourced facilities contract.

In addition the system workflow is integrated with archivists’ conventional workflows in order to support a business as usual approach.

## Preserving provenance

The preservation of provenance relies on the continued existence and accessibility of the Archive’s hierarchical catalogue. This is threatened by the gradual obsolescence of supporting technologies and by the need to migrate or roll-over the provenential information at the expiration of a management contract. Neither of these threats are mitigated by DR but must be managed through an exit plan. Supplier failure represents a disorderly exit and plans here cannot assume any supplier support. As with DR, exit plans must be regularly tested.

All archival descriptive metadata for digital documents is maintained by this catalogue.

The design response is for the hierarchical catalogue to be ISAD(G) compliant and to ensure that frequent system agnostic information exports are generated. Exported information can be “round tripped” or re-imported to simulate a recreation of the hierarchical catalogue. Importantly this is designed to be achievable without any support from the catalogue provider.

## Preserving authenticity

Digital documents are uniquely fragile. Access to their information is based on reading a stored bitstream which can become corrupted. A demonstration of a lack of corruption is a consistent cryptographic hash or message digest of the bitstream. This is known as a fixity value for the document in question. Different cryptographic algorithms provide supplementary digests.

A sufficient long-term fixity management system is the principal design challenge facing the archival system. The question of fixity arises three times, firstly when the Archive deposits the document into the store, secondly when the Archive requests the document from the store and thirdly when monitoring the integrity of a particular store. A particular instance is verifying every stored document when managing an exit.

As identified above, the Archive protects the authenticity of its physical documents by keeping them under lock and key. For digital documents an equivalent is to maintain a copy of relevant digests independently of the storage supplier. These are retained within the operational IT and are thereby covered by appropriate business continuity arrangements. The three fixity questions are addressed by;

1) on initial deposit the store returns two or more digests which the Archive compares with digests computed independently. These are retained in an operational fixity database.

2) on request the Archive computes two or more digests for the returned document and compares these with the digests retained in the fixity database.

3) periodically the storage system provides digests for a selection of stored documents which the Archive compares with the fixity database.

# Implementation

During early 2021 GCC issued a request for proposals in respect of a storage fixity manager (SFM) having the design features described above [10]. This complemented a similar procurement to implement an ISAD(G) compliant catalogue also with features as described above. The catalogue is now in production and exit plans are being tested. Here we present the implementation of the SFM.

The SFM has been implemented as a Web based service that mediates interactions with multiple cloud storage providers in order to eliminate single provider vulnerabilities and to support sustainable storage provider exit planning. Cloud based commodity storage is used in order to benefit from both economies of provision and geo-diversification. In addition to “upload” and “download” the service provides progressive reporting of current fixity values.

This service is complemented by an independent (open source) desktop digital curation application [11] used by the archivist to both interact with the SFM and to access the operational fixity database. Archivist workflows replicate practice in respect of non-digital documents when accessioning documents and storing them in strongrooms – our business as usual approach. In particular the desktop application also supports the creation of OAIS archival information packages and dissemination packages.

# Conclusion

The successful implementation of the ISAD(G) catalogue, storage fixity manager and desktop digital curation application provides the necessary attributes of an authentic preservation system that includes planning and testing the management of disorderly exits. This authentic preservation system is also sufficient. It mimics existing procedures and practice, in particular mediating access to documents and for working with closed documents. It is anticipated that any future challenges when rendering authentic bitstreams will be addressed by paleographers skilled in archaic digital formats as well as archaic scripts.

# Disclaimer

The views and opinions expressed in this paper do not necessarily represent those of the institutions to which the authors are affiliated.

# REFERENCES

1. Schellenberg T. R. Modern archives: principles and techniques. [Midway reprint] Chicago: University of Chicago Press. 1956.
2. ISAD(G). (1999). ISAD (G): general international standard archival description: adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999.
3. Jenkinson H. A manual of archive administration: new and revised edition. London: Percy Lund, Humphries & Co. 1937.
4. The National Archives. A guide to environmental management of archival material. https://cdn.nationalarchives.gov.uk/documents/information-management/environmental-management.pdf. 2017.
5. The National Archives. Consultation on a new strategic vision for the archives sector. 2016.
6. Procter M. Introduction to the English edition. In Delsalle P. A history of archival practice [Translated and revised by Margaret Procter] London: Routledge. 2018.
7. Shoemaker R. Why Naomi Wolf misinterpreted evidence from the Old Bailey Online. http://www.historymatters.group.shef.ac.uk/naomi-wolf-misinterpreted-evidence-bailey-online/.
8. Cothey V. Retaining digital information over the long-term. https://www.gloucestershire.gov.uk/media/2087704/retaining-digital-information-over-the-long-term.pdf.
9. Cothey V. Archives First: digital preservation: further investigations into digital preservation for local authorities. https://www.gloucestershire.gov.uk/media/2094490/digital-preservation-for-local-authorities.pdf.
10. Forbes H. and Collins C. Request for proposals. https://www.gloucestershire.gov.uk/media/2113461/gloucestershire-archives-request-for-proposals.pdf.
11. Cothey V. SCAT is Curation And Trust. Available from claire.collins@gloucestershire.gov.uk. 2022.

1. “document” is here used in its broadest possible sense to refer to an identifiable unit, that is, regardless of form or extent [↑](#footnote-ref-1)