# Metadata Provenance and Vulnerability

Timothy Robert Hart and
Denise de Vries

**ABSTRACT**

*The preservation of digital objects has become an urgent task in recent years as it has been realised that digital media have a short life span. The pace of technological change makes accessing these media increasingly difficult. Digital preservation is primarily accomplished by main methods, migration and emulation. Migration has been proven to be a lossy method for many types of digital objects. Emulation is much more complex; however, it allows preserved digital objects to be rendered in their original format, which is especially important for complex types such as those comprising multiple dynamic files. Both methods rely on good metadata to maintain change history or construct an accurate representation of the required system environment. In this paper, we present our findings that show the vulnerability of metadata and how easily they can be lost and corrupted by everyday use. Furthermore, this paper aspires to raise awareness and to emphasise the necessity of caution and expertise when handling digital data by highlighting the importance of provenance metadata.*

## INTRODUCTION

UNESCO recognised digital heritage in its "Charter on the Preservation of Digital Heritage," adopted in 2003, stating, "The digital heritage consists of unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are 'born digital', there is no other format but the digital object."[1]

Born-digital objects are at risk of degradation, corruption, loss of data, and becoming inaccessible. We combat this through digital preservation to ensure they remain accessible and useable. The two main approaches to preservation are migration and emulation. Migration involves migrating digital objects to a different and currently supported file type. Emulation involves replicating a digital environment in which the digital object can be accessed in its original format. Both methods have advantages and disadvantages. Migration is the more common method because it is simpler than emulation and the risks can often be neglected. These risks include potential data loss or change, in which the effects are permanent. Emulation is complex, but it offers the better means to access preserved objects, especially complex file types comprising multiple dynamic files that must be constructed correctly. Emulation also allows users to handle digital objects as closely to the "look and feel" as originally intended.[2]

**Timothy Robert Hart** (tim.hart@flinders.edu.au) is PhD researcher and **Denise de Vries** (denise.devries@flinders.edu.au) is Lecturer of Computer Science, College of Science and Engineering, Flinders University, Adelaide, Australia.



Accurate and complete metadata is central to both migration and emulation; thus, it is the focus of this paper. Metadata are needed to record the migration history of a digital object and to record contextual information. They are also necessary to accurately render digital objects in emulated environments. Emulated environments are designed around a digital object's dependencies, which typically include, but are not limited to, drivers, software, and hardware.[3] The metadata describe the attributes of the digital object from which we can derive the type of system in which it can run (e.g., the operating system), the versions of any software dependencies, and other criteria that are crucial for accurate creation of an emulated environment.

While metadata are being used to support the preservation of digital objects, there is another equally important role it should be playing. It is not enough to preserve the object so it can be accessed and used in the future. What of the history and provenance of the digital object? What about search and retrieval functionality within the archive or repository the digital object is held in? One must consider how these preserved objects will be used in the future, and by whom. Preserving digital objects is difficult if adequate metadata is not present, especially if the item is outdated and no longer supported. Looking to the future, we should try to ensure metadata are processed correctly for the lifecycle of the digital object. This means care must be taken at the time of creation and curation of any digital objects because although some metadata are typically generated automatically, many elements that will play a pivotal role later must be created manually. Digital objects also commonly go through many changes, which is something that must be captured, as the change history will reveal what has happened to the object over of its lifecycle. The changes may include how the object has been modified, migrations to different formats, and what software created or changed the object—all of which is considered when emulating an appropriate environment. Examples of these changes can be found in case studies presented in the paper.

## METADATA TYPES

The common and more widely used metadata types include, but are not restricted to, Administrative, Descriptive, Structural, Technical, Transformative, and Preservation metadata. Each metadata type describes a unique set of characteristics for digital objects. Administrative metadata include information on permissions as well as how and when an object was created. Transformative Metadata includes logs of events that have led to changes to a digital object.[4] Structural metadata describe the internal structure of an object and any relationships between components. Technical metadata describe the digital object with attributes such as height, weight, format, and other technical details.[5] Preservation metadata support digital preservation by maintaining authenticity, identity, renderability, understandability, and viability. They are not bound to any one category as they comprise multiple types of metadata, not including descriptive or contextual metadata. However, unlike the common metadata types, preservation metadata are unique from the other metadata types and are often ambiguous.[6]

