

Digital Copies and a Distributed Notion of Reference in Personal Archives

Catherine C. Marshall

Microsoft Research, Silicon Valley

Some consider the ability to make perfect copies one of the primary advantages of digital media. Copies ensure the safety of digital belongings—they can be stored in many places simultaneously—as well as rendering individual items easy to share, access from mobile devices, and reuse. Digital creation relies on the facile production of copies: we often make a copy before we make any changes we consider controversial; we take dozens of digital photos of the same scene to get the picture we're after; and re-use is considered part of contemporary art. Copies are also an incidental by-product of standard practice; for example, we might download the same file multiple times, adding a new copy to the local file system each time we view a particular artifact.

Thus, as time goes by, our personal collections of digital media become rife with copies, exact, modified, and partial.

Is this proliferation of copies necessarily problematic? It may not be. A professional archivist may advise us to simply keep them all if there is any question about the nature of the variations or their importance. Differences can be noted and preserved. Computer scientists may give us the same answer for a different reason: storage is cheap, and de-duplication algorithms can eliminate irrelevant copies at any point in the future (see for example Fetterly et al. 2005, Hoard and Zobel 2003, or Shivakumar and Garcia-Molina 1996).

In the realm of personal archiving, however, these copies may introduce unanticipated difficulties. For ordinary people, benign neglect forms the linchpin of a pragmatic archiving strategy (Marshall 2008). That is, little sustained effort is applied to digital stewardship, and it's

probably better that way. Digital belongings are simply kept (e.g. in a personal computer's file system, on removable media, on a server, by uploading them to a social media service, or by using an extended combination of different types of stores) because it's too onerous to do otherwise. Furthermore, people are notoriously poor judges of what they'll want later in their lives, and they're not necessarily technologically savvy when it comes to choosing among formats and storage media.

In other words, in an era of multivalent communication (text messages, Twitter feeds, Facebook walls, email messages, Skype conversations) and promiscuous republication (on sites such as YouTube, Flickr, blogs, and so on), when transient forms have become permanent and artifacts have become too numerous to evaluate individually, *it is easier to keep things than to cull them*. It is rare for people to go through their digital photos one by one rather than dumping them from the camera's storage to a local folder, and picking out a few to name and share. It is rare to keep track of the proliferation of these items; it is easy enough to make another copy if one is needed.

Everyone has plenty of digital belongings—stuff we own individually or communally—and we store these assets wherever is convenient and suits our overarching purposes. We don't worry much about the copies we make, or their provenance.¹ We just store the stuff and plan to sort it out when need dictates with the idea that it'll be easy enough to find the thing that we want later down the line. The good copy will be stored in one of our caches of such files—we'll instantly be able to distinguish the in-focus photo from the out-of-focus version, and the one with

¹ Neither professional archivists nor computer science research have ignored provenance (for example, see Muniswamy-Reddy et al., 2006); however, mechanisms to maintain provenance have generally been omitted from commercially-available operating systems, probably because the overhead of such bookkeeping has been considered unwarranted.

red-eye reduction from the unretouched version—and we'll know exactly what we want when we see it.

Where, then, does the “real” or reference copy of any digital artifact reside? Which version(s) do we consider archival? How will we find this reference copy when we once again want it?

Of late, systems developers say things like, "the truth is in the cloud." What they mean is that today the real, authoritative version of the data is stored on the network, possibly in a centralized repository, and that all local copies are acquired by rationalizing them with the cloud-based copy. In the most familiar case, Outlook email works this way. Your local desktop and laptop computers may each have a copy of your Inbox, your calendar, drafts of your messages in progress, your address book, and so on. Newer versions of data (updated address entries, for example) make their way to the server-side storage, and eventually replace all of the local copies that have been rendered obsolete by the change. When you want an item, you don't need to think about where it is—the data is replicated and, if everything is working correctly, the authoritative version is wherever you are.

But does this distributed systems model reflect the real situation for all of our digital assets? Is the truth actually in the cloud?

We might want to hedge our bets. Distributed systems research tells us that even when we rely on the authoritative or reference copy of the data that resides on network servers, the copy with the most recent updates—in some sense, the 'truest' copy of the data at a given point in time—may be somewhere else, on a cell phone or on a personal laptop; of course, under the

proper conditions in a replicated systems architecture, the updated copy will eventually propagate to the network store, and then to the various local copies.

But evidence from the field shows us that something else, something far more radical, may be happening in the face of imperfect replication, manual copying (the usual state of affairs), an ecology of many interacting systems and applications, and the social use of digital data: *the authoritative copy of any of our personal digital belongings is spread among many stores, and what we regard as the authoritative, archival digital item, the one we want to keep (possibly forever), has become essentially decentralized.* In other words, although it is possible to identify the digital original and all of the copies that have been derived from it (using timestamps and watermarks, for example, to trace the production of a digital original and all of the copies derived from it, or by adhering to standards that implement a more complete notion of digital provenance), it is not necessarily straightforward to point at the copy that we want to keep.

How does this happen? Many serious amateur photographers would identify the digital original as the version in the camera's native RAW format that was stored on the camera's memory when the photo was taken. But if pressed, the photographer might allow that there are important enhancements above and beyond the digital original that she might want to consider as part of the original: the metadata she has added, the reactions her peers have contributed to the photo-sharing site where she published the photo, the annotations her friends have made explaining and describing the photo, and the content modifications she has made (for example, if she has cropped the photo) to express her artistic intent. These enhancements, spread among different copies of the photo, all represent changes she'd want to save (or use as a basis for producing further copies). Thus it becomes less clear which copy is the authoritative one and which store holds it.

We might characterize this situation as one of distributed authority or decentralized reference, although most people don't acknowledge it as such. At first glance, it does not seem particularly problematic, and people choose a copy in the abstract to consider as authoritative. For example, they might think of the copy on their desktop computer, the version managed by an application designed for this purpose, as the one that is the version they would back up and archive.

