

# Novice to Know-How Module Text

## Course 5: Ingesting Digital Content

# Module 2: Metadata for Digital Preservation

The development of this course was funded by The National Archives (UK) as part of the "Plugged In, Powered Up" digital capacity building strategy.

## 1. Introduction.

The creation and capture of metadata and documentation is an essential part of digital preservation. In this module we will introduce the types of metadata and documentation needed, digital preservation metadata standards, and options for organizing and storing metadata.

Before we start in earnest, one important thing to remember when planning your approach to metadata for digital preservation is to keep it manageable. Creating complex metadata structures to cover all eventualities is tempting, but can end up being a barrier to progress. Many digital preservationists recommend a "minimum metadata" approach. Identifying what is really essential to reduce the time needed to create and manage metadata.

## 2. Metadata for Digital Preservation.

Metadata serves many purposes in long-term preservation, providing a record of activities that have been performed upon the digital content, a basis on which future decisions on preservation activities can be made, as well as supporting discovery and use.

The information contained within metadata often encompasses a range of topics. There is no clear line between what is preservation metadata and what is not, but ultimately the purpose of preservation metadata is to support the goals of long-term digital preservation, which are to maintain the discoverability, identity, renderability, understandability, and authenticity of digital content over long periods of time.

## 3. Documentation for Digital Preservation.

Documentation is the information (such as software manuals, survey designs, data dictionaries and user guides) provided by a creator and the repository that supplements the metadata and provides enough information to enable the resource's use by others.

It is often the only material providing insight into how digital content was created, manipulated, managed and used by its creator and it is often the key to others making informed use of the resource.

## 4. Digital Preservation Issues Addressed.

There are a number of issues which make metadata and documentation critical for the continued viability of digital content and they relate to fundamental differences between analogue and digital resources:

**Technology** - Digital content is dependent on hardware and software to render it on screen. Recording technical requirements allow decisions on appropriate preservation and access strategies to be made.

**Change** - While analogue materials may be preserved by predominantly passive preventive preservation, digital content will be subject to repeated actions, and there will be many different people and quite possibly different institutions influencing the management of digital content over time. Recording actions taken on a resource and changes occurring as a result will provide a key to future managers and users of the resource.

**Authenticity** – The use of metadata (checksums) in integrity checking is the major means of establishing the authenticity of digital content.

**Rights Management** - While analogue resources may or may not be copied as part of their preservation program, digital content must be copied if it is to remain accessible. It may also be changed, e.g. migrated to a new format. Managers need to know that they have the right to copy or change files for the purposes of preservation, what (if any) technologies have been used to control rights management and what (if any) implications there are for controlling access.

**Future Re-Use** - It may not be possible for others to use the material without adequate documentation and metadata. It will tell us things like how to use the digital content, what the information means, how it is structured, and what context it was created in.

**Cost** - It is expensive to create metadata manually and preservation metadata may not always be easily generated automatically. Additional metadata for digital preservation needs therefore requires careful cost/benefit trade-offs.

## 5. Capturing Metadata and Documentation.

When starting out in digital preservation it can be useful to focus efforts on capturing metadata and documentation at two levels:

1. At a file level, generating as much metadata as necessary using characterization tools (e.g. DROID). This should cover information such as file formats, size, location, and checksums for integrity checking.
2. At a high-level, such as collection-level, to capture information including provenance, rights, ownership, risks, value, and retention.

One possible approach to capturing file level information is the creation of Verifiable File Manifests for groups of digital content, and at the higher level, a document called a Digital Asset Register. We will cover in detail the creation of these two types of documents in future modules.

## 6. Introducing PREMIS.

When considering good practice for digital preservation metadata, it is useful to look at implementing established standards, one of the most widely used is PREMIS. PREMIS (PREservation Metadata: Implementation Strategies) is a standard for metadata to support the preservation of digital objects, manage rights, and ensure long-term usability.

Developed by an international team, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems.

PREMIS does not aim to cover all metadata you might need to manage and provide access to digital content. It focuses on preservation processes and rights but it can be used in conjunction with other standards, e.g. cataloguing standards like Dublin Core or ISAD(G). A link to the PREMIS standard is included in the course resources.

## 7. Implementing PREMIS.

PREMIS defines which of the semantic units within the standard are mandatory or optional (and even those that are mandatory are not necessarily so in every circumstance). This means that there is quite a lot of flexibility to allow you to tailor your PREMIS implementation to your local context.

PREMIS is a large and complicated standard, so if you do decide to implement it it is wise to remember the advice from earlier to only aim to use the parts of the standard that you deem to be necessary. You do not want to create an overly burdensome approach to metadata.

As mentioned most commercial digital preservation systems allow for PREMIS implementation, but they too allow customization of metadata requirements to meet your needs.

## 8. Introducing METS.

Another useful standard to be aware of is the Metadata Encoding and Transmission Standard (METS). METS is a standard for encoding descriptive, administrative, and structural metadata for digital content. It is implemented using an XML-based schema.

Using METS provides a standardized method for structuring metadata. METS is complimentary to PREMIS and is sometimes used as a wrapper for PREMIS metadata. It can be implemented with the help of tools such as Bagger, which creates a "bag" for digital content which will include the necessary files and their accompanying metadata and documentation.

## 9. Module Wrap-Up.

Metadata and documentation are essential parts of digital preservation. They provide the information we need to:

- Manage digital content over time
- Make informed preservation decisions
- Provide meaningful access

We do, however, need to be pragmatic in our approach to metadata and documentation so that we do not set requirements that are ultimately a barrier to success. Standards like PREMIS and METS can help guide us in our choices about the metadata and documentation to capture, but don't let them be a barrier to starting to make progress in digital preservation.

In the next two modules we will look at two useful ways to capture metadata about digital content: verifiable file manifests and digital asset registers.