In 2012, the developers of version 2.2 of the *PREMIS Data Dictionary for Preservation Metadata* saw descriptive metadata as less crucial for preserving digital objects; however, they did state it was important for discovery and decision making.[7] While version 2.2 allowed descriptive


https://doi.org/10.6017/ital.v36i4.10146

metadata to be handled externally through existing standards such as Dublin Core, the latest version (2017) of the dictionary allows for "Intellectual Entities" to be created within PREMIS that can capture descriptive metadata.[8] Thus, while digital preservation does not require all types of metadata, the absence of contextual metadata limits the future possibilities for the preserved object.

Hart writes that because the multimedia objects are dynamic and interactive, and often composed of multiple image, audio, video, and software files, descriptive metadata are increasingly important because they can be used to describe, organise, and package the files.[9] It is also stressed that content description is of great importance because digital objects are not self-describing, which makes identifying semantic-level content difficult; without description metadata, context is lost.[10] For example, without description metadata to provide context, an image's subject information and search and retrieval functionality is lost. Without this information, verifying whether an object is the original, a copy, or a fabricated or fraudulent item is impossible in most cases.

### *Metadata Vulnerability—Case Studies*

Digital objects that are currently being created often go through several modifications, making it difficult to identify the original or authentic copy of the object. Verifying and validating authenticity is important for preserving, conserving, and archiving objects. The Digital Preservation Coalition defines authenticity as

> The digital material is what it purports to be. In the case of electronic records, it refers to the trustworthiness of the electronic record as a record. In the case of "born digital" and digitised materials, it refers to the fact that whatever is being cited is the same as it was when it was first created unless the accompanying metadata indicates any changes. Confidence in the authenticity of digital materials over time is particularly crucial owing to the ease with which alterations can be made.[11]

Tests were undertaken to discover how vulnerable metadata can be in digital files that are subject to change, which can lead to loss, addition, and modification. The tests were conducted using the file types JPEG, PDF, and DOCX (Word 2007). The tests revealed what metadata can be extracted and what metadata could be present in the selected file types. Furthermore, they revealed how specific metadata can verify and validate the authenticity of a file such as an image. For each test, the metadata were extracted using ExifTool (http://owl.phy.queensu.ca/~phil/exiftool/). Alternative browser-based tools were tested and provided similar results; however, ExifTool was selected as the primary testing tool because it produced the best results and had the best functionality. Some of the files tested provided extensive sets of metadata that are too large to include, but subsets can be found in Hart (2009). Note that only subsets are included because some metadata was removed for privacy and relevance reasons. The process and method for each test was conducted in the following manner:

- Case study 1—JPEG
    - Original metadata extracted for comparison
    - Image copied, metadata extracted from copy and examined for changes
    - File uploaded to social media, downloaded from social media, extracted and examined against original
- Case study 2—JPEG (modified)
    - Original metadata extracted for comparison
    - Image opened and modified in photo editing software (Adobe Photoshop), metadata extracted from new version and examined against original
- Case study 3—PDF
    - Basic metadata extraction performed to establish what metadata are typically found in PDF files and what types of metadata could be possible
- Case study 4—DOCX
    - Original metadata extracted for comparison
    - File saved as PDF through Microsoft Word and metadata compared to original
    - File converted to PDF through Adobe Acrobat and metadata compared to original

**Case Study 1**

This case study investigated the everyday use of digital files, the first being simply copying a file. It was revealed that copying a file creates an exact copy of the original file and no changes in metadata aside from the creation and modification time/date. Thus, the copy could not be identified against the original unless the original creation time/date was known. The second everyday use was uploading an image to Facebook. The metadata-extraction tests revealed that the original file had approximately 265 metadata elements. (The approximation is caused by the ambiguity of certain elements that may be read as singular or multiple entries.) These elements included, but were not limited to, the following:

- dates
- technical metadata
- creator/author information
- color data
- image attributes
- creation-tool information
- camera data
- change
- software history

