File Formats

Preservation and Curation of ETD Research Data and Complex Digital Objects

INSTITUTE of Museum and Library SERVICES

EDUCOPIA INSTITUTE

Photo by Anthony Martino on Unsplash

File Formats

An email arrives. “We’re excited to tell you … As part of the acquisition, we will shut down your account in two weeks. Per the terms of the user agreement, it is your responsibility to retrieve your data from the service.” What promised to be the latest-and-greatest way to get your work done has now trapped your work. The only export option is a .xyz format you’ve never heard of and can’t import into anything else. ■

Rationale and Motivations – Why

File formats give shape to the way you store and use your data. Different file formats let you use the same information in different ways. For example, when you submit your thesis or dissertation, you will likely do so as a "PDF" file so that it will look the same on every computer screen. But, as you are writing and editing, you will probably use a range of different formats. Depending on the software used, you might notice the different extension at the end of your file name, such as "docx" for Microsoft Word or "tex" for LaTeX or "pages" for Apple Pages. You also may notice a different icon for different file types.

Many file formats exist, and most digital content types can be created and stored in multiple format types. For example:

- Images: jpg, gif, tiff, png, ai, svg, ...
- Video: mpeg, m2tvs, flv, dv, ...
- GIS: kml, dxf, shp, tiff, ...
- CAD: dxf, dwg, pdf, ...
- Data: csv, mdf, fp, spv, xlx, tsv, ...

There is no perfect file format. Each will have advantages and disadvantages depending on your research uses. However, it is important to find a file format, or set of file formats, that helps you complete your research now, and that you can access again in the future. This is true both for your research outputs (what you create) and your research inputs (materials you use in the research process).

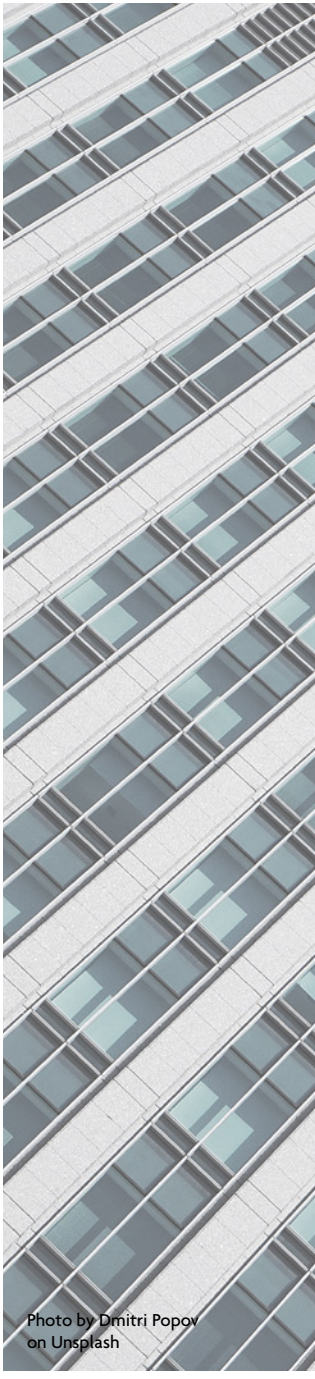

Photo by Dmitri Popov on Unsplash

The Basics – How to Do It

File format choices are often bundled with and determined by the software you choose to use in your research. It is important to understand which formats each of your chosen software packages will both save to and open, especially if multiple pieces of software will be used during the research process (e.g., Microsoft Office, SPSS, Adobe Photoshop).

You also need to consider what might happen if you can no longer use the software. Whether the software publisher goes bankrupt, the latest version refuses to read older data, or you can't afford a personal license for it after you graduate, the end result is the same. Losing access to your software can mean losing your data, especially if it is the only software that can read your data. An initial task is to determine how much you can trust and depend on the software you want to use. This can be difficult to judge, but a good metric is if you can save or export your data to a format used by other software.

Many disciplines have default formats that are used, and sometimes recommended, by practitioners. Following community practice means that challenges you might face with your choice of format are problems that others are likely to have already faced and probably solved. It also enables easier collaboration with others, since you do not have to migrate data from one format to another to exchange it with colleagues.

How you will use the data is another consideration in choosing your formats. If you are actively collaborating with colleagues, it is good practice to store data in a format that can be easily shared and fully edited. On the other hand, once you publish your research, you may wish to save your data to a format that is more stable. For example, while you may work with a PowerPoint file while creating a presentation, you might save it as a pdf once you have delivered it. Finally, if you are saving data to a temporary file only to import it to another piece of software, you may not need to worry about the long-term use of the temporary format at all.

You can rarely find a perfect format, but you can at least choose a good file format.

- Use software that imports and exports data in common formats.
- Ask advisors and colleagues which formats they use.
- Choose a format with functions that support your research needs.
- Save final versions of your content in multiple formats in order to spread your risk across multiple software platforms (e.g., docx, pdf, and txt; or mp4, avi, and mpg).

Tools – What to Use

Tools are available to help you make formatting choices and to accomplish formatting migrations. Here, we focus on three functions and some of the tools available for each.

Format Choice

The choice of file formats depends on your field, subfield, or specialization workflows. While few tools exist to evaluate your choices, two potential resources are the Sustainability of Digital Formats site and the Recommended Formats Statement, both created by the Library of Congress. These resources document the quality and functionality of many file formats, grouped by the type of content they can store.

