On Preserving
Typecasting and Linecasting Documents
in the Digital Era

American Typecasting Fellowship
2012 Conference, August 15-20
C.C.Stern Type Foundry, Portland, Oregon
www.metaltype.com

Dr. David M. MacMillan
Mineral Point, Wisconsin
www.CircuitousRoot.com
dmm@Lemur.com
(608) 623-2286

> "... whereas the records of all other arts and trades are effected by means of typography, yet the records of its own progress are singularly deficient, and, for a trade of such antiquity, the data available are most meagre."
> - Lucien A. Legros. *Typecasting and Composing Machinery.* (1908)

Summary: We are at a critical point in the preservation of technical and historical information about typecasting and linecasting. The move to digital information has become precipitous. The print-based media used previously to preserve the craft may soon be limited in their availability and uncertain in their continuity (e.g., print libraries may soon be seen as too expensive; they are already being phased out in favor of digital repositories). Moreover, print has never served type well as a means of preservation – the bulk of typecaster and linecaster technical documentation is *already* in the landfill. Perhaps paradoxically, the move of most text media to digital presents the first opportunity to preserve in a systematic way the technical literature of typographical casting. This must be done in order to keep casting alive now that we have passed out of the era of oral tradition and craft apprenticeship. Yet there are many obstacles to this: technical literature is often not highly regarded; experience practitioners may ignore it as documenting what is obvious to them; the sometimes arbitrary and complex issues of digital preservation may be alien to those with the best access to materials; issues of copyright are changing rapidly and in ways which, if not understood, threaten the preservation of this literature in the long term; finally, in the long term, digital preservation is itself simultaneously the best and the worst way to try to save a document. Digital documents can last "forever," or vanish in an instant. This present paper presents my own personal thoughts on these matters, together with many of the details necessary to accomplish digital preservation. The good news and short summary is that it can be done, but just as with typecasting itself the details are important.

(This paper was to have been presented as the backup text for a much briefer discussion/Q&A at the 2012 Conference of the American Typecasting Fellowship. While I was unable to attend the Conference, it was distributed there.)

1. Three Principles

I find that the following three principles summarize a useful attitude toward this process:

1. Nobody else is going to do this. We must.
2. It is important to preserve even boring things.
3. Integrate digital preservation into your regular work.

Digitizing a substantial body of printed literature and presenting it in a way that will preserve it for the future may seem complicated at first, but in fact it is easy. It merely takes a bit of time and understanding to come to grips with a few details. It's just as easy to understand these and do it well as it is to do it poorly.

2. What to Save?

The simple answer is: everything.

The more someone knows about a subject, however, the harder it is to understand this. All too often, the detailed technical documentation which is really necessary to understand a subject from scratch is discounted as unimportant by those who understand the subject best. The first thing you learn when starting to preserve documentation is to look in the backs of the bottom drawers. You're looking for the things which are in fact so important to running a shop that the operators forgot they were ever worth documenting. If nobody has taken out the trash in a few decades, check the wastebaskets.

The corollary to this is that you should not be afraid to be boring. It is the nature of technical information that it is the most boring stuff in the world - until you actually need it. Then it becomes very interesting indeed. There is no way to tell beforehand.

It is also important to avoid thinking that someone else has already saved something. Perhaps this is so, but perhaps not. I've only been at this for about four years now, but already I've lost count of the number of times that I've discovered what is almost certainly the last remaining copy of a document. It is much better to save something and find that it's already digitized elsewhere than to lose the last copy. It takes little effort. A copy of every known surviving document for the Thompson Type-Caster, for example, would take between two and three inches of shelf space.

Recognize also the importance of different editions. They wouldn't have made a new edition if there weren't some information in it not present earlier. In some cases, also, the same title can indicate very different documents. For example, the *Instructions to Linotype Operators and Machinists*, although printed as a booklet, contained a collection of individual items appropriate at the time. I've identified at least four different versions.

When investigating or cleaning out an old shop, get into the habit of looking in obscure places. Look in the drawers. Look behind the drawers in cabinets. Look under things. I discovered the 1956 edition of the Thompson Type-Caster manual in the bottom of a box, underneath a stack of moldy Monotype case arrangement drawings, itself underneath a pile of rusted spare parts filling the box, at the bottom of a stack of boxes on the broken floor of a damp basement just minutes before the owner's grandchildren came through with instructions from the owner's children to throw all of that junk into the dumpster. As I write this, it's been downloaded 136 times; it was worth looking in that box.

You should be interested in preserving not just books, but anything which contains information about the technology or its history. This includes

- Books and Booklets
- Drawings (engineering or hand-scribbled)
- Notes (the operators knew a thing or two)
- Paperwork
- Packing and labels

You are trying to preserve a technical culture so that it can survive into the future. Technology is all about the details.

3. How to Preserve It?

3.1 Archival Preservation

3.2.2 Metal and Acid

Just because you are going to preserve something digitally does not mean that you should discard the paper original. As I'll discuss later, digital preservation is in its own way precarious. Nothing paper lasts forever, but we should still be attempting to preserve paper documents at least as backup sources in case the digital versions disappear.

Archival preservation is a professional subject in itself, and I am certainly no expert in it. If you really wish to accomplish the proper archival preservation of a document, you should make inquiries and contact a professional.

My own exposure to this field was simply my first slug-level job as a graduate student, working on the papers of the poet George Oppen in a university library special collections department. It was good training, because Oppen's materials were in very poor condition. The basic rule I learned there is that to preserve paper documents you must, *absolutely*, remove all metal from them. Metal will rust, and it will destroy the paper around it. This is a significant problem with older magazines and some trade literature. I routinely unbind these to remove the staples. This also makes them much easier to scan. (I must admit that I discard the staples; future historians of the staple will curse my name.)

It is best, in my opinion, to unbind such an item by hand. This can be laborious. I know of those who, instead, simply chop off the spines to get easily scannable copies. I cannot bring myself to do this.