Many of the metadata elements had useful information for a range of situations. Even so, several metadata elements were missing that would require a user input for creation. Once the file had been uploaded to and then downloaded from social media, approximately 203 metadata elements were lost, included date, color, creation-tool information, camera data, change, and software history. It can be argued that removing some of this metadata would help keep user information private, but certain metadata should be retained, such as change and software history. These


https://doi.org/10.6017/ital.v36i4.10146

metadata make it easier to differentiate fabricated images from authentic images and to know which modifications have been made to a file. For preservation purposes, the missing metadata is what may be needed to provide authenticity. This case study aims to make users aware of the significant risk of metadata loss when dealing with digital objects. If metadata are not identified and captured before the object is processed within a repository, the loss could be irreversible.

**Case Study 2**

The second case study revealed how the change and software history metadata can be used to easily identify when a file has been modified. In the test conducted, it was evident by visually comparing the images that changes were made; however, modifications are not always obvious as some changes can be subtle, such as moving an element in the image that completely changes what the image is conveying. The following example displays the change history from the image used in case study 1, revealing how the metadata can easily identify modification:

- **History Action**—saved, saved, saved, saved, converted, derived, saved
- **History When**—The first saved was at 2010:02:11 21:59:05, the last saved was at 2010:02:11 22:12:01 with each action having its own timestamp
- **History Software Agent**—Adobe Photoshop CS4 Windows for each action
- **History Parameters**—Converted from TIFF to JPEG

Further testing was conducted with simple photo manipulation using an original image to see firsthand the issues described in the initial test. The image contained approximately 178 metadata elements, including the typical metadata that were found in the first case study. Once the image was processed and modified with Adobe Photoshop CS5, the metadata were no longer identical. The modified image had approximately 201 metadata elements. The new elements included Photoshop-specific data, change, and software history. However, extensive camera data were lost. It can be argued that the camera data are not important for digital preservation because the lack of it will not hinder the preservation process. However, once the file is preserved and those data are lost, important technical and descriptive information can never be regained. For example, consider a spectacular digital image that captures an important moment in history. If that image is preserved for twenty years, in that time cameras and perhaps photography itself will have advanced dramatically. How digital images are captured and processed might be completely different and will most likely provide different results. Should someone wish to know how that preserved image was captured, they would need to know what camera was used, lens and shutter-speed data, lighting data, and other technical information. Preserving those metadata can be almost as important as preserving the file itself because each metadata element has importance and meaning to someone.

As most viewers of online media are aware, photos are often modified, especially on social media. This is often performed on "selfies," pictures taken of oneself. These can be modified to make the person in the photo look better or to hide features they see as flawed. Small modifications, such as covering some blemishes or improving the lighting have little effect on the image's context, but some modifications and manipulations that can mislead people. These manipulated images often


CC BY

take the form of viral hoax images circulating around the web. For example, Figure 1 displays how two images can be combined into a composite image that changes the context of the image.

**Figure 1.** Composite image. "Photo Tampering throughout History," Fourandsix Technologies, 2003, http://pth.izitru.com/2003_04_00.html.

The two images side by side are original photos taken in Basra of a British soldier gesturing to Iraqi civilians to take cover. In the right image, the Iraqi man is holding a child and seeking help from the solider; as you can see, this soldier does not interpret this as a hostile act. The image above is a composite of the two that changes the story. In this image, the soldier appears to be responding with hostility toward the man approaching. With basic photo manipulation, this soldier who is protecting innocent civilians is portrayed holding them against their will. Images like this circulate through media of all types, and although the exchangeable image file format (EXIF) metadata may not identify what has been done to the image, it would eliminate any doubt that the image has been modified. Unfortunately, these data are not made available. Making users aware of this vulnerability may improve detection of file manipulation at the time of ingest to better ensure only accurate and authentic material is being considered for preservation. Donations received by digital repositories such as libraries must be scrutinised by trained individuals. With this awareness and knowledge of metadata, they can perform their duties to a much higher standard.

