Digital Preservation: A Quick-and-Dirty Do-It-Yourself Guide

Marc Demarest
The Emma Hardinge Britten Archive
October 2013
marc @ ehbritten.org

Abstract

Digital preservation is no longer the province of specialist firms and library science. Individuals in possession of rare, scarce or degrading books, periodicals and ephemeral can, with relatively inexpensive tools and a little care, produce high-grade digital versions of these materials: versions that are easily stored, circulated and shared, and that ensure that the rare, scarce and degrading material will not be lost to posterity. This article discusses the tools, techniques and processes used by people working with the Emma Hardinge Britten Archive, and the International Association for the Preservation of Spiritualist and Occult Periodicals, to preserve paper-based materials.

Overview

Ordinary people are often in possession of extraordinary printed material: rare books, scarce pamphlets, photographs, fliers, broadsheets, newspapers and other ephemera that are of interest to others, or that represent important-but-unavailable parts of textual corpora studied by people, all over the world.

In many (if not most) cases, that material is degrading, not because of how it is cared for by its owner, but because of how it was made, originally. High acidity paper, poor-quality ink, iron-staple binding -- all of these materials degrade, and in degrading compromise the quality, and the life, of the printed material. This degradation is irreversible, and cannot be arrested usually, except with high-end preservation techniques not available to ordinary people.

Digital preservation -- the conversion of physical printed artifacts to digital representations -- not only captures the material in its current state, but allows multiple copies -- potentially, thousands of copies -- to be in use, across the planet, simultaneously, without further damaging the original artifact.

The tools required to do a creditable job of preserving printed materials in digital form are, these days, available to nearly everyone, at reasonable prices.

This quick-and-dirty do-it-yourself guide describes the tools and processes used by preservationists at the Emma Hardinge Britten Archive, and the International Association for the Preservation of Spiritualist and Occult Periodicals to produce digital representations of Spiritualist and occult printed material, providing a guide for owners of printed materials who are interested in preserving those materials digitally.

A note: library scientists and professional preservationists often take a dim view of amateurs. While we will note some of the issues of interest to professional preservationists in this document, this document is not intended to be a substitute for a thorough grounding in bibliographical and preservation science, to provide the framework for a preservation policy, or to serve as a guide for people working to preserve manuscript material or incunabula. Googling “digital preservation” (with the phrase quoted) will provide the curious reader with access to the arcane world of professional digital preservation.

What Are You Preserving? And Why?

To know best what to do with the scarce material you have, it's important to answer three questions:

- What are you preserving?
- Why are you preserving it?
- What will the digital representations you produce be used for?

In cases where you have an absolutely unique artifact -- say, Albert Einstein's copy of *The Secret Doctrine*[1] with his marginal notations on hundreds of pages, and with inserts of different kinds placed by Einstein at various locations within the text, including what appears to be a proof for Von Westerlaak's second hypothesis -- it is clearly important to preserve not only the text of *The Secret Doctrine*, but also Einstein's marginalia, and the slips and bookmarks he added to the text. This is a job for professionals, and you should seek professional assistance.

This is not the normal case. In the normal case, you will have a rare, scarce or important text, in some form, that may be degrading, or may be of interest or use to others, and you want to reproduce it digitally, both to preserve it, and to circulate it, in digital form, without risking damage to the physical artifact.

If you see yourself as preserving **the physical artifact in its entirety**, then you're going to want to provide, along with a "scan" of the text, a rather large amount of metadata (data about the artifact in question), including:

- scans of the cover, spine and all end-papers (even blank end-papers), if those exist
- physical measurements of the artifact (height, width and depth)[2]
- a comprehensive bibliographical description of the artifact[3]

---

[1] An entirely imaginary example.

[2] Photographs of the front, spine and back of the item, with height and width rules, incorporated into the digital version of the artifact, usually suffice. For a crude example of physical measurement, see: http://www.noumenal.com/marc/jcfmf/index2.html

[3] Bibliographical description is not simple, but it isn't rocket science either. If you intend to do this, obtain a copy of Fredson Bower's Principles of Bibliographical Description (http://www.amazon.com/Principles-Bibliographical-Description-Pauls-Bibliographies/dp/1884718000). For a simple example of bibliographical description, see: http://www.noumenal.com/marc/jcfmf/index2.html.

- information about the provenance of the artifact, in cases where the bibliographical description of the artifact does not answer questions of origin, ownership or transmission.[4]

If you see yourself **as preserving the text of the artifact** in its entirety, your job is a bit simpler. You can certainly choose to provide information about the physical artifact as part of your digital representation, but it's not necessary if you see your job as producing a useful digital representation of the text of the artifact, as long as you provide a complete representation of the text, including all information that would be required for a user to construct a complete and accurate bibliographical citation of material drawn from your digital representation.

You also need to have formed, in your own mind, a clear sense of why you are preserving the article in question. Are you producing a digital representation of the artifact, for your own use, because your original is too delicate to be used? Do you intend to circulate the digital representation of the artifact, for use by others?

If you are creating a private copy, for your own use, you're of course free to do as little or as much as you think you need.

But, to the extent you intend to let your digital representation loose in the wilds of the Internet, you have a responsibility to make sure that your digital representation is complete (that it accurately represents all relevant, information-bearing aspects of your original), because others will not have the luxury that you have: of referring back to the physical artifact, when information-bearing aspects of the original are missing from the digital representation.

