November 2022

apac Association of Performing Arts Collections

Digital Preservation Series

PART 4: File Formats

Prepared by APAC Digital Preservation Working Group

Table of Contents

INTRODUCTION

As has been touched on in earlier parts of this guidance, file formats are very important in the digital preservation process. Gathering information on which file formats you hold is key to the planning of your subsequent digital preservation processes. This section of the guidance will take you through how to audit your file formats, familiarise yourself with performing arts specific file formats or challenges, identify which assets may be at high risk due to their formats and point you towards useful resources in this area. This section will focus on file formats pre-preservation.

UNDERSTANDING YOUR FILE FORMATS

The first step in this process is to understand which file formats you hold. Creating a comprehensive list of all the formats in your repository is a good place to start. There are several free tools you can use to undertake this process on your files including DROID and Jhove2. The National Archives format registry PRONOM provides authoritative information relating to over 1,400 formats and file extensions. The International Council on Archives also has a useful page on file format analysis tools available here including their main features and operating system requirements. Once you have a clear understanding of the formats in your repository, you will be in a better position to move forward and identify the make-up of your collection, those formats that are most at-risk and those that are more stable. Having a definitive list of your file formats is likely to assist with advocacy to senior management for long-term support with digital preservation. IT departments will also be in a better position to provide you with advice if they are aware of the formats you wish to preserve and provide access to.

PERFORMING ARTS SPECIFIC FORMATS

Performing arts archives are likely to hold a wide range of formats within their collections. Alongside standard document formats, performing arts archives may also have a higher volume of audio and video formats, in the form of performance recordings or recordings of talks and events. The types of audio-visual formats likely to be found in performing arts collections were covered in part 3.4 of this guidance. It is worth highlighting the lack of an agreed standard within archives for the preservation of video. The International Association of Sound and Audiovisual Archives have however produced some useful technical guidelines in this area.

As the recording and transmission of live performing arts (e.g. NT Live, RSC Live) becomes more sophisticated, archives may be asked to store a greater range of technical AV file formats, including rushes, and those that cannot be rendered without specific software, for example Avid media files. It may be useful to discuss any unfamiliar formats with the record creator in order to understand how they are used and any software dependencies they have, particularly if you would like to render them for access. Alongside specific file formats required for filming live performances, performing arts archives may also hold a range of file formats relating to technical performing arts work, such as digital drawings and Computer Aided Design (CAD) files. Other areas to consider include internal databases, such as a staff intranet, email, and social media content.

WHAT TO DO WITH HIGH RISK ASSETS

All digital files are to some extent 'high-risk' as they are all vulnerable to both obsolescence and proliferation. File formats can become out of date quickly and multiple generations of the same format can proliferate within an organisation's storage systems. It can quickly become difficult to identify for organisations which files are most 'at risk'. The Digital Preservation Coalition (DPC) defines three areas that can constitute 'risk' to a digital file:

Content Type - where the risk is caused or aggravated by the structure of the content (e.g. sound and vision, gaming, social media)

Media Type - where the risk is caused or aggravated because of the media used (e.g. portable media such as magnetic or optical media)

Context - where the risk is caused by the context of production and use (e.g. research outputs, legal records, engineering data).

The DPC outline a number of risks of failing to preserve 'at risk' digital content:

- Data loss/corruption
- Disruption of business/organisation function
- Damage to reputation
- Unable to meet regulatory requirements

There are a number of techniques that can be employed in order to guard against the above:

Bitstream Copying - This is the most basic form of preservation and consists essentially of making a copy of your data. This guards against hardware failure and malicious/accidental damage caused to the original file but does nothing to guard against format obsolescence.

Migration - Migrating a digital file from an unstable format to a more stable one is a popular technique to ensure authenticity and continued access. The PRONOM database maintained by The National Archives is a useful tool containing information on potential migration pathways for certain file formats. If possible, the original file should be saved along with the migrated file as there may be a loss of characteristics or functionality when the file is migrated.

Popular software used to undertake migration: FFMPEG (For AV content), ImageMagick (for images), Ghostscript (for PDF), LibreOffice (for word processed files).

Emulation - This reproduces the behaviour of certain files in a computer of different design - so files from obsolescent machines can run on contemporary hardware. There is a strong benefit of being able to access a digital file within its original context, however there are legal concerns with regard to emulation. For example, obtaining the relevant licences to use legacy software.

Normalisation - All files of a certain type e.g. images, are converted into a single agreed format that is agreed to be the most 'stable'.

USEFUL RESOURCES ON FILE FORMATS

Digital Preservation Coalition handbook - file formats

Digital Preservation Coalition Bit List - Endangered digital content

PRONOM - The National Archive registry

Also, please see the table on pages 6-7 of guidance Part 3 for brief explanations of relevant file formats.

Acknowledgements

Acknowledgement for the creation and development of this digital preservation series go to:

Erin Lee - National Theatre
Malcolm Mathieson - National Theatre
Arantza Barrutia-Wood - University of Sheffield
Bethany Johnstone - University College London
Robyn Greenwood - The Royal Shakespeare Company

with special thanks to the wider APAC digital preservation working group.

Stay tuned for

Part 5: Workflows & Mapping against NDSA Standards

to be published Winter 2022/23

Contact

info@performingartscollections.org.uk
@apac_ssn
performingartscollections.org.uk