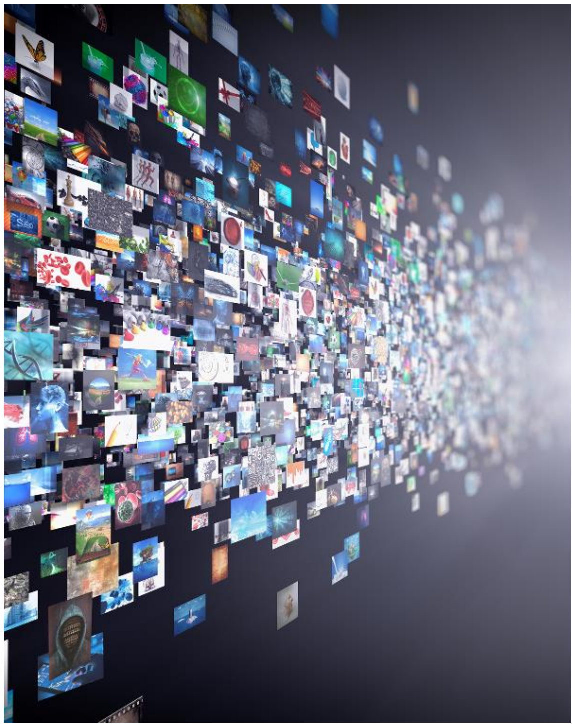


DIGITAL DATA ARCHIVING

A guide to permanent access through digital format preservation

ABSTRACT

There are many challenges involved in digital data archiving. Here we examine format monitoring and migration and consider an alternative: format preservation.

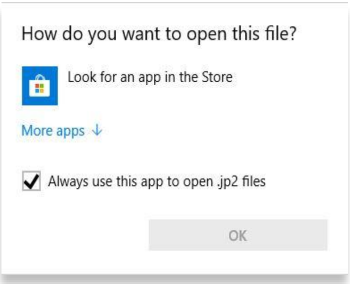
Maja Bystrom

CEO, Bevara Technologies, LLC

Introduction

While we may not be able to – and may not want to – archive all of the zettabytes of data (Cave, 2017) produced per year, it is certainly vital to preserve some of those data. Thus, the challenges of a successful archiving program include: determining which data to save, choosing the appropriate metadata, determining the most useful archival format, hosting the data in reliable system, and maintaining the content. Here we address the third issue: how to determine the most useful and cost-effective archival format and consider the challenges and work-arounds to format obsolescence.

Format obsolescence, or lack of format support, is pervasive. Even "reliable" formats may not always be accessible. Consider JPEG2000, which is recommended by NARA (National Archives and Records Administration, 2018) as an archival image format. Popular platforms, such as Microsoft's Windows 10 OS, often don't come with applications to display JPEG2000 (JP2) images (see Figure 1).


How do you want to open this file?

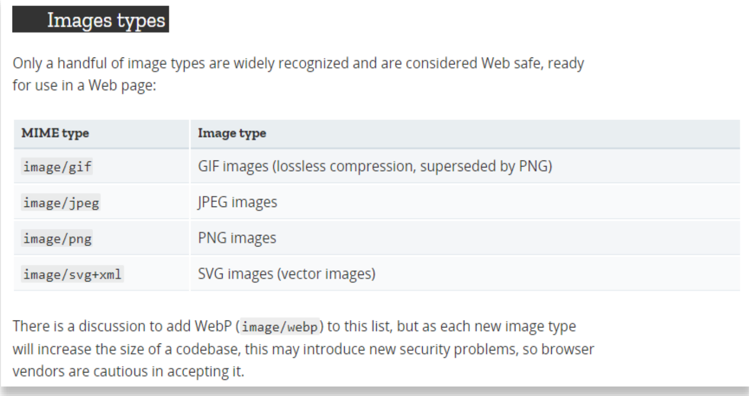
Look for an app in the Store

☑ Always use this app to open .jp2 files

Figure 1: Even common formats may not be supported by popular platforms.

Similarly, browser software developer guidelines discourage the use of all but a handful of image and audio/video types (see Figure 2). This presents a formidable obstacle to archivists and content generators who wish to make their data available in a native format to their customers. Either the data are produced and stored in multiple formats, migrated as new formats are adopted, or each provider requires browser plug-ins for each format type.