You also owe it to yourself and your users to specify, very clearly, within the digital representation itself, what rights your users have, with respect to the material. Particularly, if you do not want users to republish your artifact, for free or for fee, you need to specify the limits of users' rights in the digital artifact itself.[5]

In any case, you need to form a sense of your user, her needs and what you need to put into your digital representation to meet your users' needs.

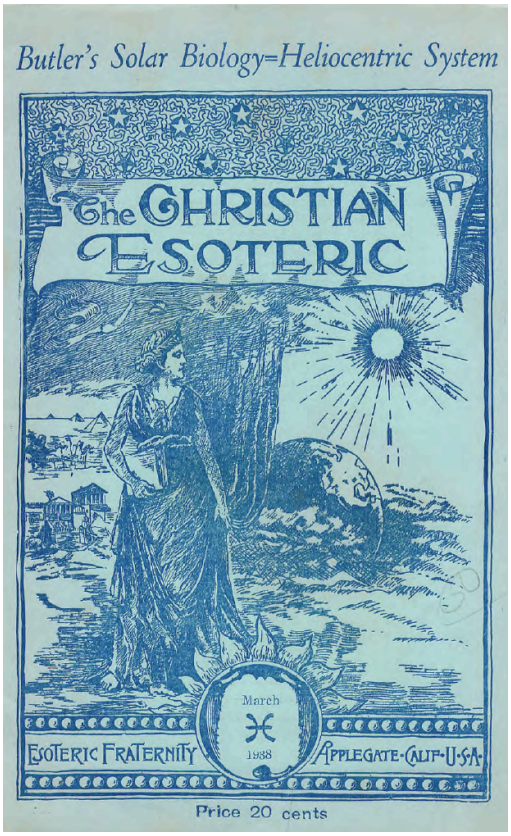
At the Britten Archive and IAPSOP, we have a simple model for preserving Spiritualist and occult materials, and our tenets can be summarized thus:

---

[4] For a fun discussion of provenance, see: http://www.prbm.com/interest/provenance-a-b.php. For serious resources, see: http://libguides.princeton.edu/content.php?pid=256933&sid=2147187

[5] Creative Commons licenses are a great way to do this. See http://creativecommons.org.

1. *Preserve the text*.[6] In most cases, the material we work with is valuable primarily for its content, and not its physical attributes. In many cases, we are working with periodical material that has no significant physical attributes beyond the text, and when that material does have significant physical attributes beyond the text -- the covers of pamphlets and magazineas for example -- we include those in the digital representation of the physical object. We consider tipped-in advertisements and publishers' catalogs to be a part of the text, and always include those. We preserve partial (damaged, incomplete) texts when we cannot locate complete texts.



Example: The cover of an issue of The Christian Esoteric, in the IAPSOP archive. These covers are information-laden and are preserved.

2. *Preserve the other aspects of the physical artifact when they have significant information-bearing or aesthetic value*. When we are working with bound books, we frequently do not preserve the binding of the text, or the blank end-papers of the text, and we do not provide bibliographical descriptions of the physical original, unless there is something about those components that is information-bearing. Other preservationists make different choices in this regard.

---

[6] Preserving "the text" is itself problematic, in many cases. Understanding the various editions and states of the text you are preserving is always important. For more information, see: http://www.noumenal.com/marc/etext.html

3. *Any digital copy, no matter how incomplete or damaged, is better than no digital copy at all*. Because the materials we work with are in most cases scarce and in danger of vanishing, we preserve whatever we get our hands on, even if the material is damaged and/or incomplete.

4. *Damaging the original in the creation of the digital representation is often unavoidable, and in some cases necessary to produce the highest-possible-quality digital representation. Don't damage physical artifacts you don't own. Disbind physical artifacts you do own, to produce the optimal digital representation of the artifact, and preserve the disbound original*. We routinely unbind books we have purchased, and trim the unbound pages prior to scanning those pages in a sheet-feed scanner in order to get a clean, crisp and shadow-free image of every page of the book: page images that offer the best chance of a good optical character recognition (OCR) run. Many people find this practice abhorrent (and immoral), as the original physical artifact is irreparably damaged by this process. There are alternative non-destructive methods to preserve bound books, and you may choose to pursue those instead. We discuss both, later in this document.[7] Below are examples of the relative merits of destructive and non-destructive scanning methods.

[7] Disbinding is a charged issue. Feel free not to disbind books, if that offends you. Our belief is that preservation of the text trumps any quasi-religious notions about the privileged status of a book.

