Title

Assess, Annotate, Export: Quick Recipes for Archiving Your Personal Digital Life

Permalink

https://escholarship.org/uc/item/9mm6s1bs

ISBN

978-0-8389-1605-6

Authors

Emmelhainz, Celia
Wittenberg, Jamie

Publication Date

2018

Assess, Annotate, Export:
Quick Recipes for Archiving Your Personal Digital Life

Authors: Jamie Wittenberg & Celia Emmelhainz

In The Complete Guide to Personal Digital Archiving, ed. Brianna H. Marshall. 2018.

This chapter offers step-by-step instructions, or “recipes,” for archiving different kinds of digital objects. Because other chapters in this volume offer approaches for preservation, these recipes focus on extracting digital objects from platforms in a state that facilitates preservation. The goal of PDA recipes is to provide practical, immediate guidance that will be applicable to a broad range of archiving circumstances. They do not require extensive technical knowledge or a strong theoretical understanding of digital archiving. This makes them very useful as templates to pass on to patrons with immediate archival needs.

Many of these recipes target cloud or web-based applications because these tools can require less conventional methods of extracting digital objects. Furthermore, it is increasingly common for digital items to live online, rather than on internal and external hard drives or storage devices like CDs. While each recipe addresses a specific category of digital object, there are general principles that apply across these categories. The three-step “Assess, Annotate, Export” are broadly applicable and represent core practices of preparing materials for archiving; assessing the current environment, capturing and/or adding metadata, and retrieving the digital object in an appropriate format.

These recipes often reference ‘proprietary’ and ‘open’ tools. This language refers to the accessibility of the software’s source code and the user’s ability to examine or adapt the tools. There are many software distribution models, and proprietary/open does not necessarily correlate to whether the software publisher is commercial or non-profit. Both models have advantages and disadvantages. These include:

Proprietary *[Software with its copyright and/or patent rights retained.]*

> *Pros*: often have support staff, bug fixes are developed by the same organization that created the software, larger development budget so may have more advanced features, in some cases software has larger user community, documentation is often user-oriented
>
> *Cons*: If the company fails, the software is unsupported and no one can fix problems, some software companies make exporting content difficult in order to ‘lock’ users into software package, usually requires that a license be bought in order to use the software, terms may change unexpectedly (ie, you used to be able to use these features for free, now you have to pay)

Open *[Software with its source code accessible to users.]*

> *Pros*: community of users contribute to the code, often has many more eyes on bug fixes, transparent mechanics, if community falters, standards will still be available, no incentive to prevent users from moving content between platforms
>
> *Cons*: sometimes not as widely used, bugs may not be fixed as quickly because there is less incentive to do so among a small community, smaller development budget may mean less advanced features, documentation may be highly technical

When evaluating the platform you are currently using, you'll want to consider first how to archive what you've put into that website, including captions, views, and other contextual information. You'll also want to consider whether you're better off moving to another platform because it lets you more fully preserve and share. Questions to consider:

1. Can you easily review and **assess** everything on the platform? Is there an overall view of how many posts, notes, or videos you have? Can you filter and see which may be most useful to save? Although many websites let you do a bulk download, you may be better off to be more selective in what you archive for the future.

2. Can you **annotate** your files easily while online, or will you need to go back and merge information about your files with your files manually once they're stored on your computer? If you export, will it include everything, such as captions or comments? Is there a way you could document your files before uploading to the cloud, so you don't have an enormous job later?

3. When you **export** your files or online presence, can you easily get it into an open format, one that you or your loved ones may be able to open, print, or re-use long into the future? Will you be able to export in bulk, or will you have to do it one by one? Is the website committed to allowing you to fully export your files, or are you obliged to use an external application to pull out your data, one that's possibly risky website to your computer?

There's no easy answer to the question of tradeoffs between using proprietary but popular websites, or using more open formats that may be less likely to be used by other people in your social circle. However, we hope to give you some guidelines to get you thinking about what you currently use in your online life, as well as whether it will be something that you can store or re-use in the long run. New online tools will always come along, and may be very useful for your social, business, and personal life. We won't cover everything in these recipes, but we'll give you a place to start.

Archive Your Web-hosted Video

Web-hosted video includes moving images and associated audio that is hosted by a streaming service. Much of this guidance will also apply to born digital video more generally.