With any of these options, the cost, either of maintaining multiple data versions or migrating across data formats, can be significant. Not only are there costs associated with format monitoring and migration, there is the ongoing problem of conversion loss and significant risk in relying on what are often proprietary solutions to data access.


Images types

Only a handful of image types are widely recognized and are considered Web safe, ready for use in a Web page:

MIME type	Image type
image/gif	GIF images (lossless compression, superseded by PNG)
image/jpeg	JPEG images
image/png	PNG images
image/svg+xml	SVG images (vector images)

There is a discussion to add WebP (image/webp) to this list, but as each new image type will increase the size of a codebase, this may introduce new security problems, so browser vendors are cautious in accepting it.

Figure 2: Only a small number of image types is recommended for web browsers (Mozilla, 2018).

Historical Approaches and Their Challenges

Status Quo

The simplest method of handling formats is to maintain data in the original, native format. While low-cost, this approach may be risky. Each type of data file, whether a simple image on website or a complex spreadsheet, needs software to interpret and interact with it. Not all formats are natively supported on all devices, and may require additional apps, plug-ins, or some sort of intermediate software that does not automatically come installed on the device for display (see Figure 3).

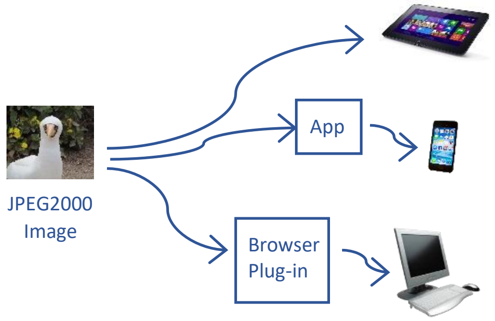


Figure 3: Not all formats are natively supported on all devices. Some may require plug-ins or apps for access.

Not only do customers have the inconvenience of downloading and installing potentially pricey intermediate software, there is inherently reliance on device or platform developers to maintain software through generations of formats and devices. Support of common formats is up to any manufacturer, in fact, manufactures can easily choose to not support a given format, e.g., Apple and Flash (Jobs, 2010).

Monitor and Migrate

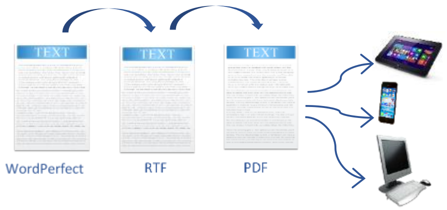


Figure 4: Migration over generations of popular document formats.

An increasingly popular method of mitigating format obsolescence is to repeatedly convert into newer widely-supported formats (see Figure 4). This approach requires format monitoring and conversion expenses, and often results in loss of fidelity, features, and metadata due to the conversion process. Embedded metadata easily goes astray and original formatting can be lost. Additionally, there may be no support of various data file components, e.g., the lack of embedded audio or video support in PDF/A. However, when there is no concern about loss of fidelity or features, this is a straightforward preservation approach.

There are services that will automatically handle the migration of popular formats; however, it is important to be aware that they cannot mitigate the conversion loss challenge and may not handle less popular or proprietary formats.

Alternate Solution: Format Preservation

An alternative to the migration approach that still allows data to be viewed or played on any platform is to store an Accessor, such as a viewer or player, with each data file.

As shown in Figure 5, each file is packaged with its own specific, platform-independent Accessor software. The Accessor software stays with the file, so customers do not have to search for, or pay for, programs to read data. When the data file is to be displayed, the Accessor software automatically loads the data, interprets or decodes it, and displays or plays the data in the file. As an extended feature, the Accessor can include interaction; for instance, the Accessor software can be a spreadsheet program that allows the user to modify cells or formulae.