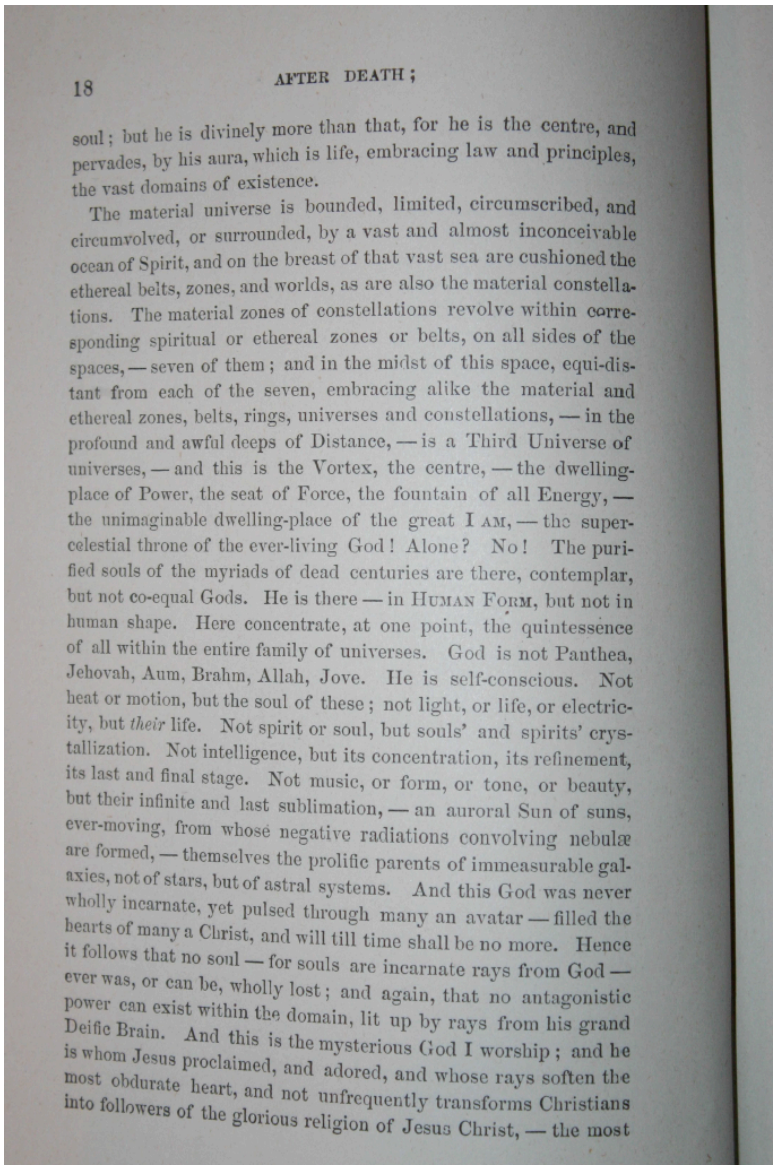
he is differenced constitutionally from the lowest savage: a race thus constituted might be established on Earth. The adept perceives the law; but he also perceives, that if the methods of the law were to fall to the handling of those whose characters are evolved, so as to form, within an educated intellect an evil or inversive spirituality, results would ensue that would be pregnant with unnamable disasters.

6. Already the outward science almost touches the confines of this secret domain. Hence, if the adept puts forth at any time the concentrated forces at his command, he exerts them silently, invisibly, to combine the chemical elements, by their finer qualities, in such style that they shall blind the eyes and baffle the methods of the unauthorised investigator; interposing barriers between him and the goal of his pursuit. Placed thus as a guardian, the servant of the Law must be prepared, in emergencies, to chemicalise the arch-fluids from his person, and overcome by means of them the intruder who would sacrilegiously grasp the potent elixirs that are concealed within the laboratory of nature. Thus he must not only conceal his own knowledges of the divine method, but must also seek, by every endeavor, to protect those arcana from unwarranted approach and unlawful appropriation: and this is no easy task at the present time.

7. There are laws and processes of law, which if known would enable a combination of scientific experts to divert the rains from any region of the globe for a period; and again to flood that region to the entire destruction of animal and human life.

8. The fluid known as electricity is a compound ether, itself the reservoir and medium of more subtle and potent agencies. These are capable of many combinations, in hands that are educated to the unbinding, recombining and directing of them;—capable, in one quality and effect of force, of being discharged upon a land as a deadly pestilence sufficient to destroy life down to the very germs within the soil;—capable, in another form of force, of inducing a common madness in the minds of a people;—capable of making the bed sterile;—of dissipating and rendering void the fecundative elements of

Example: A page from an instance of T. L. Harris' The Wisdom of the Adepts, captured destructively by disbinding and scanning each page in a sheet-feed scanner.

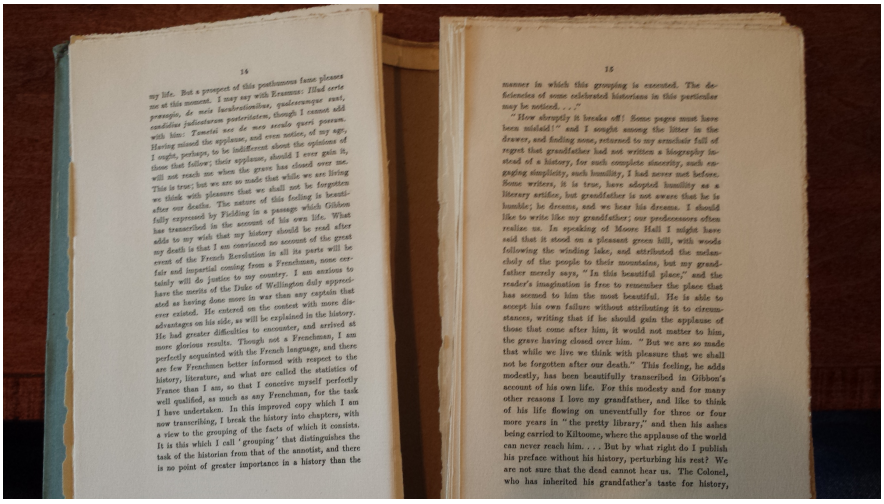


18 AFTER DEATH;

soul; but he is divinely more than that, for he is the centre, and pervades, by his aura, which is life, embracing law and principles, the vast domains of existence.