Instructions	Context
Assess Before archiving video, determine who holds the copyright for the video and determine whether you are licensed to archive it. Once you determine that you may legally archive the video, identify whether you can download directly from the hosting site or whether you will need to use a third party application to download it. Identify the highest-quality version of the video that you can find.	Consequences for violating copyright law can be severe. It is not recommended to archive video if you do not have a license to do so. Typically, commercial hosting sites like YouTube will allow you to download your own videos in the video manager. If you are unable to access the account, a third party application may be effective. These applications are ubiquitous and some may contain malware. Be cautious.
Annotate Document the original format of the video, any migration or conversion you do, the date of download and the date of creation in a text file. Describe the content of the video, the creator, and the subject. Include contact information for the copyright holder. If you are the copyright holder, consider applying a creative commons license.	Because video files are often large and can be time or resource-intensive to play back, written descriptions of videos will help future viewers find the video they are looking for, or potentially make a case for allocating resources towards recovering a corrupted file so that it can be played back. A Creative Commons license will allow others to legally save and share your video.
Export Download the video that you wish to archive. If your video is in a format that is not widely supported or difficult to play back, like the Flash container format FLV, use a converter to migrate the video to a widely-used format with well-supported playback software and open standards. Ensure that documentation is stored with the video files.	The Library of Congress recommends the following formats, in order of preference, for personal digital video archiving[1]: - MPEG-2 - MPEG-4_AVC - MPEG-4_V - MPEG-1 - Compressed in wrappers like AVI, QuickTime, WMV, etc.

[1] Sustainability of Digital Formats. Last updated: 2014. http://digitalpreservation.gov/formats/content/video_preferences.shtml

Archive Your Notes

There are many flavors of note-taking applications that can be used in-browser, offline, or on mobile devices. Notes are usually organized into one or more notebooks and are automatically saved locally and to the cloud.

Instructions	Context
Assess Take inventory. How are your notes currently organized and what kinds of media do they contain? Text? Photos? Audio recording? If you have separate notebooks for various purposes, like work and personal, consider archiving these separately.	Taking inventory, or assessing your content, helps you make decisions about what you want to keep and what you don't. It also gives you important information about how much you have and what kinds of items you have. This kind of information will inform your preservation decisions.
Annotate If you have photos, audio, drawings, or other media, ensure that they have descriptive names, ideally using a consistent naming convention. Be sure to put the creation date of a note into its text or title, as it will otherwise be stripped away when you export.	Descriptive file names are crucial during this process because programs like Evernote will automatically group all of the media from a single note together and move it into a separate folder. During this process, the media may lose some of its context.
Export To export one or more notes, select the notes that you wish to export. Then, navigate to 'Export' (usually in the File menu). Choose a text-based format like HTML or TXT. You will then have an opportunity to choose a descriptive name for your folder. All of the notes you selected will be files in this folder. Usually the media embedded in the note will be in a subdirectory in this folder.	When selecting a format for your notes, consider how you would like to access them or how you would like others to access them. If there are important formatting considerations, HTML or even PDF-A may be preferable to TXT. However, if you are only interested in the content itself, a TXT file is much more expedient and inexpensive to preserve.

Archive Your Online Photos

Flickr is a major photo-sharing site; patrons may also be storing photos in iCloud, Facebook, or Instagram—or on their mobile device. Apple devices may be set up to back-up photos to iCloud.

Instructions	Context
Assess Take inventory in Flickr. How are your photos currently organized in sets or by batch uploaded? Do they have descriptive captions? Your primary option for sorting Flickr photos is 'date taken' or 'date uploaded,' although a Magic View (!!) attempts to automatically identify birds, dogs, architecture, and other photo features.	Assessing your photos helps you decide which need downloaded, and what caption editing you may want to do before you back them up to your computer. In the future, it would be wise to add captions on your computer before uploading, and back up the photos to an external hard drive.
Annotate Ensure that your photos on Flickr have descriptive captions, including date, name, and place. The sub-caption description will not be downloaded with the file, nor does a caption seem to be retained from your original upload. The date of the original photo will remain available in file information under "modified" or "date taken" but after downloading the file, it will be listed as "created" on today's date.	To have photos with sortable file names, make up a system to list dates first, location first, person first, or other method of sorting. Unless you download each batch separately and save to a separate folder (recommended) all your photos will be in one .zip file.
Export In the past, Flickr users ran third-party programs like Bulkr or Flickr Downloadr. As of 2015, Flickr offers native support for batch downloading. To download and back up your files, login and click Camera Roll, and select either by individually tapping on photos or choosing Select All above all uploads on a given date. A download icon will appear at the bottom. Click download to save a .zip file to your hard drive containing all the photos. Files are named according to your caption in Flickr (but not the description), so ensure that photos are properly captioned (above) before downloading.	Photos can be downloaded from Flickr as .JPEG files, which are the most popular standard for personal archiving (Severson's Digital Photos chapter). This will be adequate for most users, but .TIFF provides archival quality for professionals. ?]

