Personal Digital Preservation: Issues and Approaches

D. Randall Wilson, Ph.D.
FamilySearch
RootsTech 2012

Abstract. Many individuals and organizations have valuable photos and documents that need to be preserved, organized and shared. However, it is nearly impossible to preserve such "digital artifacts" in a way that guarantees that they will be around and accessible 50 years from now. People often have shoe boxes full of old family photos, which are priceless treasures when the individuals are identified, but become nearly worthless when nobody can remember who they are. Other people probably have many old photos that would be valuable to you, but it is currently very difficult for you to find out that they exist or to get a copy. Others in the family could help you identify who the people are in some of the photos you have, but there isn't a convenient way to get their input. Many excellent products and services have been created that address parts of the problem, but there remains a need for industry standards and free (or perhaps pre-paid) long-term preservation services that continue to preserve data and metadata even if the user passes away or a company goes out of business.

This paper reviews the requirements for long-term preservation of personal digital artifacts, discusses existing standards that might help (such as Adobe's XMP metadata format and the Metadata Working Group's definitions), and suggests what still needs to happen to make personal digital preservation reasonable and attractive to the masses.

1. Introduction

There are many, many people in this same boat:

1. They have a bunch of old family photos, and they can scan them onto a hard drive (perhaps with help from a computer-savvy family member).
2. They know who some of the people in the photos are, and are willing to tag them, but they don't know who some of the people are, and wish we could ask our relatives for help in tagging them.
3. There isn't a widely accepted standard for storing "person tags" in image file metadata, so tagging in one system (iPhoto, Flickr, Picasa, etc.) probably won't transfer to another one.
4. They don't know of a permanent place to put photos where they won't be lost if they stop paying their subscription and/or pass away. (Children may wipe their hard drive before donating their computer to a charity, for example).
5. They can share photos on Flickr, DVD, FaceBook, etc., but it's very ad-hoc and short-lived.
6. They may need to remember how the photos were organized when they found them; but they want to reorganize them into a more useful arrangement. Yet a hard drive usually only allows one arrangement.
7. They don't always know when or where each picture was taken, but they could often at least organize them roughly by time ("these photos are before those"). But they don't have a great tool to do that (creating sub-folders and renaming them so that they happen to sort right in an operating system is pretty tedious).

The same problem exists for other types of "physical artifacts" besides photos, such as reels of movie film, documents, family bibles, journals, etc. These can usually be digitized into "digital artifacts", such as images, PDF files, audio files, video files, etc. But organizing, tagging, describing, archiving and sharing them face the same issues as for images. (In this paper whenever "photos" and "images" are discussed, the same discussion almost always applies to these other types of physical and corresponding digital artifacts as well.)

This paper discusses the above problems and suggests possible solutions, including standards that remain to be defined. It also mentions some products or standards that have been developed that address parts of the problem. Section 2 discusses issues around scanning and organizing collections of images. Section 3 goes over new standards that could start making face-tagging data interoperable. Section 4 discusses long-term preservation and Section 5 discusses the sharing of digital resources. Section 6 summarizes the challenges that remain for making truly long-term preservation a widespread reality.

2. Digitizing and organizing collections

When gathering, organizing and digitizing collections, it is important to remember the archival principles of *respect de fonds* and *respect for original order* [1]. The former has to do with understanding the entire group from which materials are drawn, and the second has to do with remembering the ordering within those groups. Often the grouping and ordering of materials help give them powerful context that gives the materials meaning. A set of slides in the order they were taken is more valuable than a pile of slides loose in a box, because they imply chronological order and thus help you know when and where they were taken and even who is in them.

When digitizing images, one approach is to create a folder hierarchy in the computer's file system that reflects the physical groups you found the images in. For example, my grandfather's slide collection came in 19 large metal boxes, so I arranged those in chronological order and scanned them into 19 folders on my computer representing these. Some arrangement information isn't important enough to save. For example, these 19 metal boxes were in several larger cardboard boxes, but their inclusion there was completely arbitrary, so I didn't bother capturing that.

As another example, I scanned many small boxes of slides that each represented one roll of film. It is important to remember the *original physical arrangement* for a variety of reasons. For instance, if I discover that an image in one roll was taken before an image in another roll, chances are that the rolls are out of order. Knowing which images were on each roll helps fix problems in organization. Remembering the original physical arrangement retains significant contextual information related to how the images were first created (i.e., they were taken in sequence within the same camera).

On the other hand, collections of images are often more useful if rearranged into a *logical arrangement*. For example, a particular vacation to Hawaii may cover 7½ rolls of film. That group makes a handy collection. As another example, I have scanned old family photos from various shoe boxes, but there are pictures of my father as a baby from different boxes that were clearly taken on the same day. A more logical arrangement would put these together.

Ideally, therefore, it would be nice to be able to remember the original physical arrangement of a set of images but be free to rearrange them into better groupings.

One way to accomplish this would be to *tag* images with information about what collections or sub-collections they were part of originally, and then arrange them on the hard drive (and tag them, too) according to another logical arrangement. Then if anyone is handed an assortment of these images, they could use the metadata in the image files themselves to reconstruct both the original physical arrangement and the alternate logical rearrangement(s) as needed.