The material universe is bounded, limited, circumscribed, and circumvolved, or surrounded, by a vast and almost inconceivable ocean of Spirit, and on the breast of that vast sea are cushioned the ethereal belts, zones, and worlds, as are also the material constellations. The material zones of constellations revolve within corresponding spiritual or ethereal zones or belts, on all sides of the spaces, — seven of them; and in the midst of this space, equi-distant from each of the seven, embracing alike the material and ethereal zones, belts, rings, universes and constellations, — in the profound and awful deeps of Distance, — is a Third Universe of universes, — and this is the Vortex, the centre, — the dwelling-place of Power, the seat of Force, the fountain of all Energy, — the unimaginable dwelling-place of the great I AM, — the super-celestial throne of the ever-living God! Alone? No! The purified souls of the myriads of dead centuries are there, contemplar, but not co-equal Gods. He is there — in HUMAN FORM, but not in human shape. Here concentrate, at one point, the quintessence of all within the entire family of universes. God is not Panthea, Jehovah, Aum, Brahm, Allah, Jove. He is self-conscious. Not heat or motion, but the soul of these; not light, or life, or electricity, but *their* life. Not spirit or soul, but souls' and spirits' crystallization. Not intelligence, but its concentration, its refinement, its last and final stage. Not music, or form, or tone, or beauty, but their infinite and last sublimation, — an auroral Sun of suns, ever-moving, from whose negative radiations convolving nebulæ are formed, — themselves the prolific parents of immeasurable galaxies, not of stars, but of astral systems. And this God was never wholly incarnate, yet pulsed through many an avatar — filled the hearts of many a Christ, and will till time shall be no more. Hence it follows that no soul — for souls are incarnate rays from God — ever was, or can be, wholly lost; and again, that no antagonistic power can exist within the domain, lit up by rays from his grand Deific Brain. And this is the mysterious God I worship; and he is whom Jesus proclaimed, and adored, and whose rays soften the most obdurate heart, and not unfrequently transforms Christians into followers of the glorious religion of Jesus Christ, — the most

Example: A page from an instance of P. B. Randolph's *After Death*, captured non-destructively with a document camera rig. Note the shadow, text distortion and gutter.



Example: An instance of George Moore's Hail and Farewell, disbound and trimmed for scanning as part of the author's project to put the entirety of the Carra Edition of George Moore's work online.

5. *Preserve the text in photofacsimile form*. We do not believe that converting a physical artifact to machine-readable text (or HTML, or XML, or Word) is an act of preservation; it is an act of translation. Instead, we choose to reproduce artifacts digitally as **Adobe Acrobat PDF files**, which can be converted into machine-readable text (or into sets of pages images) at will, by the users of our materials.[8]

6. *Preserve the text in indexed and searchable form whenever possible*. We assume that our users will be as interested in finding things within the preserved text as they are in reading the preserved text, end to end. We also assume copies of our digital representations will be stored on the Internet, and should be discoverable by search engines, which will in most cases use indexes and embedded textual representations found in PDF files.

7. *Release the text into the wild, to ensure the greatest possible chance of its survival*. As many people have pointed out, the Internet is a copying machine, and promiscuous copying of our digital representations increases the chances of the survival of those materials. We encourage people to make copies, to repost (subject to legal restrictions) and in general to spread our digital representations far and wide.

[8] Arguments are made that HTMLifying or text-ifying a photofacsimile document has two benefits: (a) when posted on the Internet, search engines do a better job of indexing the document's content, and (b) people can read the document online, in a browser, without waiting for long PDF download times on slow connections. Both arguments are sound. We release PDFs that can be converted to HTML or text, and posted, by anyone inclined to spend the time to do so. We believe that photofacsimile renditions of physical printed material are superior preservations, because they capture the look of the printed original, and because they can be converted easily to other forms. Neither of those features is the case with HTML or text.

8. Release the text with the minimum legal constraint, to promote circulation and use. We release our materials with a Creative Commons license that allows broad non-commercial use of our material. The only activities we forbid are those that seek to make money by republishing our materials for a fee.[9]

The point is: think about what you're doing, and develop a policy and a set of guidelines to constrain your work. If you're making choices that won't be obvious to your users, document those choices, in the digital representation itself -- using a sheet or two for commentary, at the end of the document.

[9] In our estimation, the reprint publishers will seek to make money by publishing in physical form any digital representation of any text they can obtain, gouging consumers who could obtain those texts at no charge in electronic form. If you do not explicitly exclude this activity, you should expect it to occur.

The Tools of Our Trade: What We Use To Preserve Texts

For IAPSOP's work, the hard and soft toolset includes:

- sheet-feed scanners
- flatbed scanners
- document cameras
- Adobe Acrobat Pro
- Image processing software
- miscellaneous utilities.

Sheet Feed Scanners

Sheet feed scanners, which cost between $500 and $5000 USD, are very useful for high-quality, flat and accurate reproductions of loose sheet material, including the disbound trimmed pages of bound books and pamphlets.