You should also become familiar with a couple of basic archival products - especially acid-free folders and buffered acid-free tissues. For proper preservation, if a document appears to be fragile or highly acidic (where one sheet is likely to stain another, or the sheets are likely to decompose) it should have buffered acid-free tissue interleaved between the sheets. The document should then be put into an acid-free folder. This in turn is put with other folders into a box made of acid-free cardboard, and these are stored on metal (not wood) shelves in a temperature and humidity controlled environment.

I do not go quite this far in my own library, primarily because the expense is too great. However, I do use buffered acid-free tissue liberally ("buffered" means that it is prepared so as to be slightly basic; this counteracts the acidity typical of poor paper). I've also found that acid-free insert folders are much cheaper than regular acid-free folders.

My source for these is University Products (www.universityproducts.com), a major supplier of archival materials. One can be relatively confident that a product sold as "acid free" by a professional supplier such as this is in fact acid-free.

3.3.2 Order

It is also often necessary to decide what to scan first on the basis of the state of preservation of the original. Not infrequently, the originals are in such condition that they won't survive much longer at all - scanning is the only way to save their content. This must be judged on a case-by-case basis, of course, but I would note:

- Old fused-toner photocopies will stick to each other and often destroy themselves in this way. I have things I photocopied only 20 years ago that cannot be recovered.
- Old diazo/whiteprints are often in bad shape due to their chemistry.
- Old mimeographs are even worse.
- Old spirit duplicator copies are the worst; scan them immediately, lest they disappear before you get a chance.

I have spirit duplicator copies from the 1980s which can no longer be read. The book in my personal library which will outlast all the rest was printed in 1794.

3.3.3 Disposition

The final, and perhaps the most difficult, question to decide is where to preserve the original. To those with no experience in the matter, the obvious answer is to give it to a museum or library. While I hope that there are exceptions, this is almost certainly the wrong answer.

The first problem is that museums of technology have an exceedingly poor record in past decades for preserving technology (which is just what you'd think they should do). This comes from an unfortunate combination of good intentions and hard times. On the side of good intentions, it has become a part of accepted museum practice that museums must not simply stash stuff away but actively interpret the past for the benefit of their visitors. In itself this is a good thing. The problem is that it has come at a time when budgets are decreasing. You don't actually need physical objects - much less working physical objects - to "interpret" the past. All you need are pictures and interactive displays, which are much cheaper. This has led to a great deal of "deaccessioning" of objects - and if one doesn't need a typecaster or linecaster, one certainly doesn't need the technical documentation behind it.

Libraries, similarly, are subject to decreasing budgets and changing times. If you visit a university library today, you'll find that it is for the most part a convenient public place to use your computer which happens for some reason to have a bunch of books cluttering things up. It is increasingly easy to wander the stacks without getting in anyone's way because there are fewer people there. This situation is not supportable in the long term. Indeed, a couple of years ago The Internet Archive discovered that certain university libraries (they won't say which ones) were beginning to deaccession books after they'd been scanned by Google. This is troubling, since the Google scans of technical books are frequently incomplete and unreadable. From the point of view of the preservation of documentation that even we think is sometimes boring, it makes libraries a risky choice. (The Internet Archive responded by beginning a project to archive books in cold storage simply to make sure we have a backup copy of civilization.)

The problem with the alternative - keeping them yourself - is that they'll vanish in your estate sale.

I do not know the answer to this problem. Actually, I do not believe that there is presently an answer. If an answer is to emerge it would have to be in the form of institutions committed to the long-tem preservation of expensive physical objects (books, in this case, which require shelving space and climate control) when those objects have very limited popularity. This is a very hard problem.

3.2 Scanning

3.2.1 Resolution and Depth

I've encountered at least six attitudes toward scanning, each characterized by a notion of what the purpose of the scanning is.

1. Scanning for OCR. The purpose of scanning as done by Google Books, and by the "Million Books" project, is to process as much text as possible with the assumption that the important thing is to do optical character recognition to convert these into electronic text (= word) documents. This is fine so far as it goes, but it means that much is lost in the process. I don't find this to be appropriate to our purposes here. (As an aside, this is the reason that the "Million Books Project" scan of the 1923 ATF catalog, online at The Internet Archive, is so bad. It was scanned to extract its words, but it consists primarily of images.) The implication here is that we must scan at a sufficient resolution (and gray or color depth) to represent images.

2. Scanning to reprint a new edition. I know of at least one person reprinting old technical documents (Teletype manuals in this case) who scans so that what he produces at the end of his process is essentially a new edition. This means that he scans bi-level (not grayscale) and edits out anything that he feels inappropriate. The end results look very nice, but they are really new editions prepared out of the old ones, not the old editions preserved. I do not find this method to be appropriate, as it involves making editorial decisions about what is or is not part of the reprint. The implications here are the same: scan at sufficient resolution and gray/color depth.

3. Scanning just to get rid of stuff. This is what the USPTO did with its archive of patent specifications. They scanned the entire archive at 300dpi bi-level (not grayscale). This is not a sufficient resolution or depth to preserve the information of the original documents (example: one cannot discern the actual construction of Benton's matrix depth gauge from the official USPTO

digitization of its patent). What makes the USPTO instance tragic is that after they did this they deliberately sent the entire paper archive to the landfill. In this case, digitization was simply the excuse for vandalism of the public record.

4. Scanning for the Web. My friends who are graphic designers cannot understand why I scan at the resolution I do, because it is much higher than what is reasonably presentable on computer screens (which are actually still quite low-resolution). To them, scanning is a means of showing a pretty picture. This is, obviously, not what I'm trying to do.

5. Scanning for full archival preservation (see below)

6. Scanning for acceptable preservation given current constraints (see below)

It is my opinion that in scanning documents for long-term preservation (as we are here; it isn't likely that anybody is going to do this a second time), it is necessary to think in terms of preserving not only the literal information that a document is "supposed" to have, but also some of the document itself. The scan should at least give some suggestion of the paper quality, when viewed at full resolution, because only then can we be sure we're actually preserving everything.

What, then, is sufficient?