Archive Your Social Media Accounts

We increasingly conduct our social relationships online, and social media sites such as Instagram, Facebook, and Twitter document our family and friend relationships. These won't be saved for the future, unless you follow this handful of steps to download.

Instructions	Context
Assess What social media accounts are you currently using? How full of data are they likely to be? Do you keep regular backups, or could you set up a schedule to do so?	Because social media accounts come in and out of fashion, it's good practice to keep track of the ones you use most frequently, and assess the kinds of information that would be helpful to export. If the site itself doesn't allow a full export, sometimes third-party sites or browser add-ons may be able to access the data and provide you with an alternate view.
Annotate Social media accounts like **Instagram** will export (through Instaport.me) a folder of jpg images. However, they include no captions, and recreating those manually could be difficult.	Social media sites vary on how much information they export, and how well linked it is. When a site includes only partial contextual image or captions, you can go back and add that information to a file manually. Other sites include full metadata in their export. In that case, do your best to describe photos and the people therein when first posting.
Export In **Facebook**, go to *Settings* and *download a copy of your Facebook data*. Open the .html file in a browser to see your photos by album, with comments; messages; timeline posts (but without links!); plain text list of friends, and more. In **Twitter**, go to the bottom of your Settings page and request your archive. It will include a .csv file for analysis in a spreadsheet, as well as an .html file for viewing in a browser.	Social media sites come and go, so it is good to make a backup (if available) at least once a year. You won't be able to do this for all accounts, but put exports of your data on your calendar so that you're preserving your digital record. You may even want to open exports (e.g. html files) and save the images to a new format like PDF.

Archive Your Documents

Microsoft Word dominates the market, followed by Office for Mac and Google Docs. All are in proprietary formats and should be moved to .rtf for long-term storage.

Instructions	Context
Assess Open a folder of files to see what format they currently use: doc, docx, or pages? Check the file size, and look inside a few files to see if they have a header with more information on the purpose of that file. You may also check for descriptive file names that will allow people to easily sort and find what they want.	Word processing programs go out of style, and you don't want to always be monitoring your file formats. Open Document Text (.odt) or Rich Text Format (.rtf) are formats likely to survive into the future. odt is more fully open but not as widely used as rtf, which is produced by Microsoft. Beware that complex features like comments and track changes may not be support in open formats, so save those to PDF.
Annotate Before or after saving to an open format, make sure your file is named descriptively so you know what's inside without having to open every file! You might also add a paragraph at the top of each document explaining its purpose and any necessary contextual information.	While you want to keep file names brief, you also want them descriptive: *2017-10-21_widget_inventory_CE_v1.rtf* will tell you when a file was created, what it's about, the author, and the version.
Export In Google Docs, select File → "Download as" and select rtf (rich text format) or odt (open document text). In a desktop program like Microsoft Word, select File→ "Save as" rtf format, an open file type that preserves both formatting and accessibility. if you don't need to save highlights, bold, and the like, .txt is an option and is easier to analyze in large batches.	In the future, you can change the options in any word processing program (in MS Word it's File → Options → Save → Save file in format). Saving automatically to rtf will save you time later. Also remember that no hard disk or flash drive lasts forever; make sure to copy and back up your files every 3-5 years.

Archive Your Emails

You likely use Mail on an iOS device, Microsoft Outlook, or Google Mail to manage your emails. These can be exported in bulk bundles, or individually as text documents.