There's plenty of evidence that reveals this simplification to be a think-o². One manifestation appears as people fall victim to circular reasoning: 'I don't have to worry about the copy of this photo that I've stored in a network service—I have the authoritative copy on my local disk drive.' At the same time, the person might also be thinking: 'I don't need to back up my local drive. All of the important photos are out on the network, where I've put them to share them.' But the problem is much worse than that. What's really happened is that the version on Flickr has the best descriptive metadata; the version on DeviantArt has the best photographic metadata; both shared copies have attracted valuable (and not-so-valuable) comments and annotations; the version on the camera's flash card is the only one at full resolution; and the version on the local disk is the one to which transformations have been applied to 'correct' the base photo (i.e. rotation, cropping). Some transformations are best thought of as archival only in the abstract: a better red-eye reduction algorithm might come along in a couple of years, and that would be the one the photographer ultimately wants to apply in lieu of the original correction.

In short, the way we work with and share digital media, documents, and datasets has left us with complicated notion of which copy is the reference copy. To return to the systems developers' perspective: The truth is not in the cloud; rather, the truth *is* a cloud.

² A think-o is the conceptual equivalent of a typo.

To tease out the nuances of how copies complicate our personal archives, first I will discuss some of the reasons that people copy files, and will chart an observed case of copying in practice. This evidence in turn is used to form the foundation of a taxonomy of copies that builds on existing abstractions such as the model presented by Currall, Moss, and Stuart (2008), with the pragmatic aim of ensuring we handle the observed cases. I will conclude with a brief look at the implications of these stored copies, both in terms of technology development, and in terms of stewardship.

Why people copy files

People copy files for a variety of reasons. Not all copies are created equal; they are often created with different purposes in mind, and these purposes guide the degree to which the copies are identical. There are at least five reasons people copy files: (1) for use on different personal devices; (2) to share or publish the files; (3) to backup the files as part of a ‘best practices’ IT regimen; (4) to create a permanent archive of digital assets; and (5) to re-use the content without modifying the original.

Why should we care about the various motivations behind the production of digital copies? After all, the result is more or less the same, or so it seems. But if we look more closely, each form of copying has different consequences. For the sake of this argument, we’ll concern ourselves with two kinds of copying in particular: copying files to use them on different personal devices (say, on a desktop computer, a laptop computer, a phone, and an MP3 player), and copying files to share or publish them. It is these two reasons for copying files (or more generally, any kind of digital items) that are the most apt to lead to interesting types of diversity and—potentially—useful convergences as the different copies are brought back together.

When copies are spread across devices, it is usually to edit (or simply use) the content itself rather than its metadata. We usually (but not always) want these content changes to propagate so that all devices end up with the most current version of the object. In systems research, we refer to this cycle of replication, change, and synchronization as achieving *eventual consistency* (Ramasubramanian et al. 2009). That is, we would like all of the copies to end up in sync again after editing has taken place.

Ideally, the replication and synchronization of files is managed by a systems infrastructure that uses some combination of specialized metadata and/or logs to control the propagation and reconciliation of different versions. For example, systems such as Cimbiosys (Ramasubramanian et al. 2009), Ficus (Guy et al. 1990), PRACTI (Belaramani et al. 2006), Perspective (Salmon et al. 2009), and EnsemBlue (Peek and Flinn 2006) all implement mechanisms and protocols to replicate and synchronize items among devices. Further research takes into account the differences in object fidelity required by different devices and different use situations; for example, Polyjuz (Veeraraghavan et al. 2009) allows users to maintain parallel stores of objects at different fidelity levels.³ Meanwhile, as systems research progresses, field investigations of these mobile devices in use reveal that people often synchronize the devices by hand, primarily because they don't quite trust the outcome of the automated replication processes (Dearman and Pierce 2008).

This lack of trust can't be attributed to bad software; usually the software works in a predictable way. More often trust breakdowns arise because it's difficult to match operation

³ What this means is that, for example, an address book record in your email may use many fields to describe an entry; the comparable record on your cell phone may only store a small subset of these fields. Or a photo may be high resolution on your desktop computer and lower resolution when it is stored and displayed on your digital photo frame. A mechanism is used to synchronize changes to the parallel stores.

semantics and human expectations. Although simple operations follow a carefully articulated set of rules—deletes propagate, for example, so that deleting an object on one device has the expected outcome of deleting it from the corresponding data stores on every device—people are often less than delighted with the outcome of inter-device synchronization. Furthermore, to people with different backgrounds, the whole process is unpredictable and uncontrollable: synchronization is more or less magic. As with other aspects of computing, *things just happen*. Thus people take the management of copies across different devices into their own hands. Smaller-scale synchronization processes—say, between an iPod and a Mac, mediated by Apple’s iTunes software—are used (and are useful), but they exist in isolation, and don’t extend further than they absolutely must. Automatic synchronization is regarded with skepticism and is tolerated rather than celebrated.

This propagation of versions is further complicated when changes come into conflict. That is, if one copy of an object is edited on device A and another copy is edited on device B, the conflicts need to be resolved. Sometimes conflict resolution is simply a matter of merging the changes. For example, if I’ve edited the introduction to a paper on device A and you’ve edited the conclusion on device B, it’s pretty obvious how the conflicting changes may be resolved; the new copy should have my introduction and your conclusion, although the resulting version of the paper may no longer make sense. Likewise, if red-eye reduction has been performed on a photo on device A and rotated 90 degrees on device B, it is easy to resolve the conflict by combining the operations. On the other hand, if you and I edit the same passage on the two devices, or rotate the photo in different directions, manual intervention may be required to resolve the conflicts.⁴

This model of eventual consistency (wherein changes to copies of a collection propagate

⁴ Thousands of person-hours may be sunk into automatically reconciling concurrent changes. Microsoft Word has such a facility that attempts to integrate the concurrent changes made to individual copies of a document.

throughout a network) and the need for conflict resolution mechanisms drives aspects of distributed systems research.