Example: A Fujitsu ScanScap S510M sheet feed scanner, used by one of the IAPSOP principals. This device is the workhorse of our digitization efforts. This particular unit cost $495 in 2008, and has scanned more than 200,000 pages of printed material for IAPSOP, without a single service event.

Flat Bed Scanners

Flat bed scanners, with prices starting at under $100 USD, are useful for scanning books, pamphets, photographs and newspapers. For books and pamphlets, flat bed scanners offer a non-destructive option to produce high-quality page images with a minimum of distortion and shadowing, if they are used properly.

For newspapers and other large-format artifacts, A3 flatbed scanners (with beds measuring 11 inches by 17 inches, and prices starting at around $250 USD) offer a mechanism for reproducing large-format items, with a minimum amount of image reconstruction. Broadsheet items can be reproduced on A3 flatbed scanners by doing a half-sheet per scan, and stitching the resulting half-images back together, using a stitching utility.

Example: A Canon Lide 100 9x12 inch flatbed scanner, used by one of the IAPSOP principals. This particular device cost $99, is highly portable, draws its operating power from a USB port (no power cord) and can produce ultra high-resolution (2400x2400 dots-per-inch) images. It can be carried in a computer bag, and used in libraries that permit the use of such technology, without much effort. This particular device has produced more than 250,000 page images for IAPSOP and its Standard Spiritualist and Occult Corpus project.

Document Cameras

Document cameras, which can be acquired for as little as $100, or built using standard digital SLR cameras[10] and tripods, offer a non-destructive method for reproducing pages of bound material.

Document camera attachments and rigs can be purchased, or constructed, to provide direct non-angled lighting for the page, to reduce shadowing and gutter noise.

Example: the Ipevo Ziggi document camera, used by several IAPSOP preservationists to preserve fragile or tightly bound material non-destructively.

[10] If you have a steady hand, and an assistant, you can produce usable images of a printed artifact with a high-end cell phone camera, provided it is 5 megapixels or higher in resolution, and you have the storage space for the resulting images on that phone, or a cloud service to which the phone is connected. This strategy works well when you have to preserve artifacts in situ -- in the library where they are held -- and you cannot bring other equipment into that setting. High-end cell phones, like the Nokia 11xx series, have 40+ megapixel cameras and their image quality and detail exceeds that of digital SLRs.

Example: A professional document camera rig, with a 13 megapixel Canon EOS digital SLR serving as the page image capture device. Similar, less expensive rigs, with light sources, are usually available from digital SLR camera manufacturers.

Note: We do not recommend using the flatbed scanners built into so-called “all-in-one” devices (printer/fax/scanner combo devices). These scanners tend to be configured, in software, to produce low-resolution images in undesirable file formats, with lots of artifacts (dead and distorted pixels) in the image. Choose better tools.

Adobe Acrobat Pro

The most expensive part of IAPSOP’s toolkit -- clocking in at roughly $500 a copy -- Adobe Acrobat Pro includes all the features required to take raw TIFF images (the file format in which IAPSOP scans material) and convert those images to one-page, fully-indexed PDF files that are highly portable and guaranteed to be usable, at no cost, by our users.

The format in which you preserve materials matters, significantly. You have to imagine users, 50 or 100 years from now, with radically different technology than we have today, making use of the digital representations you create today. We use Adobe Acrobat, despite its high cost and Adobe’s lousy reputation for support and customer care, for a few simple reasons:

1. Acrobat is a ubiquitous file format, with free readers available on every modern computing platform, in most if not all national languages.

2. The Acrobat file format, Portable Document Format (PDF), is likely to be either directly usable, or convertible, two or three decades from now.

3. PDF can be converted, by Acrobat and third party tools, to a wide variety of file formats, including HTML, text, and several different image formats, include the Tagged Image File Format (TIFF), which is a 30-year-old file format that is almost certainly the most widely-used, widely-supported file format in the world.

There are rough substitutes available for Adobe Acrobat Pro available in the market; we have used most of them, and found all of them to be, in some form or fashion, suboptimal as compared to Acrobat Pro.

Image Processing Software

In our workflow, it is sometimes necessary to post-process page images (TIFF files, in our case) before those page images are combined in Adobe Acrobat Pro. To process TIFF images, we use GIMP, an open-source image processing tool that is available, for free, for most computing platforms.[11] GIMP has most of the functionality of Photoshop, and none of the cost, or the hassle of dealing with Adobe.

Miscellaneous Utilities

We use a wide variety of special-purpose utilities to solve particular problems in our workflow. For example, to index images that Acrobat refuses to recognize as text, we often use PDF OCR X, an OS X implementation of the Tesseract OCR engine.[12] To produce ultra-high-fidelity textual representations of PDF objects, we use Omnipage Professional, the best commercial OCR engine available today (and one of the most expensive).[13] In general, you will find there is an open-source or a commercial utility for any special-case problem you encounter in your digitization process.

---

[11] http://www.gimp.org/

[12] http://solutions.weblite.ca/pdfocrx/

[13] http://www.nuance.com. ABBYY is also a fine OCR engine, and costs less than Nuance products. Because Acrobat's in-built OCR engine is terrible, use high-resolution images to enable your users to (a) dump the images from the PDF file and/or (b) use a third-party OCR engine on the images.