I'm sure that this is a subject which has been studied professionally, and in extreme cases it can become a major project (see, for example, the work being done to apply advanced imaging techniques to recover the text of the Archimedes palimpsest). But in simpler terms I would suggest that the ultimate limit would be that of the resolution of the human eye. This was established many decades ago in photographic research as about 100 lines per millimeter. This works out to 2540 dpi (which will be a familiar number to some).

Unfortunately, at the present moment that scanning resolution, while sometimes attainable with the optical resolution of a better scanner, is not yet feasible in terms of resulting file size.

For my own work, I have found the following compromise to be useful: Typically, I scan at 600 dpi, either in 8-bit grayscale or, more often, in 24-bit color. For particularly important images, I will make additional scans at 1200dpi. The files this produces are still too large to present online (example: my scan of Legros & Grant's *Typographical Printing Surfaces* comes to about 25 Gigabytes). I save these as lossless PNGs (see discussion below), but then process them into smaller versions for online presentation. I keep the original scans, however, and at some point when network capacity has increased sufficiently I hope to present them online as well.

Summary: Minimum 600 dpi; more ok. Scan grayscale or color - *never* scan bi-level. Save original scans as PNG images.

3.2.2 File Formats

This starts to get into details of computers, so I'll summarize it first: Save original images as PNG files, and present them online (usually) as JPEGs (or as PDF files which wrap up JPEG page scans).

Image files are large. If you store them without compression, they are huge. There are two ways an image may be compressed: lossless and lossy.

Lossless compression turns an image file into a smaller file but does not lose any information in the process. It works because most images are redundant - you can eliminate the whitespace, for example. If you uncompress such a losslessly compressed image, you get back exactly the same file you started with. There are many formats for this, but three are most common: GIF, TIFF, and PNG.

You want to use PNG. It is modern and well supported. GIF is a bit older and lacks some features of PNG at a low level. I have no idea why TIFF still exists. Years ago I had the unpleasant experience of working in it at a low level; it's a mess. But because these are all lossless formats, you can convert from one to the other with no loss (for simple scanned images without fancy features such as transparent areas). So if your scanner only saves TIFF, don't worry - it's ok.

Lossy compression is strange and wonderful. It's hard to understand, and it "shouldn't work." The best way to think of it is this: A lossy compression takes a picture and creates from it a completely different picture which (a) just happens to look to human eyes to be almost identical to the original picture, and (b) is much smaller in file storage size. JPEG compression is a lossy compression method. If you take a picture, convert it to JPEG, and then try to convert it back to the original - you can't. The original image can never be recovered, exactly, from a JPEG version of the image.

Obviously you'd never use JPEG/lossy compression for storing, say, your bank account. But for images it works quite well. The trouble is that while the "new" picture produced by JPEG transformation is very, very close to the original, it isn't quite the same. So if you do this twice in a row, it's further removed from the original. If you edit a JPEG image repeatedly, storing it as JPEG at each step, in a few steps you can see visible degradation of the image.

For the first generation of an image this is probably ok. My DSLR camera can save photographs either as (large) lossless images or as JPEGs, and I usually set it to saving as JPEG even when documenting machinery. But for scanners, when given the option, I prefer to save losslessly as PNG. If you're making a smaller, derived version of an image for presentation online, however, JPEG is usually the better choice for the final result.

3.2.3 Unsharp Mask

Your scanner may have a feature called "unsharp mask." Turn it off. What it does is blur the image in a clever way so as to make it look better to human eyes. This is great for a final result (you can apply "unsharp mask" processing in software later), but it is not what you want to do when capturing the original image. It may look better in preview mode, but you'll have lost data.

3.2.4 Alternative Process: DIY Book Scanning

The big scanning projects (e.g., Google Books) don't use office flatbed scanners. They're much too slow. They use photographic rigs. (If you download volumes from The Internet Archive which were digitized by Microsoft's now-abandoned scanning program, and look at the original page images, you can see some of the photographic setup.)

There is a small movement to establish methods and resources for do-it-yourself book scanning using similar methods. This involves two main components: (a) figuring out the camera setups, and (b) software to process the images into things that look like books, not like pictures of books. This is being done, at least in part, within the emerging "open hardware" and hackerspace/maker movements (themselves vibrant subcultures of great interest). This is an area I have not investigated in detail. In the past at least the resolution has not been sufficient for me (whereas a scanner is really an extremely high-res camera) and the manipulation of the images into e-books has transformed them to a degree where the spirit of the original document was, to me, missing. But these are merely personal thoughts, and quite likely they are outdated thoughts. The methods they are developing are very fast, and the software they're using is really quite impressive. It also works better than scanners for tight bindings. To investigate this field, you might wish to start at: http://diybookscanning.org/

4. Processing Scans

4.1 Goals

Scanning is independent of presentation. Scans, as noted above, should be as good as you can make them. The presentation or digital reprinting may satisfy different goals. As long as you still have the original scans, you can do a different presentation later.

The basic trade-off for presentation is quality vs. quantity.

It is possible to do good looking digital editions. It is even possible, I think, to do "fine" or near-fine digital editions. The problem, though, is that doing this takes a great deal of time. For a fine digital edition, every single page needs editing. It will need to be rotated (scans are rarely perfectly aligned) and cropped. Cropping will have to be uniform between scans (or you get odd page size and/or scaling effects). You may wish to do image cleanup (or not). In general, preparing such an edition of a single work is easily a matter of several days for even a short work - more for longer works. This might, of course, be exactly what you want to do. I would encourage it; the world needs more fine digital editions of arcane old typecasting books!

What I realized, however, is that my own goal was to create a relatively large body of useful work rather than a smaller body of good-looking work. There's just too much that people actually need to run these machines (I can see this by looking at the download counts of those documents I have at The Internet Archive. Who would have expected that Linotype Parts Catalog No. 52 would have been downloaded 358 times so far?) So for all but a few works I have adopted a procedure for generating useful, ugly, digital versions relatively quickly from the page scans. These are made using automated scripts under Linux. (I'd be happy to share these with anyone who wants them; they're simple and rather crude examples of Linux shell scripts).