Thus replicating files among one's own mobile devices introduces a set of thorny problems as the content of the parallel files potentially diverges and must be reconciled.

On the other hand, copying a file to share it or publish it is apt to lead to other issues as metadata diverges, and other content changes are made that are not intended for propagation to all copies (for example, content resolution is reduced). When files are shared or published, descriptive metadata may be added to reflect the needs and interests of the audience (for example, tags and captions may be added to videos or photos); social metadata also accumulates as the digital object develops a life of its own online (for example, viewers may add ratings or comments; the service itself may keep track of the viewers' characteristics—where they are from, or more commonly, the number of viewers a particular online object attracts).

The content changes that accompany publication or sharing are usually different than those made when copies are replicated among one's personal devices, which is why people don't necessarily think of the published copies as archival. Resolution may be reduced to conserve either storage or bandwidth or both. Changes may be made for the sake of intelligibility or privacy (for example, a photo may be cropped or an email response may be quoted to remove the personal portion of the correspondence). Hence we see different kinds of derived files being produced.

Publication—even informal publication to the cloud—is associated with fixity (Levy 2001); inter-device replication is associated with fluidity and changing content [see for example (Ramasubramanian et al. 2009)]. The reverse is often true for metadata. The idea of replicating

content among devices is to arrive at consistent metadata (reconciliation is part and parcel of synchronization). Table 1 summarizes this tension between the reason the files are copied, and the effect this motivation has on content fixity and fluidity as well as metadata accumulation and divergence.

Table 1. The effect of copying on content and metadata

	inter-device replication	publication and sharing
content	changes ideally propagate and are reconciled among copies (so they are eventually consistent)	changes usually propagate in only one direction (to the publication site)
metadata	metadata converges (by design)	metadata diverges

Deriving a pragmatic taxonomy of copies in action

To investigate the effect of copying digital content, it may be productive to look at a real-world example. About three years ago, I interviewed an animator who had just finished art school in southern Taiwan. Taiwan was not her homeland, so while she was a student, she published a widely read blog about her life and adventures, emphasizing cultural differences and her experiences exploring the Taiwanese countryside. In addition to personal journal videos reporting current events and describing sights she'd seen, she included short animated works in her blog.

Let's look at one of these animations, a music video, focusing on how she published and shared it, how she and others reused it, and how she tried to protect herself against the vagaries of digital storage. Earlier she had lost a series of podcasts because a social media site (primary storage for the large audio files) had gone out of business abruptly, without adequately warning its users; this experience made her wary, and prone to store extra copies in various hospitable places on the Web. We'll also consider the copies that turned up in different places, reused by others.

The animated work I have selected for this example, *Satellite*, was produced as a music video to go with a song by one of her favorite bands, The Motion Sick. The band liked the video enough to reuse it to promote their music.

For our purposes, I'm going to divide the music video into three different kinds of data: (1) the primary data, which consists of the video itself and all derived forms (i.e. individual frames stored as still images) and different versions of the content; (2) descriptive metadata used by the artist and other publishers to characterize the video (e.g. tags, narrative, structured attribute-value pairs); and (3) social metadata, the structured and unstructured metadata (e.g. number of views, comments, and ratings) that accumulated as the music video was used on a variety of sites.

First let's examine three important aspects of the primary data (the content). The first is *versions*. In this case, the band published an earlier (preliminary) version on their website to whet the appetite of their fans; this version was not maintained once the final version of the music video became available. The second is *variations*. Variations are published in parallel, rather than being replaced.⁵ Finally, there are *derived forms* such as the photos and thumbnails.

During our interview, the artist said that she made copies to protect herself from loss. As of this writing, there are at least ten full copies of the video on social media sites; three of the copies are variants that include 10 seconds of title frames and credits.⁶ A 3:06 variant is found

⁵ It's easy to see analogs in other endeavors. For example, in scholarly publishing roughly the same ground might be covered in a technical report, a conference paper, and a journal article. None supersedes the others. The technical report might have pseudocode that wasn't published in the other two forms; the conference paper might be an important stake in the ground, published considerably earlier than the journal paper; and the journal paper might have an evaluation section that is a substantial contribution beyond the results reported in the conference paper.

⁶ There may have been other copies of the full content, but for this example, we'll focus on the six that were easily found using search engines. The artist may have additional copies on old computers and on removable media.

under the artist’s name on the Internet Archive in three different formats (QuickTime, Ogg Video, and MPEG4); an identical video is found on YouTube under the artist’s name (and stored with her other videos); a third identical video is found on Google Videos (playable from the site, but downloadable in MPEG4); and yet another copy is found on awnTV (a video site devoted to animations). The 3:15 variant is copied on three sites, and seems to be primarily used and propagated by the band. This variant is on the band’s MySpace site, on YouTube under another identity, tubetubetube99, and on Facebook on the Motion Sick’s fan page. There is also an earlier version of the video on the band’s website; it is stored on a temp directory, so there is some evidence that the band intended to replace it.

If we consider the artist’s intent—sharing and saving her short film—the nine complete copies are comparable (albeit in different formats, which may involve lossy transcoding operations) even though they are not the same. It would be easy to reproduce the segment that renders them inexact copies or to reverse the transcoding operations.

Table 2 compares the descriptive metadata for the seven different sites where the ten videos are stored. The animator’s multiple online identities are consolidated as *artist*, and the band’s online identities are referred to as *band*.

Table 2. Comparing metadata on seven websites where the music video is stored.