Note: the workstation on which you do preservation work should be relatively powerful: a multi-core CPU, several gigabytes of main memory (RAM), and several hundred gigabytes of free disk storage, as well as a current version of your operating system. Image processing is CPU-intensive (lots of calculations), memory-intensive (whole images are read from disk into main memory) and disk-intensive (lots of reading and writing of bits from and to the disk).

IAPSOP preservationists use both OS X and Windows workstations; the files created on one can be read and used on the other.

Our Workflow: How We Digitize A Physical Artifact

Your workflow -- the process you use to convert a physical object to a digital representation of that object -- is more important than the tools you use to do it. The workflow is, ultimately, the guarantor of the fidelity of the digital representation, and of its quality as well.

Our workflow varies depending on the project, but is, in summary, roughly as follows:

1. Select a digitization strategy.

2. Convert the physical object to a set of 600 DPI images, one page per image, or two pages per image, depending on the scanner used and our ability to part out the original.

3. Inspect each image for quality and pagination.

4. If necessary, duplicate each image (when the images are two-page images and we are producing a single-page per image digital representation).

5. Combine the TIFF images in Acrobat.

6. Rotate the pages in Acrobat as necessary.

7. Crop the pages in Acrobat as necessary.

8. Carefully check pagination, flow and image quality in Acrobat.

9. OCR in Acrobat.

10. Dump OCR'd text in Acrobat to assess OCR quality.

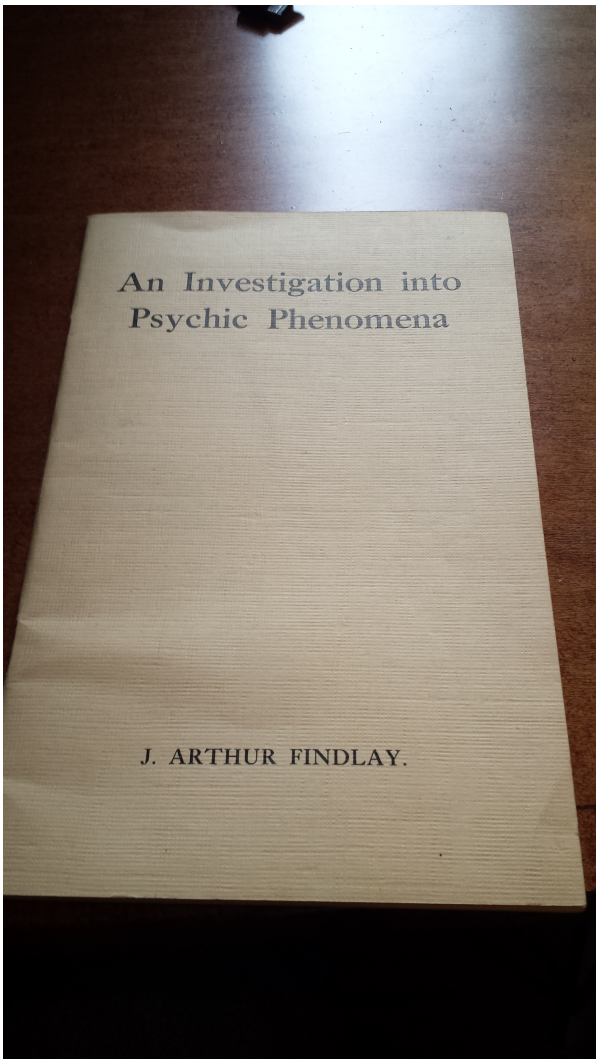
11. Have someone else review the PDF for quality.

12. Release the PDF into the wild.

To illustrate this workflow, we'll use a recent project we did at IAPSOP -- the preservation of some brittle Spiritualist pamphlets from the early decades of the twentieth century, donated for preservation (with the expectation that the originals would be returned) by the librarian of a prominent Spiritualist organization.

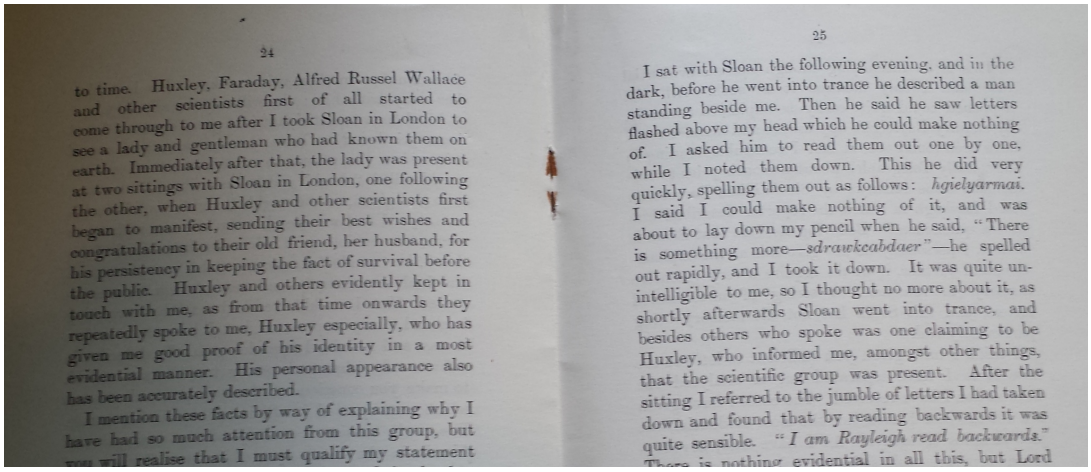
1 -- Select A Digitization Strategy

Among the pamphlets we digitized during this project was this one:



Example: J. Arthur Findlay's *An Investigation into Psychic Phenomena* (c. 1924)

This pamphlet's cover and pages are printed on high-quality, low-acid paper, and the pamphlet is in a remarkably good state, for a 89-year old artifact. However, the first signs of degradation are already present, in the center of the pamphlet: oxidation of the staples.