### Case Study 3

The PDF metadata extraction provided interesting results. Over a range of tests on academic research papers, the main metadata identified consisted of PDF version, author, creator, creation date, modification date, and XMP (Adobe Extensible Metadata Platform) data. These metadata


https://doi.org/10.6017/ital.v36i4.10146

were not present in every PDF tested; in fact, the majority of PDF files seemed to be lacking important metadata. The author and creator fields were generally listed as "administrator" or "user" and bibliographic metadata was usually missing. However, PDF openly supports XMP embedding, therefore, bibliographic metadata could be embedded into the PDF. Through further testing, bibliographic metadata linked to the PDFs were discovered stored in online databases.

Bibliographic software such as Endnote and Zotero allow metadata extraction, which enables users to import PDF files and automatically generate the appropriate bibliographic metadata. For example, Zotero performs this extraction by first searching for a match for the PDF on Google Scholar. If this search does not return a match, Zotero uses the embedded Digital Object Identifier (DOI) to perform the match. This method is not consistent: it often fails to retrieve any data, and in rare cases it retrieves the wrong data, which leads to incorrect references. Given what we saw happen to metadata when a file is uploaded such as in case study 1 and the nature of a PDF's journey through template selection, editing, and publishing, it is no surprise that metadata are lost or diluted along the way.

**Case Study 4**

The fourth case study conducted on DOCX files provided an extensive set of metadata, some of which are unique to this file type. Creating a new Word document via the File Explorer context menu and attempting to extract metadata resulted in an error as there were no readable metadata to extract until the file was accessed and saved. Once the file had some user input and was saved, the metadata were created and could be extracted. Microsoft Office files contain external XML files that holds information about the document, such as formatting data, user information, edit history, and information about the document's page count, word count, etc. Picture a DOCX file as an uncompressed directory. However, using ExifTool on the DOCX file allowed retrieval of the metadata from all the hidden files.

The metadata included creation, modification, and edit information, such as number of edits and total edit time. Every element within the document (e.g., text, images, tables, etc.) has its own metadata attached that are crucial for preserving the format of the document. The next step in the test involved converting the DOCX file into PDF using the following two methods: (1) converting the document via the "Publish" save option within Microsoft Word; and (2) "right clicking" the document and selecting the option to convert to an Adobe PDF.

The results of the two methods varied slightly. Method 1 stripped all the metadata from the document and generated only default PDF metadata consisting of system metadata (file size, date, time, permissions) and the PDF version, author details, and document details. Method two behaved the same way except that some XMP metadata were created. Both methods resulted in no informative metadata remaining as the majority of the XMP elements were empty fields or contained generic values such as the computer name as the author. All formatting and metadata unique to Microsoft Word was lost. This case study is an enlightening example of what can happen to metadata when a file is changed from one format to another.


CC BY

## HUMAN INTERVENTION

The human element is a requirement in digital preservation as certain metadata, such as descriptive and administrative metadata, can only be created by humans. In fact, as Hart notes, user input is needed to record the majority of the digital preservation metadata. [12] The process can be tedious, as described by Wheatley.[13] One of the examples described included following the processes in a repository from ingest to access, beginning with the creation of metadata and the managerial tasks that are necessary. These tasks include using extraction tools and automation where possible. Using frameworks to record changes to metadata is required, and in some cases metadata must be stored externally to their digital objects. This allows multiple objects of the same type to utilise a generic set of metadata to avoid redundant data. However, although using a generic metadata set is convenient, a large collection of digital objects could be affected if the metadata is lost or damaged.

The human element increases the risk of error drastically because there are numerous steps to metadata creation. Misconduct is also possible. Therefore, the less digital preservation is reliant on humans (and the easier the tasks are that require human input), the better. This can only be achieved by automating most process and training people to ensure they handle their responsibilities accurately, consistently, and completely. Learning the results from the case studies like those described in this paper will better prepare users working with digital objects.

## DISCUSSION

To achieve the most authentic, consistent, and complete digital preservation, institutions must revise their preservation workflows and processes. This entails ensuring the initial processes within workflows are correct before processing digital content. The content must come from a credible source and have its authenticity approved. Participation from the donor of the digital content might be beneficial if they can provide information and metadata about the content. This information could provide additional context for the content as well as identify its history (e.g., format migration or modification). This is not always possible as the donor is not always be the creator of the digital content. If the original source is no longer available, as much information as possible should be gathered from the donor about the acquisition of the content and any information regarding the original source.