4.2 Formats

The file formats you use are important. Why? Because they will necessarily become obsolete.

One of the sad things one learns when studying ancient texts is that all things are destroyed. We have only a handful of texts from classical antiquity - most have been lost. Actually, all of the documents themselves have been lost (aside from a few inscriptions on stone and scraps of mummy wrappings and midden material from arid regions). Typically, the earliest extant copies date from perhaps the 10th century - multigenerational copies of copies made over a thousand years after the originals. Every document will perish; only copies will survive.

Digital documents perish as well - they just perish much faster, so copying is more important. This is obviously true of the "physical" copies of electronic documents (files on disk), as anyone who has ever had a disk crash knows - or who has found an old disk backup unreadable. But it is true of formats as well. My personal collection of computer media dates back to the early 1980s, and I have many files from those years which can never be read again simply because no software now exists to read them.

It's therefore important to figure out how you're going to process scans, and what format you will save them in and present them in.

Image scanning produces image files (of course). While I believe that some systems can be set up to produce page scans grouped together into a single document, it is best (I think) to set things up so that each pages scans to a separate file. This gives you the maximum flexibility. You can then use file (page scans) individually, or group them into documents, or process them into other forms (usually smaller in file size) and then group them into documents. Image files will be in PNG (or TIFF, or sometimes JPEG) format; see above. PNG (and TIFF and GIF and JPEG) are established and documented open formats from which copies my be made indefinitely (even after they become uncommon).

When assembling page scans into complete documents, it is important to avoid proprietary formats for the overall document. In the end, all computer formats will be replaced by different formats, but if your document is in a non-proprietary, open, format then in the future someone will be able to convert it to whatever new format comes around. If its format is proprietary, then it is more likely that your document will become unreadable.

The best known, and right now probably the best, nonproprietary format for documents is PDF. It was developed by and for Adobe, of course, but they have published the specifications and third-party open-source implementations exist. There are open-source tools for doing just about anything with a PDF file (e.g., the "pdftk" PDF Toolkit program; also pdfimages and various other open-source utilities). Whatever comes after PDF, it will be possible to convert PDFs to it.

Proprietary formats such as Microsoft Word are best avoided, as documents in older versions (and sometimes sufficiently new versions) will be unreadable in third-party readers. The only way to guarantee the continued maintainability of a document kept in a proprietary format is to maintain in working order a complete chain of computers with every version of every operating system and word processing program so that you can convert "up the chain" to the latest version. (I have one friend who has done this for Macintosh computers.)

It is also best to avoid formats specific to proprietary online file sharing services (e.g., scribd) as these lock you into a format you cannot control with no options for migration.

When I began putting things online, I tried using a composite format where I did most of the text as text wrapped up in HTML (that is, as a web page) with images as individual embedded images. At the time it made sense, because even a small PDF file was a big deal back in the 1990s. Now I find that splitting the document up in this way just gets in the way. It also loses the visual integrity of the original document, and contains a whole set of unacknowledged editorial decisions about what I kept and did not. I now find it best to present a single document as a single document.

Finally, it is best to avoid device-specific formats for the obvious reason that they are specific to devices. (The Internet Archive will automatically generate derivative versions of documents you store there for Kindle, etc., but the original documents you upload are in PDF and thus portable.)

4.3 Tools

This section depends entirely on your own computing environment. I have no idea what tools are available for use on Windows and Macintosh machines. I'm sure they must exist, but every time I've seen friends try to use them they seem cumbersome.

I suppose that my basic tool is my office setup. I do much of my daily work at the computer, as so many of us do. So I've set things up to integrate scanning into my routine. It's a U-shaped setup with a standard IKEA swivel office chair inside the U. At my left is a cheap Epson office scanner ($49 on sale). In front of me is my "main" computer (with two large monitors, which is rather nice). At my right is a secondary computer and monitor at which I do all of my online work. Each scan takes some time, so while I'm working I just slap something down on the scanner, do some work, scan the next page, do some more work, etc. This routine enables me to scan small documents and individual sheets easily. It isn't fast (and anyway you'd go crazy just sitting at the scanner feeding it documents), but by the end of a few days I generally find I've managed to do quite a bit of scanning.

For my own work, I use various tools under Linux. It runs everything necessary for these tasks, and is free (free both in the sense of "free lunch" (no cost) and in the sense of personal freedom (you can modify the system as you wish)). I'll just summarize some of the tools I use. You could download and install a Linux system and use these tools to do the same thing, or use the tools of your choice.

Operating System: Kubuntu Linux (I may soon look at Arch Linux, as (K)Ubuntu is becoming bloated.

Scripting Language: bash

Scanner Software: I use the proprietary "iscan" program that you can get from Epson for its scanners, but in most cases now I think the Linux software (xsane) might work as well or better.

Manual Image Processing: The GIMP. (I find some aspects of the latest version (2.8) annoying, but in general it does a splendid job.)

Automated (script-driven) Image Processing: ImageMagick. This is an exceedingly complex program which can do just about anything to an image - cropping, resizing, etc. from the command line. I tend to figure out very simple things I need to do with it and then use scripts to automate the processing. For example, I have a short script which can take an entire directory of hundreds of images, scale them all, convert the scaled versions to JPEG, and then assemble them into a PDF.

(Aside: The ImageMagick "convert" program is useful for converting JPEG2000 images, not currently well-supported in The GIMP, to PNG images.)

PDF Assembly: The "convert" program of ImageMagick.

PDF: pdftk. This is especially useful for extracting a single page out of a PDF. (Example: when I re-present a Google Books PDF, I like to have a page image of its cover or titlepage as an icon. I use pdftk to extract that page. Then I use either The GIMP or ImageMagick "convert" to render this one-page PDF into an image. Then GIMP or ImageMagick to manipulate the image as necessary.)

PDF: pdfimages This is useful to extract all of the images out of a PDF. This can be handy when you need to use an image from a Google Books scan. (But note that the Google Books PDFs are small marvels of image compression; they use a combination of lossless (JBIG in this case) images for bi-level text and lossy (JPEG2000) images for most pictures. But some pictures in the PDFs are actually split between multiple pictures in their underlying representation. N.B., when you see a Google Books PDF with part of the page lighter in shade, that part is JPEG2000.)