One open issue here is to decide on a standard way to tag images with an identifier (e.g., a URI) that identifies a collection the image is part of, plus a *sort key* that puts the image in the right order within that collection. If the ID is a URL, then it could optionally point to an online resource that has additional information about that collection, such as a nice title, description, citation, paged list of contents, etc. One trick used recently at FamilySearch is to have a URL that returns HTML (i.e., a user-viewable web page) if the "accept" header is HTML (as it would be from a web browser); but returns XML if the "accept" header is "application/xml" (as it could be from an application). Thus a URI can act as the unique identifier of the resource, the address of a web page that displays the resource, and the address of an XML resource that describes the resource.

A standard is needed for collection and arrangement information so that multiple clients can be built to help users organize their images, and the results can be shared with others or migrated to other software systems without having to repeat all the work.

3. Tagging Faces

Family photos can be precious treasures if you know who is in them, or nearly worthless if you don't. Most photo editing software allows users to add captions to photos in which they could describe who the people are in a photo, and these captions tend to transfer pretty well from one application to another. However, it is not always obvious which name goes with which face. For example, I tend to list people

left to right (comma-separated), back to front, with a semi-colon between rows. But that is just my convention, so others may not be sure what I meant.

Several software systems allow users to tag faces, including iPhoto, Picasa, Facebook, F-Spot, Photoshop, Flickr, Photoloom, Mundia, 1000memories, myheritage, heritagecollector, and many more. A few of these even have automatic face recognition to help this go more quickly. Typically the face tag information is stored in a proprietary database and not included in the metadata of the actual image file. Therefore, any work tagging faces in these systems is currently lost if the images are exported to another system. This, in turn, removes much of the incentive to spend any time tagging faces in collections of old family photos, since, someday, all the work will probably need to be repeated.

In November 2010, however, the Metadata Working Group (MWG) defined a standard for tagging faces in images and storing the information in the XMP metadata tags of image files. XMP is the eXtensible Metadata Platform proposed by Adobe for storing metadata in image files. This standard has gained quite a bit of support, both in Adobe's own products, and in many other products as well. The MWG's standard for face-tagging, therefore, has a chance at being widely adopted in the industry.

Under this specification,

- A region in an image is specified as a rectangle, using a center point, a width and a height, using "relative" (0..1) coordinates. Alternatively, the region can be defined using just a center point, or using a circle, specified using a center point and a radius (using the smaller of the width and height of the image for its relative coordinate).
- The original image width and height is kept.
- Compliant "changers" have a specified way they should handle rotations, cropping, scaling, etc.

Once a small number of software products adopt the face-tagging standard, users would be free to use those, with some amount of comfort that the work they are doing will transfer when the times comes that they want to move or copy their images to another service.

One further extension that needs to be made to this face-tagging spec, however, is the inclusion of some sort of *external identifiers*. For example, when a face is tagged in Facebook, the user can just type text, or they can select a "friend" from a list. If they do the latter, they do more than just save themselves a few keystrokes—they inform the system of which real person they're talking about. It isn't just a "Bob Smith", but it is that *particular* Bob Smith (i.e., the one who has that Facebook account).

In this case, along with the knowledge that there is a "Bob Smith" at a particular rectangle in a photo, Facebook should also include a URL that uniquely identifies that particular Bob Smith user in the Facebook system so that there is no ambiguity as to which Bob Smith is identified. A similar strategy is used by several sites that allow association of a face with an individual in a family tree (e.g., Photoloom, 1000memories, etc.)

A face tag should accommodate any number of external identifiers, indicating who the person is in various systems, including Facebook, Photoloom, FamilySearch, Ancestry, etc. This removes ambiguity, and also allows various systems to automatically attach faces to individuals in their database without users having to repeat such work if it has already been done.

4. Preservation

Even if a collection of photos are organized well and tagged properly, it is a serious challenge to make sure these images will be preserved in the face of perils such as the following.

1. Hard drives can fail, so a backup is very important.
2. Houses can burn down, or computers can be stolen, so offsite backups are important.
3. Media can degrade. CD-ROM disks begin losing bits after a few years, for example.
4. Data formats and forms of media change over time. JPEG may not be the big thing in 20 years. 5.25" floppy disks (and drives that read them) have become nearly extinct, and CD-ROMs and DVDs will someday go the way of VHS and floppy disks. So data needs to be migrated to new media and new formats over time.
5. Any service that depends on continued payment of fees can't work long-term, because users eventually pass away.

6. Apathy or ignorance. When you pass away, your computer may get wiped clean before being donated to charity, and it is unlikely anyone will continue paying any subscription fee that is required to keep your archive preserved online. Your children or grandchildren may not know or care that you have such a precious set of images. As an example, one lady I know went to her grandfather's house after he passed away, and before she could get there, her sister had thrown away the large set of journals that her grandfather had kept throughout his whole life. Not everyone sees the value in these things.

The first four issues have solutions if you are diligent, but the last two are especially problematic. There are several possible approaches that might help here.

