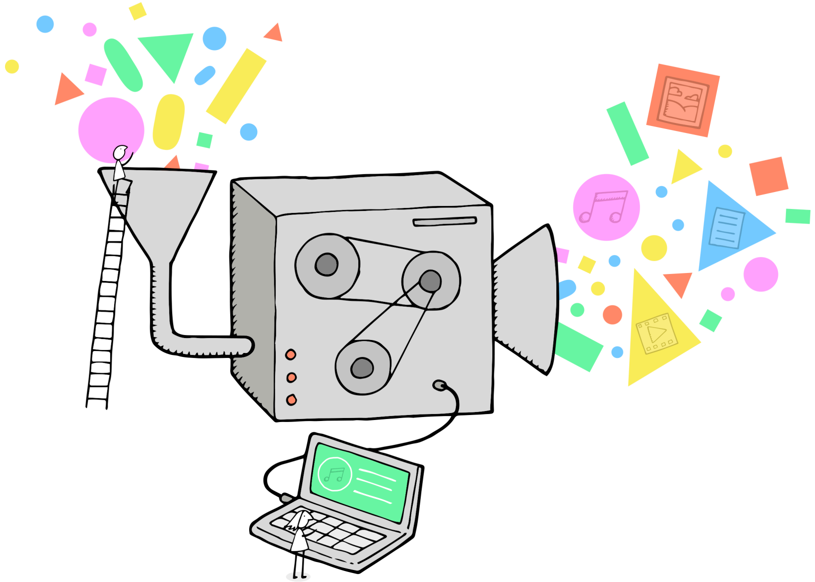
Introduction to Digital Preservation

Exercise Booklet & Glossary

Bodleian Libraries
23/03/2018
Version 1.0



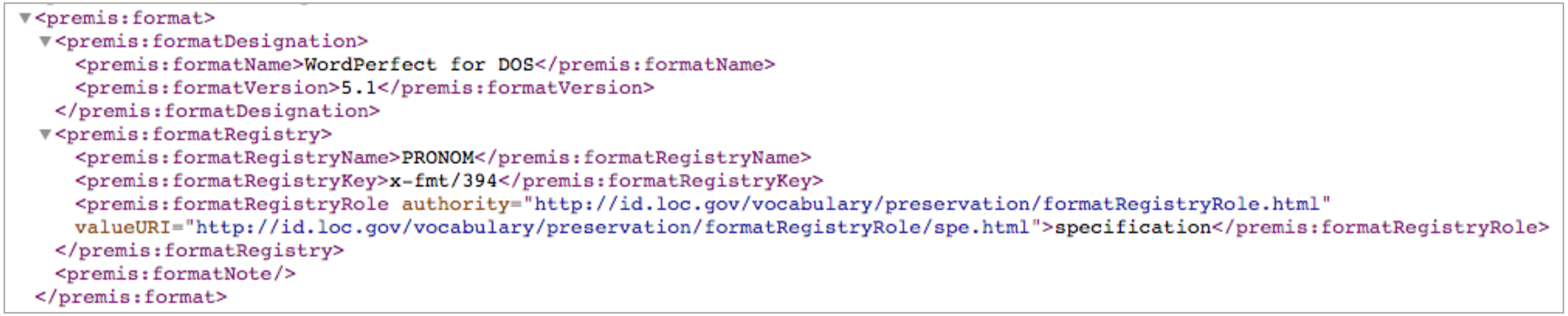
Digital Preservation Business Case Toolkit http://wiki.dpconline.org/, CC-BY-NC 3.0

Activity: Stories of loss

Share you own story of digital loss; this can be work-related, third-party or personal. Use this space to make notes if you wish, but remember that the stories we are sharing might be confidential though important. Please consider the privacy of others when taking notes. Loss of digital materials can happen any time or anywhere. Remembering these losses and why they happened is important to stop mistakes from being repeated in the future.

Activity: Preservation metadata quiz

1. When would this kind of metadata be important? Write an example below.



```
<premis:format>
  <premis:formatDesignation>
    <premis:formatName>WordPerfect for DOS</premis:formatName>
    <premis:formatVersion>5.1</premis:formatVersion>
  </premis:formatDesignation>
  <premis:formatRegistry>
    <premis:formatRegistryName>PRONOM</premis:formatRegistryName>
    <premis:formatRegistryKey>x-fmt/394</premis:formatRegistryKey>
    <premis:formatRegistryRole authority="http://id.loc.gov/vocabulary/preservation/formatRegistryRole.html"
    valueURI="http://id.loc.gov/vocabulary/preservation/formatRegistryRole/spe.html">specification</premis:formatRegistryRole>
  </premis:formatRegistry>
  <premis:formatNote/>
</premis:format>
```

(Source: Library of Congress)