ID	site (depositor; creator)	description	tags/keywords	thumbnails	social metadata
C1- C3	Internet Archive (<i>artist</i> ; <i>artist</i>)	a short music video featuring members of the Motion Sick fighting evil forces. Please visit <band’s URL>	music; music video; music videos; the motion sick; motion sick; satellite; taiwan; misadventures in taiwan; rock; rock music; band; animation; cartoon; experimental; indie	7 frames grabbed at 30 second intervals	870 downloads; no reviews; no rating
C4	YouTube (<i>artist</i> ; <i>artist</i>)	an epic battle between good and evil. Music by The Motion Sick (<band’s URL>) -- the rotating cube portion in the middle was made with Google	satellite; themotionsick; musicvideo; music; rock; indie; sketchup; animation; cartoon; cutout; cutouts; instruments; flash;	NA	11,397 views; 4 comments; 3 ratings (5 stars)

		Sketchup	13aiwan		
C5	YouTube (<i>band; artist</i>)	The Motion Sick's music video for the song Satellite made by animator Gem	The Motion Sick; motion sick; satellite; gem; animation; aliens; battle; fight; space	NA	2,390 views; 1 comment; 2 ratings (5 stars)
C6	Google Video (NA; <i>artist</i>)	music video of the song Satellite by The Motion Sick <band's URL> ----- made using 2d computer animation, cut-outs, a little bit of Google Sketchup (the rotating cube portion in the middle of the song).	NA	15 frames grabbed at 15/20 second intervals	NA
C7	MySpace videos (<i>band; artist</i>)	Music video for The Motion Sick's song Satellite by Gem	indie; music, rock; Star; the; video; Wars; motion; Sick; Space; satellite	NA	46 views no comments rating 92%
C8	awnTV (<i>artist; artist</i>)	This is a music video made for the band The Motion Sick. It features the band members in an epic battle between good and evil. Animated with cut outs.	NA	NA	3 votes; no comments; 3.33/5 rating
C9	Facebook (<i>band; artist</i>)	The Motion Sick - "Satellite" Music Video by Gem by The Motion Sick (videos) 3:15 The Motion Sick - "Satellite" Music Video by Gem http://www.themotionsick.com http://misadventuresintaiwan.googlepages.com/films	NA	NA	(not accessible to non-fans—probably the main active locus for social metadata)

A human viewer can readily ascertain that (a) these are copies of the same video, with and without the short title/credits sequence; (b) they were probably uploaded by two different entities, one on behalf of the artist (likely by the artist herself) and a second on behalf of the band. Only one of the sites gives the viewer the ability to contact the artist directly, and only one of the sites (a different one) gives the artist's complete identity. Two of the sites readily support downloads (and therefore, the creation of new copies); the others only support embedding.

The descriptive metadata differs, implying that it was re-created each time the music video was uploaded, presumably to reach the sites' distinct audiences: fans of the band, who are directed via a URL to the band's website; fans of the artist, who might want to know what the video is about ("an epic battle between good and evil"); and fellow animators, who might want to know how the artist made the video ("made using 2d computer animation, cut-outs, a little bit

of Google Sketchup”). Two of the content stores use automatically-derived thumbnails as a reduced representation of the video’s content.

The tags, which presumably tell us a bit about how the depositor expects the video to be accessed, overlap, but again reveal that there is more than one audience for the video. In fact, they reveal yet another audience: the regular readers of the artist’s blog. They refer to everything from the material’s genre to the band’s and song’s name to the blog’s name to the story told in the video to the animation technique and tools used.⁷

The social metadata—the ratings, comments, and popularity, for example—differs. The artist has indicated that she’s interested in such feedback (she’s responded to some of the comments, has submitted the video to juried sites, and has said in interviews she attends to such things). Evidence shows that she is slowly accumulating this social metadata. But is it all equal? Differences in audience (does a view from a fan of the band matter as much as view from one of her own fans or from a peer?) and accessibility (some of the hosting sites are less popular than others) make aggregation of the values less straightforward than it would initially appear. The (sparse) comments on the two You Tube sites reveal that the single comment on band’s channel is directed to the band (“This is a cool band”). Or is it? The comment was made by *iluvuartsynerd*, who is also a video-maker, albeit in a different genre. It is easy to see how social metadata has an organic form, shaped by media genre and audience; it is less straightforward to evaluate its utility.

Many of the sites that refer to the video—the artist’s own blog, for example, or podcast directories—embed the video content from another storage location. Videos are necessarily large

⁷ It goes without saying that the example illustrates the need for a mechanism to normalize the tags (Guy and Tonkin 2006).

and difficult to move around; thus a video stored on several sites like this one is apt to have many more references to it than actual copies. These references may provide considerably more descriptive and social metadata than the storage sites, since they rely on the strength of this metadata to attract viewers.

I have used the rankings of two popular search engines, Google and Bing, to gather a set of sample copies of the video and references to it. There are no doubt many more than I have collected, but it is impractical to document all of them, and the sense of how they diverge is adequately conveyed by the smaller sample. One of the sites has six different references to the video, without acknowledging that they are duplicates or near duplicates, since it is likely they are de-duped by eliminating references to the ‘same’ video (that is, videos taken from the same URL). Table 3 shows some representative metadata that is used by these copies-by-reference. In some cases, the artist has added the reference to the site; in other cases, someone else (the site’s owner, the band, or a fan of the artist or the band) has added the video and created (most of) the metadata. As before, the social metadata accretes and is contributed by a number of different viewers.