24

to time. Huxley, Faraday, Alfred Russel Wallace and other scientists first of all started to come through to me after I took Sloan in London to see a lady and gentleman who had known them on earth. Immediately after that, the lady was present at two sittings with Sloan in London, one following the other, when Huxley and other scientists first began to manifest, sending their best wishes and congratulations to their old friend, her husband, for his persistency in keeping the fact of survival before the public. Huxley and others evidently kept in touch with me, as from that time onwards they repeatedly spoke to me, Huxley especially, who has given me good proof of his identity in a most evidential manner. His personal appearance also has been accurately described.

I mention these facts by way of explaining why I have had so much attention from this group, but

25

I sat with Sloan the following evening, and in the dark, before he went into trance he described a man standing beside me. Then he said he saw letters flashed above my head which he could make nothing of. I asked him to read them out one by one, while I noted them down. This he did very quickly, spelling them out as follows: *hgielyarmai*. I said I could make nothing of it, and was about to lay down my pencil when he said, "There is something more—*sdrawkcabdaer*"—he spelled out rapidly, and I took it down. It was quite unintelligible to me, so I thought no more about it, as shortly afterwards Sloan went into trance, and besides others who spoke was one claiming to be Huxley, who informed me, amongst other things, that the scientific group was present. After the sitting I referred to the jumble of letters I had taken down and found that by reading backwards it was quite sensible. "*I am Rayleigh read backwards.*"

Example: Staple oxidation in an instance of J. Arthur Findlay's *An Investigation into Psychic Phenomena* (c. 1924)

Unless this pamphlet is stored in a low-humidity temperature-controlled environment (which the donor's library does not have, to our knowledge), the staples will continue to oxidize, eventually crumbling entirely and/or eating through the paper and cover, leaving the pamphlet disbound by time.

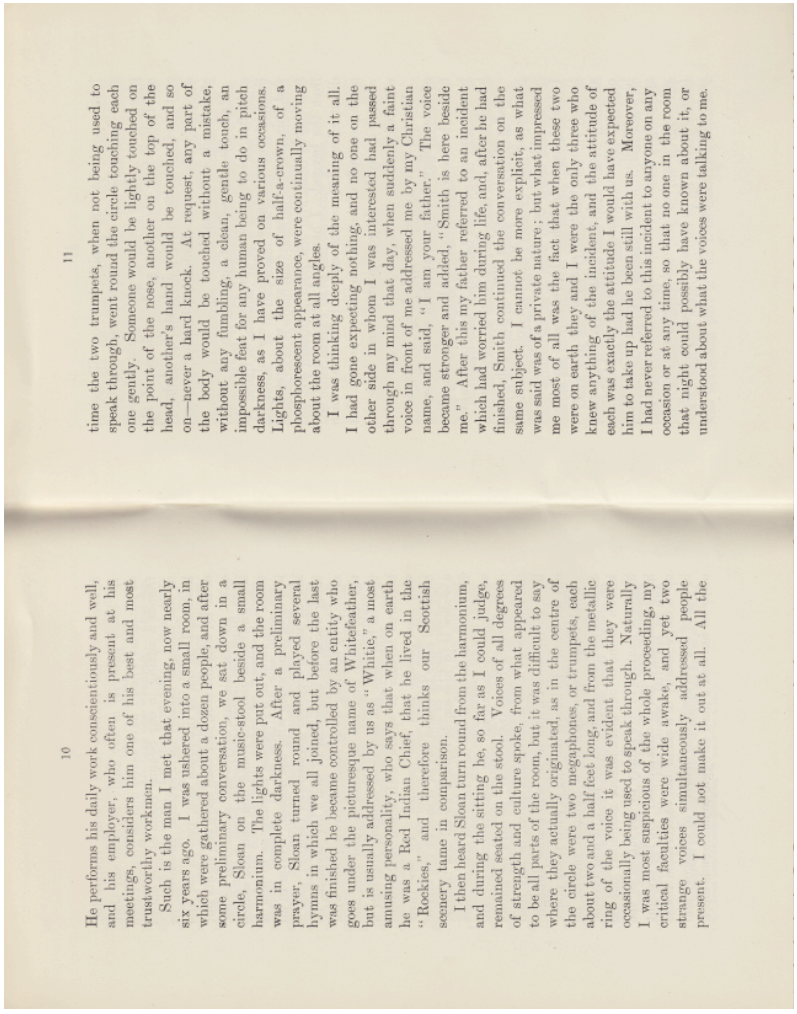
We considered pulling the staples, scanning the pamphlet as loose sheets in a sheet-feed scanner, and then restapling the pamphlet, using a long-reach stapler and modern staples, which have low (or no) iron content.

Example: A long-reach stapler capable of restapling at up to a 12" depth.