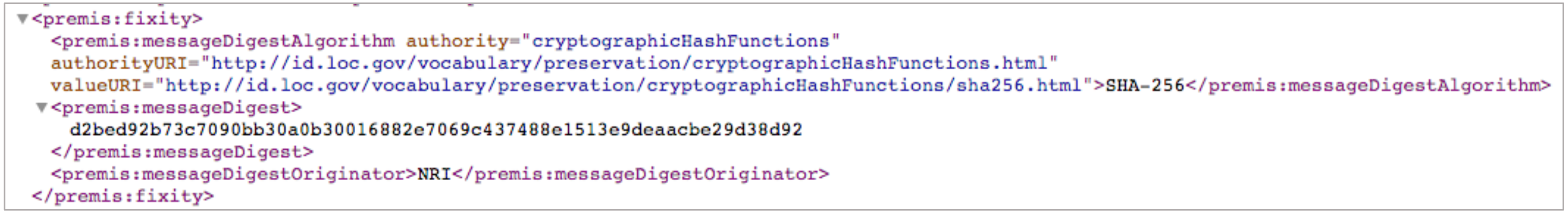
2. As part of a project, about 10,000 theses were digitized and made available online over a number of years. Due to time constraints, the images were not quality checked against the originals. A few years later, a user complained the numbers in some of the calculations as seeming incorrect, calling the work into question.

   A librarian looked into the PDF and found that the numbers had actually been changed from the original. Further research uncovered that this was due to the type of scanning hardware used to create the PDF files and the compression it was applying during image capture. It switched some of the numbers. Further look into other digitized theses showed that not all of the theses used the same scanning hardware and not all of the PDFs also has the original TIFF files.

   What kind of preservation metadata would allow librarians to find the affected PDFs and make decisions about what to do with them? Without this metadata, what are the repercussions?

3. Why would storing the file size in preservation metadata be useful, especially when you can already check the file size in ‘Properties’ of a file?

4. What is this preservation metadata telling us and why is it important?



```
▼<premis:fixity>
   <premis:messageDigestAlgorithm authority="cryptographicHashFunctions"
   authorityURI="http://id.loc.gov/vocabulary/preservation/cryptographicHashFunctions.html"
   valueURI="http://id.loc.gov/vocabulary/preservation/cryptographicHashFunctions/sha256.html">SHA-256</premis:messageDigestAlgorithm>
  ▼<premis:messageDigest>
     d2bed92b73c7090bb30a0b30016882e7069c437488e1513e9deaacbe29d38d92
   </premis:messageDigest>
   <premis:messageDigestOriginator>NRI</premis:messageDigestOriginator>
 </premis:fixity>
```

(Source: Library of Congress)

5. **OPTIONAL:** If you work with a digital collection, what kind of preservation metadata do you think would be useful to help preserve your digital collections and why? If you do not work with digital collections, think of a digital collection and use that for your example.

Digital Preservation Glossary

Authenticity – A digital object is authentic if it “is what it purports to be”. In the case of digital 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.

Bit – A bit is the basic unit of information in computing. It can have only one of two values commonly represented as either a 0 or 1. The two values can be interpreted as any two-valued attribute (yes/no, on/off, etc).

Bit stream – A stream of data in binary form. A bit stream may be a digital file or a component of a digital file. The term bit stream is particularly important in fields such as audiovisual archiving. **see also “digital file”, “digital object”, and “digital material”*

Born-digital – Digital materials which are not intended to have an analogue equivalent. This differentiates born digital material from digitized material, as it has not been created from an analogue source.

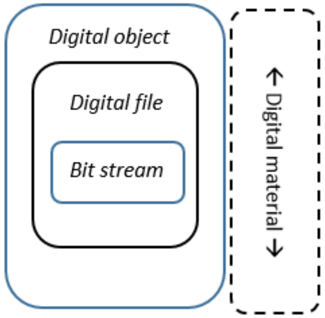
Characterization – Characterization is the identification and description of what a file is and of its defining technical characteristics. Characterization may include the identification of file formats and technical attributes such as creating software and hardware, file size, bit depth etc. Characterization is often captured as technical metadata.

Checksum – A unique numerical signature derived from a file. Used to compare copies. Sometimes referred to as a digital fingerprint due to the fact that it is meant to be unique for each digital file. Common checksums are: MD5, SHA-1 and SHA-256.

Digital archiving – This term is used very differently within sectors. The library and archiving communities often use it interchangeably with digital preservation. Computing professionals tend to use digital archiving to mean the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation.

Digital file – Binary information that is available to a computer program.