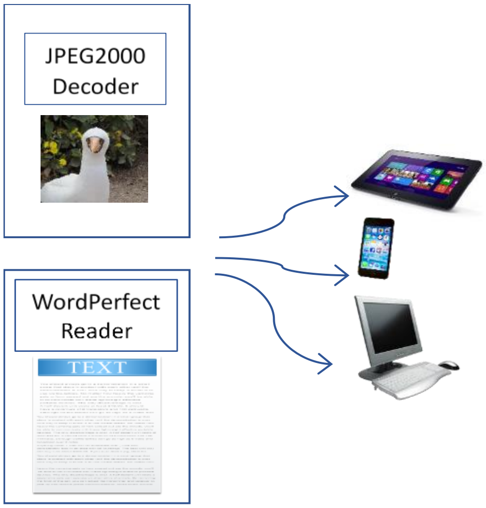


Figure 5: Storing data files with the appropriate reader allows the files to be viewed on any platform, that includes inserting non-natively supported files into browsers.

Packaging of the data files with their Accessors is straightforward. Automated software chooses the correct Accessor, such as a WordPerfect reader or JPEG2000 interpreter, for each file, and stores the data file and the reader together in a single file in the owner's data repository (see Figure 6). The combined file with the packaged data and Accessor is treated just like any other data file, it is stored with its metadata in the data asset management system, whether in the cloud or locally, and is retrieved like all other files.

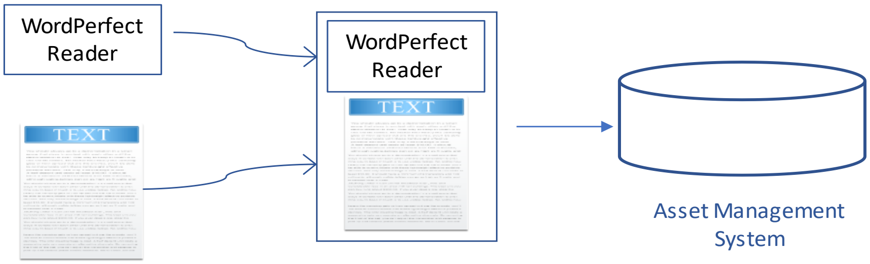


Figure 6: Packaging and storing the files is straightforward. The packaged files live in the owner's repository, just like any other file

Format Preservation: Key Advantages

- Accessibility: In addition to being easily retrievable via Bevara's free apps, the original data file in the package is extractable, that is, it can be pulled out of the package to be used in any available software.

- Metadata Integrity: The original metadata stays with the original data - it is never lost during a format conversion process.

- Format Persistence: The data file *never* changes, it stays viewable in its original, native format, so there never is the need for format conversion and there never is conversion loss. Furthermore, there is never reliance on software providers to support legacy formats.

- Resource Conserving: Format preservation eliminates the need for constant format monitoring and migration.

- Open Source: Bevara's solution is open-source. The data and metadata are always visible and always retrievable. There is no uncertainty about whether any data can be retrieved and accessed.

What Should You Do Now?

A key component in developing a strong digital data archive is to develop a technical digital data preservation strategy in addition to an organization strategy. On the technical side, the strategy involves not only selecting how and where to host the data and metadata but determining if you have formats that will present access problems in the future. Follow our blog at *www.bevara.com/blog* as we explore the pluses, minuses, and intricacies of various formats.

Finally, as we saw, approaches to maintaining accessible digital data are to either develop a migration strategy or to consider the alternative of format preservation. Bevara offers tiered packages to suit a range of needs, spanning individuals using a single format to large institutions requiring support for a host of media, documents, and scientific data types. If format preservation is the right choice for you, we will work with your team or repository provider to ensure a transparent, seamless operation. Contact *support@bevara.com* with any questions or for a preservation trial.

References

Cave, A. (2017, April 13). *What Will We Do When The World's Data Hits 163 Zettabytes In 2025?* Retrieved from Forbes: https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#2dfda165349a

Jobs, S. (2010, Apr). *Apple .* Retrieved from Thoughts on Flash: https://www.apple.com/hotnews/thoughts-on-flash/

Library of Congress. (2018, Jan 1). *PDF/A, PDF for Long-term Preservation.* Retrieved from Sustainability of Digital Formats: Planning for Library of Congress Collections: https://www.loc.gov/preservation/digital/formats/index.html

Mozilla. (2018, January 1). *MIME types*. Retrieved from Mozilla Developer: https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/MIME_types

National Archives and Records Administration. (2018, January 1). *Appendix A: Table of File Formats*. Retrieved from Records Management Regulations, Policy, and Guidance: https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#digitalphotographs