Table 3. Example copies using embedded content

ID	site (depositor)	stored	description	tags/keywords	links	social metadata
R1	The band’s website “The Motion Sick” (<i>band</i>)	C5 (YouTube)	Satellite music video by Gem	NA	Archive.org Google videos YouTube v2	NA
R2	Gem’s animation archive (<i>artist</i>)	C4 (YouTube)	music video made for The Motion Sick; animated with cut-outs, photographs, and Flash, 3 minutes 05 seconds, 2006	Animation, Films, Moving Images, cutouts, music, stop motion	YouTube v1; The band’s website	no comments
R7	ReAnimacija Festival (<i>artist</i>)	(no embedded version)	Title: SATELLITE (SATELLITE) Category: Panorama Director: Gem Urdaneta Producer: Animation Institute Animation: combination Country: Taiwan Production Year: 2006 Duration Time: 3’05” Synopsis: A music video for the Boston	NA	NA	NA

			based band The Motion Sick, featuring the band members in an epic battle between good and evil.			
R9	Blogger MIT (artist)	C4 (YouTube v1)	Satellite (a music video) 12-31-2006 update: Hooray! Satellite's been selected as part of "music videos and advertising" for the 2007 Festival Bimini in Latvia! (here's the link -scroll down, until you see "original title: Satellite/ Country of origin: Taiwan". that's the one!) ----- Satellite (music by The Motion Sick) (I made this from September to October 2006.) [also 6 thumbnails, grabbed manually]	animation, podcast, video	Bimini Festival, YouTube v1; Google video; Internet Archive (QT version) Thumbnails link to full size screen shots in zoommr	6 comments (2 are responses from the artist) note that these are copies from the original blog
R10	The band's livejournal fan site (band)	(no embedded version)	Also don't forget about our video: --- we'd like to offer you a secret glimpse at our upcoming music video for "Satellite." There may still be some buttons to push and some levers to pull to complete it, but animator Gem has helped us to recount the events that occurred recently in our battle against an evil rogue planet to save the Earth. See the latest version here and let us know what you think: http://www.themotionsick.com/temp/themotionsick-satellite.mov (video is gone; link points to the top level page for the band; this one is signed by band member Rock)	NA	The band's website	none

Metadata-only copies are cheap, and there is considerable motivation to make them. Table 3 illustrates some of the different purposes for this kind of copy. First, both the animator and the band have creative impetus to circulate copies of the video among their fans (existing and prospective); this would account for the band's website (see R1 in Table 3), as well as their Facebook, livejournal, and MySpace entries. These are sites curated by the band and their fans. The animator curates some of the by-reference sites (see R2) herself. These are minimal, but allow her to direct viewers to the 'real' sites and to accumulate social metadata. The artist maintains a blog, which she copies manually as backup (see R9). Other copies were produced for special purposes, for example, to enter them in animation festivals (see R7); although these copies seldom provide access to the actual video, they often have authoritative metadata, including the artist's full name, and production information that there is less incentive to include on the more informal sites.

Of more dubious value are the additional metadata-only copies produced by fans, viewers, or commercial sites which hope to draw attention and attract advertisers. For example, six metadata-only copies of the music video reside on a single commercial site,⁸ each originating from one of the sources shown in Table 2; the metadata has apparently been scraped directly from the source site.

Finally, one last type of copy to consider is a *derived form*. In this case, the artist has published stills from the video on several different sites as independent artistic works. Derived forms are neither versions (for example, the early form of the video, now removed from the band’s site) nor variants of the work (for example, the transcoded forms on the Internet Archive, or the band’s 3’15” music video), but rather take some other form with different characteristics.⁹

Table 5 summarizes two examples of derived forms using their metadata. One—a still—was stored in Deviantart, an art sharing site, and other stills were stored in Zoomr, a photo sharing site for professional (or serious amateur) photographers. By her own description, the artist also maintains a Flickr account, but has not published the stills there. The copies cross genres to reach another audience.

Table 4. Two sites with derived forms (frames the artist has selected from the video)

ID	Site	title(s)	refers to	caption	tags (union)	social metadata
D1	Deviant Art	Satellite	google video	The third one was made from September to October of 2006	NA	32 Deviations 15 Comments 1621 Pageviews
D2	Zoomr (gem)	screenshot from Satellite (x6)	YouTube v2	screenshot for my latest video	{misadventuresintaiwan, screencap, screenshot, animation, drawings, works} x 6; {themotionsick} x3	views 527, 497, 328, 376, 352, 229

⁸ They are referred to later in the chapter, in Figure 2, as R11-R16.

⁹ As a practical definition, the derived form *can* be produced by the source form, but the other direction (constructing the full animation from the few frames) would not be possible.

At first glance, that the information offered by this accumulation of copies skews toward different audiences seems insignificant. From the film festivals, we find the artist's full name and where she went to school; from the band's websites, we learn more about the relationship between the artist and the band, and about the video's topic; from the artist's sites, we learn more about the techniques she used to make the video. The stills give us insight into the portions of the video that the artist likes the best, or at least finds the most distinctive. From the fact that the music video is stored so many places, we discover that both the band and the animator value the film. Certain information (where and when the film was produced) is verified by the apparently independent metadata used to describe it. By the same token, we can infer that other metadata isn't necessarily reliable since it has simply propagated from site to site.

To make something of these distinct entries—to aggregate and verify the description, or to offer a summary of the social metadata, which is sometimes duplicated and sometimes independent—we would need to be more methodical and algorithmic in our gathering and harvesting.

Let's look, for example, at the tags. Tags are associated with many of the copies I gathered (including the derived forms). In aggregate, the video has 172 tags, many of them duplicate or non-normalized forms of the same tags. If we normalize the tags (for example, consistently using the singular or plural form; using spaces to delimit words in multi-word tags; and using a single tag for the band's name), we find 37 unique normalized tags, only 9 of which are used more than six times: {animation; motion sick; misadventures in Taiwan; satellite; rock; music; cutout; indie; music video}.