Image Stitching (for large prints scanned in sections): Hugin (I'm just starting to explore this).

Image Viewing: geeqie (but any viewer will do)

PDF viewing: Adobe's acroread for Linux; there are also open-source alternatives.

There is also a script available online called "pat2pdf" which automatically downloads USPTO images and assembles them into PDFs. I've hacked this script a bit for my own use so that it works with low-number patents. It's licensed under the GPL, so I'd be happy to send you a copy of my hacked version should you wish.

4.4 Backups; Where to Store It

This section will also become obsolete quite quickly.

All computer systems fail. Really. I've been programming computers since the 1970s, and my father has since 1958. I've lost track of the number of computers and storage devices I've had. They will all fail - and a lot sooner than a Linotype or Thompson!

Every scan, and any processed versions you wish to keep, must be backed up. This has been basic computer procedure for decades (yet how often it doesn't happen!) The problem for us here is that the storage space required to save reasonable-resolution scans of old documents can be large - often many Megabytes per page. As I write this, I am spinning six Terabytes of data on my desktop, and I've used five of them.

How do you back up a Terabyte of data? In practical terms, for the home user (and indeed in professional terms for many as well) the only way to backup a Terabyte disk is to another Terabyte disk.

In larger installations, this is done with something called RAID ("Redundant Arrays of Inexpensive Disks") These automatically back up one disk to another disk in a seamless, invisible way. One still must be careful, though, as I've seen data lost when nobody noticed that the first disk failed until after the second disk also failed!

For my own home system I'm much less fancy. I simply have a USB-attached device which lets me plug in a hard drive externally (I could also have bought a system with hot-swap drives, but that would have been more expensive). To back up one disk, I put an identical disk in this external device and copy A to B. (Using USB this takes all night for a big disk, but once it is done there is a nifty little utility in Linux called "rsync" which will do this by only copying new data; it works quite well.)

Actually, I never trust one level of backup, mostly because it is easy accidentally to copy B to A when you intend to copy A to B. This would be bad. So I use three levels. This means that whenever I buy one new disk, I have to buy three. Such is the price for even relative peace of mind in the computer world. I also back up all scans to DVD+R disk, but I don't really trust these backups (the reputation of DVDs for data retention is not good).

I've also found that consumer-grade hard drives fail too frequently. I've been obliged to use "Enterprise Grade" hard drives. Western Digital calls these their "RE" drives (now at "RE4" level); I'm sure that Seagate and Toshiba have equivalent products.

All of this will soon be obsolete for most users because the world is moving rapidly to "cloud" storage. This is simply the old-fashioned mainframe datacenter reincarnated (don't say that too loudly in the computer industry). Increasingly, the concept that one might have a personal computer with its own storage, wherein you maintain your own data, will be seen as quaint and old-fashioned. Within a decade, I'd guess, nobody will have computers in the way they have from the 1980s to the early 2010s;
we'll be accessing "cloud" storage maintained on remote datacenters, and for the most part running applications hosted remotely as well. For that increasingly large segment of computer users dependent upon iPhones and iPads, this is often already the case. Dependence on cloud storage will of course present its own issues, but they're not issues to be discussed (much) here.

5. How and Where to Present It?

Given that one has scanned a document and assembled it into a file (usually now a PDF) for presentation, how and where should you present it? There are two options: on your own website, or on somebody else's website.

5.1 On Your Own Website

This is relatively straightforward (it wasn't always). Build a website. Put documents on it. I'll only make two comments:

First, don't let "build a website" put you off. It is not hard, once you realize the basic principle that your website does not have to be beautiful. Mine isn't. It's horribly old-fashioned. It's an embarrassment to my friends who are trendy graphic designers. That doesn't matter one bit, since the folks who go to it looking for, say, Thompson documentation don't care.

Second, I would plead with you not to use blogging tools to build your website. Build a real website. The problem with blogging tools is that while they work well for blogging, when you try to build an actual site with content they throw 600 years of information science out the door and reduce you to a medieval level. That's not a snipe, but a simple statement of fact. A blog is organized last-in-first-out; you read first what was last posted. The only way to get to what you want within the accumulated content of the site is either to plod linearly backwards or to consult a search engine. This is exactly the way a medieval monastic library was organized - books were cataloged as received, and if you wanted to know where a book was you searched the list linearly backwards or you consulted a librarian.

I can offer little advice as to what services you should go through for building a website. That will vary with your location and circumstances. Here in rural Wisconsin, my regular ISP (a tiny local phone company) still has relatively antiquated storage limits. So I host the text part of my site with them but use another service for the images (this is why pages on my site load up as "circuitousroot.com" but images load from "galleyrack.com"). Right now I've got something on the order of 85 Gigabytes of image data loaded up on this second service (through hostmonster.com, for about $10/month). They claim that there is no limit; we'll see. At some point I'll probably migrate entirely to hostmonster.com. Your solutions here will no doubt be different. My actual connection to the net is through a third company and a radio link to the top of the local feed mill; the joys of rural living.

5.2 On Someone Else's Website

In many ways this is easier, although there are a few pitfalls.

There is much to be said for The Internet Archive (www.archive.org). They're a San Francisco based nonprofit organization devoted to putting everything they can online. They do a very good job of it. (They have no tech support, as they have a very small staff, so sometimes there are glitches; on the whole this doesn't matter.) They will take files as PDFs (assembled of images) and automatically convert them into various useful formats (original PDF, online-viewing through a pretty good interface, Kindle, etc.) The other important thing is that from the point of view of document preservation, copyright, etc. they seem always to do the right thing. They're actively scanning documents and also actively preserving physical books (in cold storage, just in case). They're trying to coordinate with various university-based scanning programs and the US Library of Congress. They're also an institution. My own website stays online only so long as I keep paying my ISPs. One hopes that The Internet Archive might have a longer lifetime.