This should be considered and carefully monitored throughout the lifecycle of digital content. Granted, if no changes are needed, devices such as write blockers can ensure this as they restrict users and any systems from making unwanted changes or "writes." However, changes are sometimes unavoidable and (although it may not affect the content) detrimental. When changes are required, it is crucial to maintain the digital history by capturing all metadata added, removed, or modified during processing, commonly known as the "change history."

Donor participation should be stipulated in a donor agreement, something that each institution offers to all donors, sometimes in the form of agreements through communication and often with a structured document. Donor-agreement policies differ for each institution: some are quite detailed, allowing donors to carefully stipulate their conditions, whereas others place most of the


https://doi.org/10.6017/ital.v36i4.10146

responsibility on the receiving institution. When dealing with sensitive or historic data of importance, policies should be in place to capture adequate data from the donor. When the content does not fall into this category, standard procedures, which should be present in all donor agreements and institution policies, can be followed. Institutions must also consider when to apply these steps as some transactions between donor and institution can follow standard protocol; others are more complex, such as donations of content with diverse provenance issues.

## CONCLUSION

We have presented four case studies that illustrate how vulnerable digital-object metadata are. These examples show that common methods of handling files can cause irretrievable loss of important information. We discovered significant loss of metadata when uploading photos to social media and when converting a file to another format. The digital footprint left behind from photo manipulation was also exposed. We shed light on the bibliographic-metadata generation of PDF files, how they are obtained, and the surrounding issues.

Action is needed to ensure proper metadata creation and preservation for born-digital objects. Librarians and Archivists must place a greater emphasis on why digital objects are preserved as well as how and when users may need to access them. Therefore, all types of metadata must be captured to allow users from all disciplines to take advantage of historical data in many years to come. Given the rate of technological change, we must be prepared; observing first-hand the vulnerability of metadata is a step toward a safer future for our digital history.

## REFERENCES


[1] "Charter on the Preservation of Digital Heritage," UNESCO, October 15, 2003, http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html.

[2] K. Rechert et al., "bwFLA—A Functional Approach to Digital Preservation," *PIK—Praxis der Informationsverarbeitung und Kommunikation* 35, no. 4 (2012), 259–67.

[3] K. Rechert et al., *Design and Development of an Emulation-Driven Access System for Reading Rooms,* Archiving Conference, 2014, 126–31, Society for Imaging Science and Technology, 2014.

[4] M. Phillips et al., *The NDSA Levels of Digital Preservation: Explanation and Uses*, Archiving Conference, 2013, 216–22, Society for Imaging Science and Technology, 2013.

[5] "PREMIS: Preservation Metadata Maintenance Activity" Library of Congress, accessed March 10, 2016, http://www.loc.gov/standards/premis/.

[6] R. Gartner and B. Lavoie, *Preservation Metadata (2nd Edition)* (York, UK: Digital Preservation Coalition, 2013), 5–6.

[7] PREMIS Editorial Committee, *PREMIS Data Dictionary for Preservation Metadata, Version 2.2* (Washington, DC: Library of Congress, 2012), http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

[8] PREMIS Editorial Committee, *PREMIS Schema, Version 3.0* (Washington, DC: Library of Congress, 2015), http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

[9] Timothy Hart, "Metadata Standard for Future Digital Preservation" (Honours thesis, Flinders University, Adelaide, Australia, 2015).

[10] J. R. Smith and P. Schirling, "Metadata Standards Roundup," *IEEE MultiMedia* 13, no 2 (April-June 2006): 84–88.

[11] "Glossary," Digital Preservation Coalition, accessed August 5, 2016, http://handbook.dpconline.org/glossary.

[12] Timothy Hart, "Metadata Standard for Future Digital Preservation" (Honours thesis, Flinders University, Adelaide, Australia, 2015).

[13] Paul Wheatley, "Institutional Repositories in the Context of Digital Preservation," *Microform & Digitization Review* 33, no. 3 (2004): 135–46.


https://doi.org/10.6017/ital.v36i4.10146