Instructions	Context
Assess How many emails are in your inbox, and are they all useful? Have you organized them into folders or labeled them? Do you know which ones are most important for you to save—and are they individual emails or whole groups?	You may not want to export all emails in one huge cluster, as that makes them difficult to view and organize later. Before annotating, you might search and delete unneeded advertisements or topics by sender or subject. You'll also want to think about which parts of your email archive you need and why.
Annotate Search for emails by subject, content, or sender and add the results to folders or labels. Going forward, you can set up certain topics to be automatically labeled as they come into your inbox. Organizing in this way makes it much easier to pull an export of only the emails you want.	Your email comes with built-in annotation in the form of *to*, *from*, *subject*, and *date*, so there's not much to do on the annotation front. Although you can export all emails, adding to folders or labels makes it easier to selectively export. You can also save to txt or PDF in most programs—make sure to name files descriptively.
Export In Outlook, look for File→ Options→ Advanced → Export and export all mail or one folder to a .pst file. This is not an open format. To save one email, use File → Save as. Google Takeout lets you bulk export mail, calendars, and contacts to open formats. You can export all mail or only certain labeled messages. On iOS, select all messages and save as "raw message source" to get an mbox file.	Most email programs can export all mail to mbox, an open file format that can be viewed in opens-source programs like Thunderbird. Use the ImportExportTools add-on to export a folder of messages to text or PDF format. Note that Microsoft's .pst format needs to be converted to .mbox to be used in this way. Individual emails can also usually be saved to text or eml, or printed to PDF.

Archive Your Spreadsheets

A spreadsheet is a document that contains data structured in a grid or matrix.

Instructions	Context
Assess Determine what software was used to create the spreadsheet(s). Identify whether the workbook (collection of sheets) contains data, graphs, images, functions, or other material. Identify whether formatting in the spreadsheet is meaningful – for example, cells with background colors or bold font.	Spreadsheets may be generated using statistical software like R, Stata, and SPSS, or using word processing software like Excel or Google Sheets. Often, documentation will be embedded within the spreadsheet, hyperlinks will be embedded, or cells will be highlighted.
Annotate Create a copy of the original workbook and alter it so that it is as platform-independent as possible. If formatting has meaning, create another column to indicate the meaning with text. If functions are used in cells to calculate value, create a column for calculated value and function used. Cells should contain only unformatted numbers and text. Repeat for each sheet in workbook.	It is important to preserve the file in its original state before taking steps to create a clean version. All major spreadsheet-generating software will allow for export to a Comma Separated Value (CSV) file, but only plain text will be preserved. Embedded hyperlinks, formatting, and figures will all be lost if the spreadsheet is not cleaned.
Export If the spreadsheet contains any charts or images, export them in a widely accessible image format like JPEG. Export your altered version of the spreadsheet to create one CSV file for each sheet in the workbook. Put the .csv files, the original spreadsheet file, all images, and a text file describing the contents of the spreadsheet(s) and the version of software in a folder.	CSV files are flat; they do not have internal hierarchy. This means that sheets within a workbook cannot be sufficiently represented within a single CSV file, and instead, the relationship between sheets must be created externally and documented. Each sheet can be represented as a single CSV file, and a combination of file names and documentation can describe the relationship.

Archive Your Websites

Web archiving is the process of saving content that has been published to web - often using programs known as ‘crawlers’ that automatically ‘crawl’ from site to site, downloading content.

Instructions	Context
Assess Determine what is within your scope to archive based on the collection guidelines and the mission of your organization. Evaluate how much you expect to acquire and what resources you have to dedicate to web archiving. If you expect to archive sites intermittently or manually, evaluate your needs for storage and access.	Institutions frequently choose to archive all content within a domain - for example, Indiana University archives all web content published in the “indiana.edu” domain. Some web archives are event-specific, for example, the September 11 digital archive collects resources from a variety of content producers.
Annotate When a website is downloaded, the source code will contain important embedded metadata, but it is also important to include contextual metadata, ideally in a readme stored with the website. Include information about the date and time accessed, the archiving institution, and any media missing from the site.	Even highly sophisticated crawlers can not capture all of the content available from websites. Sites that are database-backed, dynamic, have embedded media, gated content, or are protected by robots.txt[2] may not render in a web archive. It is important to document what is missing so that users are aware they are not accessing a complete record.
Export The format of archived web content will be dependent upon your users’ needs. For small, manual collections with a limited number of users, HTML or PDF are acceptable. For larger collections with broader user bases, the WARC (Web ARChive) file format is standard.	There are several widely-used web archiving services available from the nonprofit Internet Archive. Heritrix is a free, reliable Java-based tool and Archive-it is a fee-based user friendly service. For intermittent or manual archiving, the Wayback Machine allows users to submit a single page to be archived.

[2] The robots exclusion protocol, or robots.txt, is a standard that allows site owners to ‘tell’ crawlers that they do not wish to have their content accessed. Reputable web archiving services will honor a robots.txt file.