One thing that I do for many of my larger texts is this: upload them to The Internet Archive and also link to them at The Internet Archive from my own website. This allows me to structure the information on some subject (e.g., the Thompson) as I wish, and yet to have the underlying documents preserved by an institution.

(As an aside, I'll note a trick to uploading to The Internet Archive if your net bandwidth is low. Mine is, and it can take many hours to upload a large PDF (of, say, several hundred Megabytes). The Internet Archive "Flash"-based uploader usually fails before this, and in any case my net connection will usually drop as well. What I do is to use a Linux utility called "split" to split the file into many 50 Megabyte chunks. I upload these to my hostmonster account, where I reassemble them. If my net connection drops, I can just re-start at a particular chunk. Usually now I also present the reassembled file on my website via hostmonster. From hostmonster, I can use ftp to upload the reassembled file to archive.org.)

The other easy solution to using someone else's website would be CircuitousRoot. I'd be happy to host typecasting-related documents (provided they are copyright-clean; see that section below).

The only thing I'd suggest avoiding is proprietary file-sharing services such as scribd or Google Docs/Drive. Anything that requires that your user have a membership, anything that uses a proprietary format, or anything that allows only online reading should be avoided.

6. Copyright Issues

6.1 Who Cares?

Before delving into this I need to address the obvious question: why is this important? After all, these are all old documents from a technology of the past. Most (not all) are in fact public domain. Many of the companies involved then are long gone. It is common enough for people to think that copyright shouldn't matter here.

In one sense this is true, it shouldn't matter. But there is an emerging situation which, very unfortunately, means that it does matter. Indeed, it may matter quite a lot, and ignoring it may make a great deal of work in digital preservation useless. It's a bit like being in a war zone; it doesn't matter that it isn't your war when the bombs start falling on your city.

This particular war has been going on for some time, and it is now reaching a rather critical stage. On the one side in this war is the entertainment industry, which is trying to secure its rights to a relatively few very profitable properties. It turns out that many un-profitable things (such as documentation for old typecasters) get caught in the fallout.

There are three aspects to this:

1. Understanding the actual legal copyright status of a work.
2. The implications of this status in online reprinting
3. The implications of claiming new copyright on public domain works

6.2 Determining Copyright Status

The basic point in understanding the copyright status of a work in the US is that you cannot rely on intuition. Really. Don't. US copyright law does not make sense; it is the result of decades of inconsistent legislation, in each case guided by the needs of powerful corporations and trade groups. Don't rely on what you think it should be, and don't rely on most online gossip. That having been said, it is actually pretty straightforward to research the status of a work in the US.

(In Europe, the situation is simpler but, ironically, because European copyright law is based on the lifetime(s) of the author(s) it can in some cases be impossible to determine copyright status. Example: Until someone finds the gravestone of John Cameron Grant, it is not possible to determine the copyright status of *Typographical Printing Surfaces* in the UK.)

The basic resource for understanding and checking US copyright status is Cornell University's

"Copyright Terms and the Public Domain in the United States" at:
http://copyright.cornell.edu/resources/publicdomain.cfm

For a more extensive treatment, see also Cornell's *Copyright and Cultural Institutions: Guidelines for Digitization for U.S. Libraries, Archives, and Museums* . (Ithaca, NY: Cornell University Library, 2009). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495365

To check copyright renewal status (if any) in the US:

> The copyright renewal database at Stanford, useful for works from 1923 to 1963:
> http://collections.stanford.edu/copyrightrenewals/bin/page?forward=home
>
> The US Copyright Office database (renewals from 1978 on):
> http://www.copyright.gov/records/
>
> UPenn list of first copyright renewals for periodicals:
> http://onlinebooks.library.upenn.edu/cce/firstperiod.html

The status of overseas works in the US is trickier. Until 1996, if they had not been published in the US and did not comply with the regulations then in effect, they were in the public domain. In 1996 that changed with the implementation of the Uruguay round of the GATT talks. This retroactively provided copyright coverage to all non-US works as if they had been covered in the US for the maximum possible period (currently 95 years from date of publication); this applied to all works which were in copyright in their home countries as of 1996-01-01. This implementation was coordinated internationally with an extension of copyright in many countries. So for example on 1995-12-31 an anonymous corporate non-Crown work in England from 1945 published only in England would have been one day away from the public domain in England and in the public domain in the US. On 1996-01-01 its copyright was extended to 70 years in England (it will now expire after 2015) and, since it was once again in copyright in England, it was placed in (brand new) copyright in the US for 95 years from the date of publication (it will be in copyright in the US through the end of 2040, 25 years after it enters the public domain in England). The US does not recognize a "rule of the shorter term"; the expiration of copyright in another country has no effect on its status in the US.

The net effect of this for typecasting enthusiasts is that English Monotype, Linotype, and Intertype documentation from after 1925 (1996 - 50 > 1925, so in the public domain in England on 1996-01-01) is in copyright in the US even if it is in the public domain in England.

English copyright is based on the model, more common internationally, of providing a term of a certain number of years after the death of the last surviving author, or a specified number of years from publication date for anonymous (e.g., many corporate) works. UK Crown Copyright is a special case. The copyright status of UK patent specifications is a particularly complex mix of copyright law, copyright law for Crown copyright, and certain legal dispensations of the UK Intellectual Property Office for fair use. (By way of contrast, US patent specifications are in the public domain by law.)

UK copyright law is summarized in the "Fact Sheet P-01: UK Copyright Law" of the UK Copyright Service, a privage organization:
http://www.copyrightservice.co.uk/copyright/p01_uk_copyright_law

Canadian copyright law is summarized in a useful flowchart prepared by Creative Commons Canada: http://www.creativecommons.ca/blog/wp-content/uploads/2008/04/pdregistryca-pd_flowchart.pdf

6.3 Online Reprinting

Why does this matter for online reprints, when in practical terms nobody cares about these old documents? It matters because over the past couple of years there have been two major attempts by the entertainment industry to introduce legislation that would have made Internet Service Providers responsible for the copyright status of material put online by their users. While these two bills failed (due to strong opposition from ISPs, Google, etc.) it is not unlikely that similar legislation might succeed in the future. Indeed, it is likely that it will.