- ***Benevolent organization.*** An organization could possibly offer to preserve users' digital artifacts (and their arrangement and tagging information) long-term for free. 1000memories.com, for example, has at least stated a strong commitment to long-term free preservation, backed by the Internet Archive. FamilySearch or some other non-profit organization may also be able to provide this sort of service.
- ***Prepaid service.*** Another option is for a commercial organization to offer pre-paid indefinite storage, similar to purchasing a cemetery plot. The importance of it being pre-paid is that it still works even if you pass away. Storage will get cheaper over time, so the long-term cost of storing user's data will tend to go down, allowing such a business model to have a chance. Still, though, nobody can guarantee that their company won't go out of business eventually, so it might be best if such an organization's data was backed by some additional partner.
- ***Lots of distributed copies.*** Another approach is to have several copies of each photo distributed around the internet at several online sites, plus copies on the hard drives of several relatives. Perhaps not everyone has a copy of every photo, but enough copies are around that they are likely to survive hard drive crashes, house fires, theft, apathy or other dangers. For this to work, unique identifiers would need to be assigned to each image (and embedded in the metadata in a standard way) so that it was clear when two images are copies of the same original. Perhaps a distributed model similar to that used by source control systems like Git would work. Having centralized registries for such resources would help with making them findable and with deduping them.

There may be other approaches or business models that would make this a possibility as well. Whichever approach is used, one thing that is needed for interoperability is industry standards on how to represent arrangements, how to uniquely identify an image, and how identify who is in the photo. In addition to just preserving the raw images, it is important that the archival metadata (physical and logical arrangement information and especially face tag data) is preserved and shared via industry standards as well.

There are several additional issues that make long-term preservation of personal digital artifacts a challenge, including issues around *copyright*, *privacy* and *inappropriate material*. These might be addressed at least partially by license agreements that make it clear that users are responsible for the images they upload, and by being responsive to community requests for removing or restricting images that are illegally copied, that violate the privacy of individuals in the photo, or that are offensive.

Another problem is that *links break*, so any URLs used to identify photos, individuals tagged in images, etc., may not always work. Sometimes resources may no longer be hosted by the same organization, for example. Supporting links long-term is much more a matter of *organizational fortitude* than it is of technology. However, an approach like that defined by *Archival Resource Links (ARK)* adds one more line of defense against broken links. The concept is that if a resource moves from one base URL to another (e.g., because an organization went out of business or was forced to change their URLs), then a "resolver" could look at a permanent identifier that is part of the URL and use it to determine what URL to use to find that same resource in its new home.

Another source of broken links is when a company deletes resources or goes out of business. A line of defense against such problems is for an external organization to hold a copy of the company's images in "escrow", keeping them safe and with a license agreement that allows the organization to serve up the

images in the event that the original company can no longer do so. This, coupled with a "resolver" mechanism like that used by ARK, can help resources survive long-term.

5. Sharing

Currently a lot of image sharing does happen, but it is mostly ad-hoc. Images are sent via e-mail, posted on social networking sites, or mailed via DVD-ROM. Only a subset of images are usually shared, and often they are scaled down to lower resolutions for easier transport. Often duplicates accumulate over time, so again, a standard way of embedding a long-lived identifier would help identify duplicates.

It is hard to know which photos in another person's collection are relevant to you, especially if they are only distantly related or not related at all. The use of face tags with external identifiers would make it possible to discover images that have appearances of your relatives in it. In addition, if images are online and can be linked to from collaborative tree sites such as new.familysearch.org or other shared tree systems, then finding a relative in such a tree can help you then follow links to all the photos that have been linked to that relative.

Again, standards around face tags with embedded long-lived external IDs help to connect photos of individuals with users who care about them.

In addition, embedded metadata that identify where this photo appeared in the original physical arrangement and additional logical arrangements would allow users who obtain a copy or subset of the images to see how these images relate to each other, and would help to remember the *provenance* or history of the photo. An embedded long-lived unique identifier for the photo would help users know if they already have a copy of this photo. It would also help applications or services to synchronize or accumulate metadata that has been added to various copies of the same photo (e.g., face tags done independently on separate systems).

6. Conclusion and Call to Action

Organizing, tagging, preserving and sharing photos and other artifacts is important to a very large number of users and their eventual descendants. In order to preserve digital artifacts and the knowledge that surrounds them, this paper has identified several things that need to be done, including the following.

- Long-term free or pre-paid storage solution.
- Wide adoption of XMP/MWG face tagging standard for interoperable face tagging.
- Ability to embed external identifiers in MWG face tags.
- Definition and adoption of standards around physical and logical arrangement hierarchies.
- Definition and adoption of standards around unique identifier embedded in images.

Certainly there will be innovative products and services that will help users capture, preserve, organize, tag, share and search these resources. However, addressing the above issues would create an ecosystem in which work done using one product or service can transfer to others, allowing collaboration and truly long-term preservation of these precious resources and the knowledge that surrounds them.

References

[1] Wikipedia, *Archival processing*, accessed 11 Nov 2011. *http://en.wikipedia.org/wiki/Archival_processing*.