Would we be doing a service or a disservice were we to aggregate the tags (as, say, part of an archiving process)? Some are relative to the audience. That the video is part of the artist's blog series, *Misadventures in Taiwan*, may be useful to fans of the artist's blog, but not the band's fans (nor in some cases to other artists, who may only be interested in the artist's animations, and not her personal diary podcast entries). That the band, The Motion Sick, provides the music for the animation may not be relevant to the artist's peers, and perhaps not even to her fans. The techniques used in the animation, including cutout, is probably of greatest interest to the artist's peers; others might not even be aware of what this technique entails. On the other hand, that the genre is an animation indie rock music video and that the music video's name is *Satellite*, is probably of greater interest to the general indie music audience—perhaps even including people who are looking for new music in a genre they enjoy—than it is to people familiar with either the artist, the band, or both.

This, then, is the puzzle: how do we arrive at the version of the content and metadata that we wish to archive? It's not just a case of conflict resolution (for the content) and metadata aggregation (for the published objects).

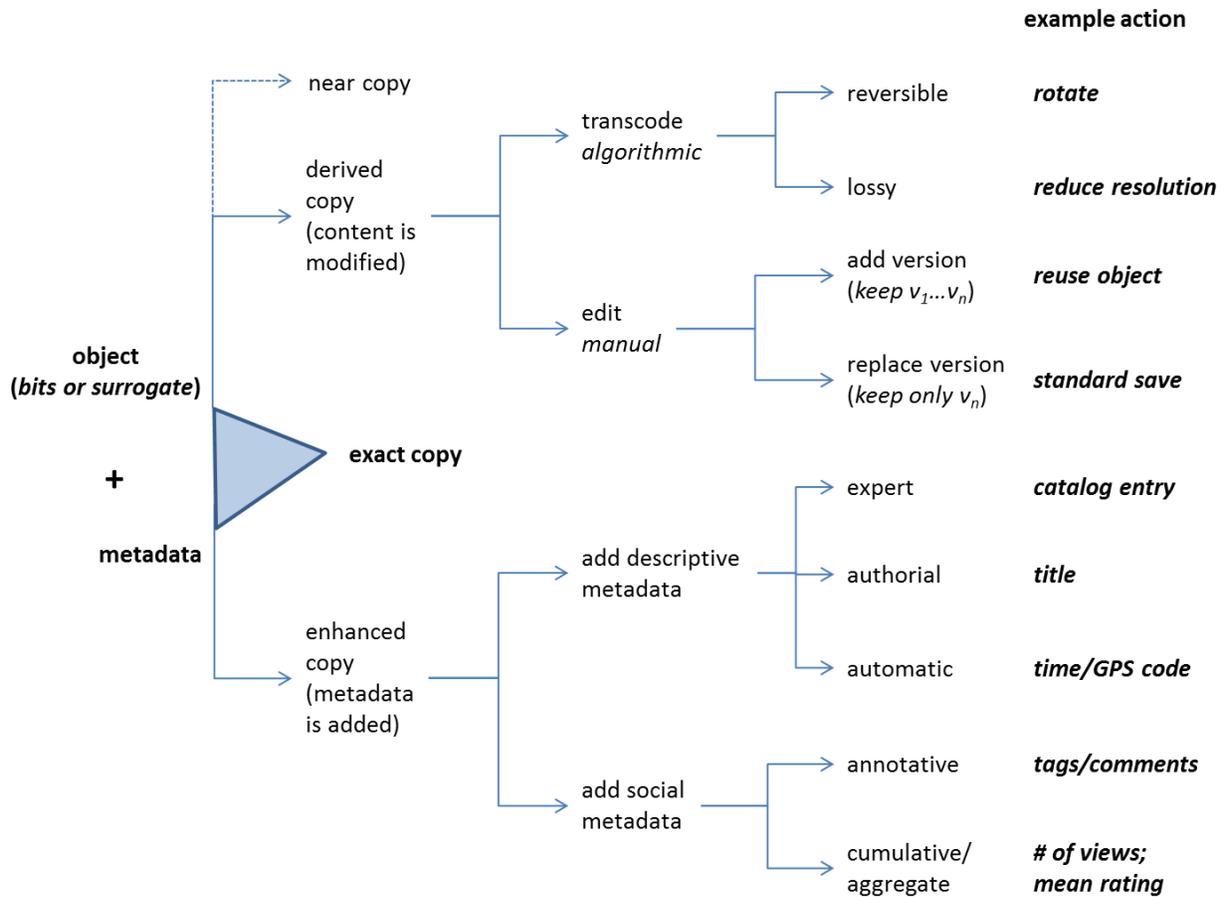


Figure 1. Observed copy practices

Figure 1 breaks down observed copy practices used to propagate the digital content. If a copy consists of the digital media plus the metadata associated with it, and the digital media is either the literal object (i.e. a copy of the bits) or a surrogate (i.e. a pointer to where the bits are stored), copies can either be exact (i.e. the bits are identical), derived (that is, made from the object), or a near copy (for example, several photos of the same scene taken in quick succession so that it is difficult to tell one from the other). Derived copies may be transcoded (that is, derived by an algorithmic process) or manually edited (that is, derived by human intervention). Transcoding may be reversible or lossy (for example, resolution may be reduced in a copy made by uploading media to a social networking site). Edited copies, which by definition result in a

new version, may either *replace* the old one (as Microsoft Word does when a file is saved), or form a set (additional versions are kept).

If metadata is doing its job, it enhances the value of the digital object. Thus we'll refer to a copy with added metadata as an *enhanced copy*. For the sake of keeping things simple, this metadata may describe the object, or it may fulfill a (sometimes vague) social role (e.g. it might be a rating or review). Descriptive metadata may be contributed by the author, by an expert (for example, a cataloger), or by an automated process (for example, a Global Positioning Sensor may contribute geographical coordinates, or a clock may contribute the time). Why make these distinctions? After all, we can easily think of examples in which the author is an unreliable source for the object's description, or cases in which a sensor has yielded inaccurate data. What these distinctions hope to tease out is the *anticipated value* of the added metadata and to suggest what we might want to do with it from an archival perspective.