This would give the entertainment industry the ability to shut down the business of an ISP for copyright violations of its users. If this happens, this will make your ISP care very much about what you put on your website, as it would be a matter of business survival for them. You'll have to be able to prove that whatever you put up is legal, or they'll have to shut down your site to save their business.

Now, if you actually look at it, the vast majority of all typecasting and linecasting literature published in the US in the period in which we're interested is in fact in the public domain. (I've done a lot of checking of a lot of individual cases, obviously.) Moreover, status is quite easy to check. It simply makes sense, given the risks, to be proactive about this and to check the copyright status of anything you plan to reprint (and to record that status with the reprint).

6.4 The Problem of New Copyright

The other area where copyright is important is in the issue of newly claimed copyright on a work.

You just scanned a work and processed the files into a PDF. Do you claim a new copyright on it? Should you?

In theory, if there is no substantive new work, then you cannot claim new copyright. Hard work doesn't merit copyright; only new work. Whether the digital processing you did constitutes enough new content to enable a new copyright claim is a matter between you and your lawyer.

I would like to argue, however, that in any case you should not assert a new copyright on a digital reprint of a public domain work, because that new copyright will almost certainly cause the work (and all your work in scanning it) to be lost.

The problem is in the extreme term of new copyright. At present it is 95 years, a great extension in term from its earlier days. This extension of term has been driven by the entertainment industry trying to keep profitable works in copyright; it may well be extended again.

In any event, even at a term of 95 years, it is safe to assume that the original document you scanned will be lost by the expiration of this term (either to the decay of its paper, or to a landfill after an estate sale, or the same after a library deaccessioning). Only your scan will survive.

But do you really plan to consistently market this scan, and to make it available, for 95 years? For a fraction of that period? The digital copies will be lost faster than the paper ones, and they'll need to be copied. But if your work is in copyright, then nobody can copy it (this will be the case more and more if ISPs have to start monitoring content, and as all data storage moves to "the cloud.") If nobody is copying it, and you're not providing it 95 years from now, it will be destroyed.

So I would argue quite strongly that if you go to the trouble to digitize a work in the public domain, and if you wish it to survive, keep it in the public domain explicitly. Otherwise you might as well just destroy it now.

As an afterword to this discussion, I would also suggest that for new works you write you give serious consideration to the various licensing terms put together by the Creative Commons organization. They have thought through quite well the issues involved in granting limited, reasonable copying terms for works which you wish to share but in which you wish to retain certain rights. Most of my CircuitousRoot website, for example, is licensed under the "Creative Commons Attribution-ShareAlike" license (as is this present paper). See: http://www.CreativeCommons.org/

7. Online Sources and Tools (Existing Archives)

This section isn't really about the new preservation of documents, but rather about digital resources for research. I thought it might be handy.

- The Internet Archive (www.archive.org) is perhaps the best single resource, in the sense that if they have a digital text, it is usually the best available. (Exceptions: their copies of Google Books scans, which are just the Google scans, and their scan of the 1923 ATF catalog).
- Google Books is very useful, but equally limited. In particular, if a book has fold-out plates, they never actually fold them out when scanning them. Finding complete print runs of journals is agonizing. It is best accessed through their "Advanced" search page at: http://books.google.com/advanced_book_search
- The Hathi Trust. This is a consortium of some (not all) of the libraries contracting with Google Books. They are presenting their copies of the scans Google has done of their books. The advantages are that their scans may be slightly higher resolution and that they are more aggressive than Google about establishing the public domain status of works. The big disadvantage of The Hathi Trust is that their books are viewable only one page at a time. There was a script (hathihelper.py) to download entire books, but it would appear that in late 2012 The Hathi Trust will, probably under pressure from Google, rework their security so that this script will no longer work. This is a shame.
- The Library of Congress Prints and Photographs Department's website of digitized material is a wonder. Don't go there unless you plan to spend a couple of weeks browsing. They also do the right thing and do not claim new copyright on the works they digitize. http://www.loc.gov/pictures
- Espacenet is the portal for searching for European patents. The Europeans have been much better about digitizing them and categorizing them (the USPTO is terrible for older works), but they do not have a free-form search engine. http://worldwide.espacenet.com/advancedSearch?locale=en_EP
- American patent records were destroyed in a fire in 1836. Many have been reconstructed from before then, and are termed "X" patents (because they fall outside of the new numbering

scheme established after the fire). There are two useful databases which have some information on these: http://www.datamp.org/ and http://www.nonesuchtools.com/patent/shotgunx.htm

- The USPTO website is basically useless for the period of interest. Use www.Google.com/advanced_patent_search to attempt full-text searches of patents (not reliable) and pat2pdf to download them from the USPTO.
- Researching British patents from the 19th century can be tricky; I've collected some useful information http://www.CircuitousRoot.com/artifice/restools/patents/index.html
- The New York Public Library was once a useful resource (Huss used it for his research into Church's typecaster, for example). This is no longer the case. They require license charges for the reprinting of anything you pay them to research (they can do this even for public domain material; it's contract law, not copyright law). They are also incapable, in my experience, of locating volumes in their own stacks, and once you direct them they are equally incapable of doing good reproductions from them (even when paid). This is all very sad.
- By way of contrast, the UK Intellectual Property Office has acquitted itself quite well; when I pointed out to them that their reproductions of Church's drawings were unreadable, they re-copied a very nice, full-size set.
- Finally, I will note that one of the most useful things you can possibly do when becoming involved in this sort of research is to build a good working relationship with the reference librarian at your local public library. They love doing this kind of research! It is amazing what they can actually find for you via Interlibary Loan.

8. The Future

So what of the future? I've got stacks of stuff I plan to scan yet, and I'm having a great time doing it. I would encourage you to do the same. I'd also encourage you simply to preserve what you find. Even if you haven't the inclination to scan it yourself, save it. Someday someone else may scan it.