If you are using website-based materials as evidence or references, you will need to take precautions to ensure that if the content moves, changes, or disappears, you still have evidence of its existence. Current tools to help you ensure the longevity of these materials include Robust Links and Archive-It. You can also take screenshots of important digital content in order to preserve the look and feel of an object.

Conversion

If you need to convert files from one format to another, there is a broad range of available tools, some of which are safer than others. These range across proprietary, "freeware" and open source solutions. If you are working with formats that are in broad use, you may have lots of choices; if you are working with more esoteric, domain-specific, or older formats, your choices likely will be limited. In some cases, you may have to perform multiple conversions in order to get content from its current format into the format you prefer.

Be aware that saving a file in a different format (e.g., using Word to save a pdf version of your docx file) is a form of conversion that may transform your data in specific ways. Understanding the conversion options in these cases is important. Do you want the pdf to be a flat image (and thus a smaller file) or to maintain the text (making it possible to search the text or index the document)? Both options are available, and either may be appropriate according to your conversion goals.

Before you undertake any conversion, you need to identify what characteristics of your data are important to maintain during the conversion. For example, are the colors in a document or image important? Is the pagination essential? What about references? You will want to test these after your conversion is complete to ensure that you have a conversion that will meet your needs. This evaluation process may include comparing the size of an image or the length of an audio track before and after conversion. For helpful guides to these types of quality control characteristics, see the National Archives' reformatting guides.

Packaging

Once you have completed a research project (such as your thesis or dissertation), you will likely need to submit it to your institution for review and long-term storage. As you do so, you will need to ensure that you submit all of your content in formats that will enable it to be viewable in the future. Most campuses today require theses and dissertations to be submitted as pdf files; many also allow supplementary research outputs to be submitted in other formats.

As you create a pdf, you will want to give attention to the following:

- **Embed fonts.** You can embed all of your fonts as you convert your content to pdf (e.g., from Microsoft Word to Adobe Acrobat) by selecting this option as you make the conversion (e.g., "save as PDF," and then select "Options" and choose any of the PDF/A options). Note that this is often an explicit requirement (e.g., UMI ProQuest requires embedded fonts). Ultimately, this ensures that your font will look the way you intended it to look when future researchers view it.
- **Embed hyperlinks.** Hyperlinks in a document may not be maintained when that document is converted into pdf format. If you do not actively choose to convert your hyperlinks, your pdf may contain blue text and underlines that signal links, but not actually have operative links (so that you click on the link and nothing happens). In order to ensure your document's links are "live," pay attention to the settings you use, and select "convert hyperlinks" (usually in the advanced settings).
- **Stabilize hyperlinks.** For every link you include, use a web-archiving tool (e.g., Robust Links, Archive-It, or PermaCC) to stabilize the version of the content to which you are linking. Web-based materials are notoriously ephemeral. They often change, are moved, or disappear entirely. Using a web-archiving tool will help you to avert a range of problems, including content changes, link rot, and 404 errors that could compromise your research findings' legitimacy in the future.
- **Store supplementary materials as separate files.** PDF is sometimes used to bundle research materials (e.g., lab notebooks, multimedia files, datasets) into a single pdf document. Embedding multimedia components within the full text document might seem helpful for keeping your files together. However, file formats change over time, and inevitably, different components of your research will need to be converted or migrated to new formats in the future in order to be viewable. As such, storing them as separate-but-affiliated files (e.g., embedding links to multimedia within the full text and including the multimedia as separate files) will help to ensure their longevity. If you are required by your institution to embed your files in a pdf, be sure to use standard file formats such as bmp, jpg, gif, or tiff for graphics, mpeg for video, and wav or mp3 for audio.
- **Verify the PDF/A compliance.** Use the Acrobat Pro "Preflight" feature (on the Acrobat Pro menu, look under "Edit") to verify the PDF/A compliance.

Resources

All links provided were last checked 10/3/2017, and the content we reference here was saved on that date in the Internet Archive's Wayback Machine. For links that no longer work, please visit the Wayback Machine and enter the url to surface the resource.

- For more information about the range of existing file formats, see Wikipedia's List of File Formats.
- For more information about the stability and expected longevity of a variety of file formats, see the Recommended Formats Statement created by the Library of Congress.
- To understand some of the methods available for converting your files from one format to another, see the US National Archives' Reformatting Guides.

Photo by Christian Fregnan on Unsplash

Photo by Adam Birkett on Unsplash

Activities

1. Find a folder of research materials that you have collected on your computer. Look through your materials and answer the following questions.
   a. What software do you need to access these materials?
   b. Do you face a risk of losing access to that software, now or in the future?
   c. Would a colleague be able to open and use your materials if you shared it with them?
   d. Can you submit your thesis/dissertation and its related research materials using the file formats supported by the software you are using?
2. Create a diagram of your research workflow. You can do this however you prefer, with software, with pen and paper, etc. In your workflow, find the materials that you are creating, collecting, and editing and answer the following questions.
   a. What file formats are you using and producing at each step?
   b. Do these file formats support the functionality that you need for your research?

INSTITUTE of Museum and Library SERVICES

EDUCOPIA INSTITUTE