Thus, the purpose of this simple taxonomy is not to split hairs, but rather to suggest modes of archival processing, given how copies are made in the first place; what people do with them; and the think-o that was introduced early in the chapter (that the digital original is on local storage and the backup is in the cloud, or vice-versa). The graph is intended to illustrate ways in which its two branches may be traversed independently to pair stored content with autonomous instances of the metadata (growing out of a surrogate, for example). It is also meant to stimulate thinking about how copies came about (for example, are they the result of irreversible or lossy transformations?); breaking down the cases this way allows one to make decisions about whether the parent node or its child is archival, and to consider the ways in which people actually handle versions (sometimes they are only interested in the most recent version; infrequently, it make

sense to keep an ordered set of evolving versions). Then we can begin to articulate heuristics for combining metadata according to its source and characteristics.

Conclusion

Copying or otherwise replicating digital material to keep it safe is an intuitive and reasonably effective personal archiving strategy; it is a strategy that is not limited by discipline or technological sophistication. Yet the results of this seemingly simple method are far-reaching. Long-standing concepts such as authenticity, authority, and reference must be re-thought. Similarly, personal archiving technologies must be designed with copies in mind.

Distributed reference copies and decentralized authority. The primary aim in gathering copies of a digital object from the services and websites on which they are stored is to illustrate an effect: *the copies take on lives of their own*. It would be fairly straightforward to develop criteria to identify authoritative instances that may serve as reference copies or useful kinds of metadata to harvest from these copies. Although some of the metadata outcroppings are intended for a particular audience, others are general, and might be useful permanent annotations for the digital media. In the case developed in this chapter, either the band or the artist (or both) may be interested in saving some of the music video's social metadata (the comments, ratings, or view numbers the video has accrued over time).

Figure 2 maps out the music video's lineage, starting from three abstract entities: the artist's original video; an early version of the video; and the band's original video. That the artist and the band each refer consistently to one version of the video or the other—the artist refers to her version without titles and the band refers to its version with them—indicates that these

versions have developed parallel authority. In other words, there is neither a single version of the content nor a single locus of metadata that serves as the reference copy.

This case helps illustrate that it is necessary to decentralize authority given the way people manage copies of published material. Not only are there parallel versions of the video; the artist has also created two blog entries, informal copies of one another (the copies were produced manually) that describe the video. Some of the metadata on the original version of the blog (referred to as MIT V1 in the graph) and realized as R17 has been copied over to R9, R16, and R19. Dotted lines in the figure indicate virtual copies; that is, R3 and R10 (early versions of the music video), and R7 and R18 (the copies the artist has submitted to competitions) are immediately available for download (or even to watch), although all four are reasonably good sources of metadata or of evidence.

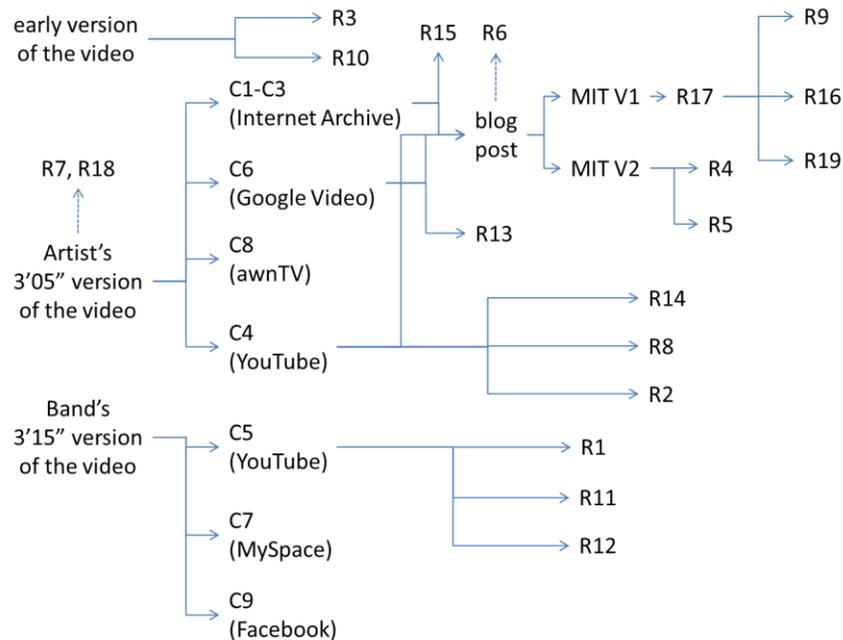


Figure 2. Where the copies came from (MIT stands for Misadventures in Taiwan, the artist's blog).

A preservation strategy. Mixed-purpose replication of digital objects has some odd effects on digital stewardship. Most people haven't thought through which version of a valuable file they would like to keep, and therefore need to maintain. The principle of benign neglect dictates that you simply keep all the files, and hope that at least one of them survives. In this case, the artist and band are both maintaining parallel versions of the video in multiple places with the implicit (or, in the artist's case, stated) aim of digital safekeeping. In an interview with the artist, she told me that she would like to save the entirety of the content that R17 represents for a long time,¹⁰ but

“the problem is i've hosted it in my school's free space for students [and] they delete the sites of students who've graduated [so] it might get deleted soon¹¹... I've backed up my videos on youtube, and the links to blog posts on Wordpress.com. The thing is many people have already bookmarked this page...I guess i can move them, like migrate the blogger page to a free blogspot page, but there are many things that will go wrong so now I'm just procrastinating... but now I have a semi-backup, misadventuresintaiwan.wordpress.com.”