Digital materials – A generic term which can refer to either a *Digital File or* to a *Digital Objec.t *see also “Digital file” and “Digital object”*



Digital object – A conceptual term that describes an aggregated unit of digital content comprised of one or more related digital files. These related files might include metadata, master files and/or a wrapper to bind the pieces together. **see also "digital file", "digital material" and "bitstream"*

Digital preservation – Refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly for the purposes of this study and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organizational change. Those materials may be records created during the day-to-day business of an organization; "born-digital" materials created for a specific purpose (e.g. teaching resources); or the products of digitization projects. This Handbook specifically excludes the potential use of digital technology to preserve the original artefacts through digitization. **See also* ***"****Digitization" below.*

Digitization – The process of creating digital files by scanning or otherwise converting analogue materials. The resulting digital copy, or digital surrogate, would then be classed as digital material and then subject to the same broad challenges involved in preserving access to it, as "born digital" materials.

DROID – A file profiling tool developed and distributed by TNA to identify file formats. Based on **PRONOM**.

Emulation – A means of overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers.

File format identification – the process of identifying a file format type based on the file signature outlined in the file format specification, the container signature or the file extension.

Fixity – The state of a digital file has remained unchanged or unaltered.

Fixity check – The process of ensuring that digital files have not been changed without prior authorization. Changes to files may occur due to human error or transmission errors. Is sometimes referred to as an integrity check. **see also "Checksum"*

JHOVE – A validation and characterization tool for digital materials.

Master file – A master file is a source file from which subsequent file versions can be created. An example of a master file could be a high quality TIFF file used for deriving JPEG access copies from. For this reason, preservation effort is generally targeted at master files, rather than derivative files which can be regenerated from the source.

Metadata – The set of information required to enable content to be discovered, managed and used by both humans and automated systems.

Migration (file format) – A means of overcoming software obsolescence, by converting files into formats which the hosting institution is able to support and render. File format migration is also referred to as "file format conversion" by some groups within the University of Oxford and can be used interchangeably.

PREMIS Preservation Metadata: Implementation Strategies. A de facto standard for digital preservation metadata. **http://www.loc.gov/standards/premis/**

PRONOM – A database of file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value. Used with **DROID**.

Validation – The process of ensuring that data is correct and useful when checked against a set of data validation rules. These might include rules for package or file structure or specific file format specifications.

Open Archival Information Systems (OAIS) Reference Model Terminology

Access – The process of end users locating, requesting and receiving digital materials from the OAIS archive.

Administration – The management of the day-to-day operations of the OAIS archive. It includes communicating with external stakeholders and users of the OAIS archive.

Archival Information Package (AIP) – An Information Package, consisting of the complete set of digital files and a complete set of metadata for the AIP (to support preservation and access) that is preserved within an OAIS archive.

Archival Store – The storage and maintainence of digital materials in the OAIS archive.

Consumer – The individuals, organizations or systems that locate, request and use the digital materials stored by the OAIS. A specific subset of Consumers is known as the Designated Community. *See also "Designated Community"*

Data Management – The maintainence, creation and alteration of databases of descriptive metadata for the digital materials stored within the OAIS archive. Administrative metadata the supports that OAIS' operation is also managed in databases under this function.

Designated Community – An identified group of potential consumers who should be able to understand a particular set of information from an archive. These consumers may consist of multiple communities, are designated by the archive, and may change over time.

Dissemination Information Package (DIP) – An Information Package, derived from one or more Archival Information Packages (AIPs), and sent by Archives to the Consumer in response to a request to the OAIS archive.

Ingest – The process of turning a Submission Information Package (SIP) into an Archival Information Package (AIP) and then putting that information package into the archive.

Management – The stakeholders responsible for the strategic planning and policy development for the OAIS archive.

Open Archival Information System (OAIS) Reference Model – A conceptual framework describing the environment, functional components, and information objects associated with a system responsible for the long-term preservation. As a reference model, its primary purpose is to provide a common set of concepts and definitions that can assist discussion across sectors and professional groups and facilitate the specification of archives and digital preservation systems. It has a very basic set of conformance requirements that should be seen as minimalist. OAIS was first approved as ISO Standard 14721 in 2002 and a 2nd edition was published in 2012. Although produced under the leadership of the Consultative Committee for Space Data Systems (CCSDS), it had major input from libraries and archives.

Preservation planning – The act of monitoring the external environment for any changes or risks to digital materials in the OAIS archive. This includes responsibility for the creation and maintainence of a preservation strategy.

Producers – The individuals, organizations or systems responsible for submitting and transferring SIPs for the creation of AIPs to be placed into (ingested by) the OAIS archive.

Submission Information Package (SIP) – An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more Archival Information Packages (AIPs) and/or the associated Descriptive Information.

Sources:
Digital Preservation Coalition Handbook, 2nd Edition. Available: https://dpconline.org/handbook/glossary
GLAM glossary. Available: https://ox.libguides.com/digitalpreservation