But all of this isn't enough. Saving the documentation is necessary, but by itself it is insufficient. We really need to save the knowledge.

I hope not to sound morbid, but we are now at a point where we have the final opportunity to tap the minds of the last generation of people who came to typecasting and linecasting as an ordinary trade, intending it to be their life's work. They'll be gone soon. So the preservation of oral history, as the C. C. Stern Foundry plans to do, is as important as the preservation of documents.

I had the good fortune to obtain a nearly-complete run of the ATF *Newsletter*. I suppose that I'm therefore one of the few who has had the chance to sit down and read it all at once, as if it constituted a single book (those of you who have complete runs have probably read it as it came out). It is a remarkable experience, and one that I would recommend. What emerges is a complete picture of a group of people who have, in fact, captured the technical culture of what is now a former era and preserved it as a living culture for the future. I only started at this four years ago, and I'm grateful to all of you for what you've done. There's a lot more to do.

Appendix1: Some of what I've Done So Far

I've been scanning old technical material for the web since the 1990s, but the current incarnation of my site really dates to around 2004 when I took the time to write my own "content management system." It's a quirky little system which I wouldn't recommend to anyone else, but it suits me. Since then I've put something over 750 pages online, at something over 80 Gigabytes. (I've also put 251 documents on The Internet Archive.) Most of this content has come since 2008, when I acquired my first Linotype and realized what I should have been doing all my life. Trying to do an overview of the whole undertaking here won't work, so I'll merely sketch some notes to address a few things which may have puzzled visitors to the site.

"CircuitousRoot" is a name my wife, Rollande Krandall, first invented for a (failed) celtic band in the 1990s. I took it over first for the name of my blacksmithing forge and now for all my hobby activities. (Insofar as I start to do typecasting professionally, I'll do that separately under the "LemurType" name so as to keep my hobby activities free and noncommercial.) I find that "CircuitousRoot" combines nicely the concepts of "taking the circuitous route" with finding the deeper roots of things (never a straight path). It also has the distinction that nobody can spell it. In a fit of lunacy I actually registered it as a trademark. During that process, I discovered not only that nobody else had registered anything close, but that nobody had tried to register a trademark, ever, with the word "circuitous" in it. It seems that I rather specialize in the obscure.

The general organization of the site is as a hierarchy of topics, going deep in preference to going wide. The consensus is that most people using the web will only click one-deep on anything. I tried, therefore, to put things at least two-deep. There will be a logic to the location at which you find something on the site - it just might not be the logic you had in mind.

There are two exceptions to this pattern of depth. I regret them both, but the structure is now too big to change. When I built my content-management software, I allowed both straight hierarchy and the possibility of multiple items at the same level. I've used the "multiple items" feature twice: in the main "Artifice" page and in the main page for the Typefoundry & Press. You may never notice this if you only click down on links, but if you click up you'll hit confusion in these two locations. Sorry.

The general organization of topics is by machine. This means that to me a Monotype specimen book which has only rental faces for display casting goes under the "noncomposing typecaster" sub-hierarchy while a specimen book which includes both composition and display goes under "composing typecasters." If anything on the site seems confusing, remember that I'm a mechanic and casterman, not a printer.

Teletypesetters, Monotypes, and Gorton pantograph engravers are particularly hard to categorize. I try to do a lot of crosslinking.

There are two main sections:

- The Typefoundry & Press: www.CircuitousRoot.com/artifice/letters/press/index.html
- The Machine Shop: www.CircuitousRoot.com/artifice/machine-shop/index.html

Slowly I'm adding a few other items related to non-typographic letter-making at: www.CircuitousRoot.com/artifice/letters/index.html

You can ignore everything else.

Within the "Typefoundry & Press" section, my original idea was to have three main sections: Surveys of the field in general terms, Mechanics of the machinery in detail, and the actual Use of the machines. I still think that this is a good idea, but in fact the middle section has ballooned all out of proportion to the other two.

The composing linecaster section was offline for over a year as I rewrote things as I was moving imagery to hostmonster.com Much of it is back online now, but not all. The Ludlow section, though, is very nearly complete for mechanical items (but not yet for type specimens).

The list of currently operating typefoundries is at:
www.CircuitousRoot.com/artifice/letters/press/tools/type/typefoundries/index.html

The list of all type specimens scattered through the site is at:
www.CircuitousRoot.com/artifice/letters/press/noncomptype/all-typography/index.html

More specifically, what I intend as a list of all known typefoundries through history (with notes and, when available, specimens) is at:
www.CircuitousRoot.com/artifice/letters/press/noncomptype/typography/index.html

and a discussion of why I consider mere Monotype shops "type foundries" is at:
www.CircuitousRoot.com/artifice/letters/press/noncomptype/typography/why/index.html

Finally, here are pictures of some of the ironmongery with which I amuse myself when I escape from my desk:
www.CircuitousRoot.com/artifice/letters/press/cr-stuff/quicklook/index.html

At some point after the ATF 2012 Conference, I'll put this present paper online at:
www.CircuitousRoot.com/artifice/letters/press/typemaking/atf/index.html

Appendix 2: Reference Summary

- Save every scrap of documentation, no matter how insignificant or boring.
- Look underneath things, etc.
- Do at least minimal archival safe-keeping work for fragile items, if needed.
- Keep the originals even after scanning.
- Unresolved: how to preserve originals in the long term?
- Try to integrate scanning into your desk work.
- Scan minimum 600dpi color or grayscale, save as PNG, no unsharp mask.
- Back up your data!   Back up Terabyte disks to other Terabyte disks.
- Decide what goals you have for presentation/quality.
- Process to JPEG and wrap as PDFs for presentation.
- Establish copyright status for each work scanned.
- Do not assert new copyright on public domain works.
- Consider CreativeCommons licenses for your own new work.
- Avoid proprietary data formats and file sharing services.
- Publish these documents on your own website, or The Internet Archive (archive.org).
- If you make a website, don't use blogging tools.

Written using LibreOffice Writer under Kubuntu GNU/Linux.
Typeface: Linux Libertine.
Revision 1, 2012-08-13.