She realizes that she's backed up many of the components of R17 (including the video), but she's concerned about her interactions with other peoples' archives, and the content she has archived by way of R5 (her Wordpress blog) is inexact and that her archiving method is, in her words, “nonscientific.” Other content that she had archived in the past on a cloud service, coupled with a copy on removable media (burned to a CD), proved to be disastrous; the service

¹⁰ R17 is the digital original of the artist's blog, coupled with the copies of the video she has stored on the Internet Archive.

¹¹ It is interesting to note that R17 is still in place at the time of this writing. The artist has not been affiliated with this institution for at least three years.

was taken offline without warning (as the company failed) and she lost the CDs. Although she continued to use the same strategy—burning DVDs and storing the files in the cloud—after she lost the first batch of files, she became relatively more cautious, making multiple copies in the cloud instead of relying on a single service.¹²

One other aspect of this case worth noting is that the purpose of the artist’s archive changed over time. Initially her intent was to save her blog, which centered on a popular series of audio podcasts chronicling her adventures as a student in rural Taiwan.¹³ As time went on—and as she reached the end of her MFA program—she transferred the emphasis of her archive to an online portfolio of her animation work.

A long term retrieval strategy. What happens when you look for an object stored this way (as a loosely connected set of copies and near-copies)? As time progresses, it is easy for us to forget just how many copies we’ve stashed away, where we’ve put them, and the differences among them. It is also easy for the copies to get away from us (McCown et al. 2009), and to accumulate troves of social metadata, some valuable, some not.

In the absence of a centralized archive and intentional archiving services, search engines (as they work today) are apt to return some combination of the most visited copy (often the copy with the richest metadata) and the copy that is the most common link destination. But is this the copy we’d most like to see? Moreover, if there were a centralized archive, would we be willing to give up all of the metadata that has accumulated on the non-archival sites after the archival copy has been safely tucked away?

¹² She refers to them as “half-copies” because she is unable to store all of the content on the same server.

¹³ It might be interesting to know if the artist has considered asking her listeners if they have stored her published files so she can get the ones she values back. Unfortunately I did not pursue this line of questioning.

Information retrieval research began developing de-duping algorithms once it was faced with the problems introduced by untidy corpora (for example, see Shivakumar and Garcia-Molina 1996). That is, the original test corpora used to develop information retrieval techniques and algorithms did not factor in copies or near-duplicates. The corpora were tidy affairs, unlike the pastiche of real data sources that even then existed in the real world. Then the pendulum swung in the other direction: the assumption behind most search engines was that not only did copies exist, but also that we wouldn't want to see them, that we'd want to be spared the messiness of real data stores.

In the long term, we will not only want to see copies, but we'll also want to harmonize them, to harvest their metadata, to select among them. Instead of relying on a simple notion—the truth is in the cloud, embodied as a single reference copy—we will want to expand our sense of what is entailed by the notion of a reference copy and turn to a distributed, social model.

References

- Belaramani, N., Dahlin, M., Gao, L., Nayate, A., Venkataramani, A., Yalagandula, P., and Zheng, J. 2006. PRACTI replication. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 59–72.
- Currall, J.E.P., Moss, M.S. Stuart, S.A.J. 2008. Authenticity: a red herring?, *Journal of Applied Logic*, Volume 6, Issue 4, The Philosophy of Computer Science, December 2008, pp. 534-544.
- Dearman, D. & Pierce, J. 2008. It's on my other computer!: computing with multiple devices, in *Proceedings of CHI 2008*, New York: ACM Press, pp. 767–776.
- Fetterly, D., Manasse, M., Najork, M. 2005. Detecting phrase-level duplication on the World Wide Web, *Proceedings of SIGIR 2005*, New York: ACM Press, pp. 170-177.
- Guy, M. and Tonkin, E. 2006. Folksonomies: Tidying up Tags? *D-Lib Magazine* 12, 1.

- Guy, R. G., Heidemann, J. S., Mak, W., Page, Jr., T. W., Popek, G. J., and Rothmeir, D. Implementation of the Ficus replicated file system. 1990. In *Proceedings of the Summer USENIX Conference*, pp. 63–71.
- Hoad, T.C., Zobel, J. 2003. Methods for identifying versioned and plagiarized documents, *Journal of the American Society for Information Science and Technology*, v.54 n.3, p.203-215, February 1, 2003.
- Levy, D. M. 2001. *Scrolling forward: making sense of documents in the digital age*, New York: Arcade Publishing.
- Marshall, C.C. 2008. Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field. *DLib Magazine*, 14, 3/4.
- McCown, F., Marshall, C.C., and Nelson, Michael L. 2009. Why Websites Are Lost (and How They're Sometimes Found). *Communications of the ACM*, November.
- Muniswamy-Reddy, K., Holland, D.A., Braun, U., and Seltzer, M. 2006. Provenance-Aware Storage Systems. *Proceedings of the 2006 USENIX Annual Technical Conference*, Boston, MA.
- Peek, D. and Flinn, J. 2006. EnsembleBlue: Integrating distributed storage and consumer electronics. In *Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 219–232.
- Ramasubramanian, V., Rodeheffer, T. L., Terry, D. B., Walraed-Sullivan, M., Wobber, T., Marshall, C. C., and Vahdat, A. Cimbiosys: A platform for content-based partial replication. In *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)* (Boston, MA, Apr. 2009).
- Salmon, B., Schlosser, S. W., Cranor, L. F. and Ganger, G. R. 2009. Perspective: Semantic data management for the home. In *Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST)*.
- Shivakumar, N., Garcia-Molina, H. 1996. Building a scalable and accurate copy detection mechanism, *Proceedings of the first ACM international conference on Digital libraries*, New York: ACM Press, p.160-168.
- Veeraraghavan, K., Ramasubramanian, V., Rodeheffer, T., Terry, D. B., and Wobber, T. 2009. Fidelity-aware replication for mobile devices. In *Proc. of the ACM Conference on Mobile Systems, Applications and Services (MobiSys)* (Krakow, Poland, June 2009).