If we owned the pamphlet, that is what we would have done, but as it is not IAPSOP property, and we did not wish to alarm the donor, we used a flatbed scanner to scan the entire pamphlet in its bound state.

2 -- Convert The Physical Object

We converted the pamphlet on a flat-bed scanner, producing two-page 600 DPI TIFF images, as exemplified below, cropping the images at the outer bound of the page during each two-page scan, to give ourselves the maximum outer gutter to work with. Note the gutter shadows in the center of the image; these will be minimized or removed during processing.



10

He performs his daily work conscientiously and well, and his employer, who often is present at his meetings, considers him one of his best and most trustworthy workmen.

Such is the man I met that evening, now nearly six years ago. I was ushered into a small room in which were gathered about a dozen people, and after some preliminary conversation, we sat down in a circle, Sloan on the music-stool beside a small harmonium. The lights were put out, and the room was in complete darkness. After a preliminary prayer, Sloan turned round and played several hymns in which we all joined, but before the last was finished he became controlled by an entity who goes under the picturesque name of Whitefeather, but is usually addressed by us as "Whitie," a most amusing personality, who says that when on earth he was a Red Indian Chief, that he lived in the "Rockies," and therefore thinks our Scottish scenery tame in comparison.

I then heard Sloan turn round from the harmonium, and during the sitting he, so far as I could judge, remained seated on the stool. Voices of all degrees of strength and culture spoke, from what appeared to be all parts of the room, but it was difficult to say where they actually originated, as in the centre of the circle were two megaphones, or trumpets, each about two and a half feet long, and from the metallic ring of the voice it was evident that they were occasionally being used to speak through. Naturally I was most suspicious of the whole proceeding, my critical faculties were wide awake, and yet two strange voices simultaneously addressed people present. I could not make it out at all. All the

11

time the two trumpets, when not being used to speak through, went round the circle touching each one gently. Someone would be lightly touched on the point of the nose, another on the top of the head, another's hand would be touched, and so on—never a hard knock. At request, any part of the body would be touched without a mistake, without any fumbling, a clean, gentle touch, an impossible feat for any human being to do in pitch darkness, as I have proved on various occasions. Lights, about the size of half-a-crown, of a phosphorescent appearance, were continually moving about the room at all angles.

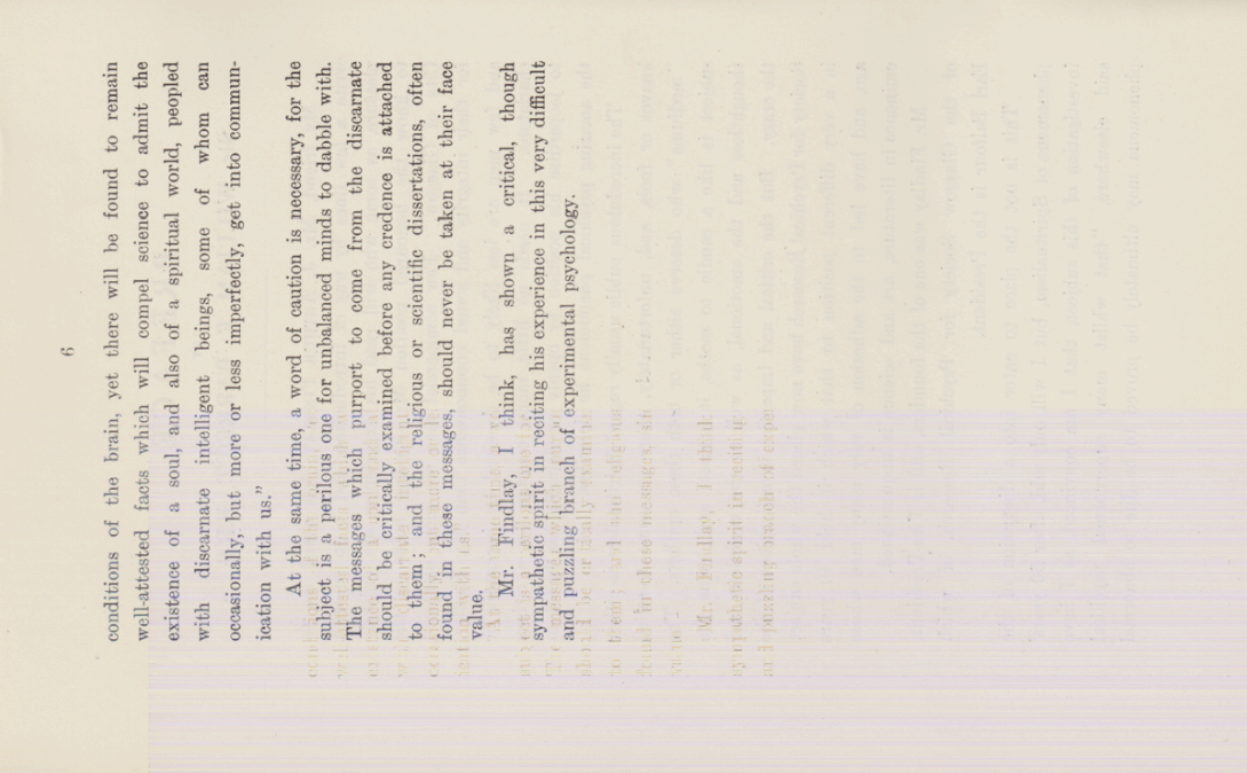
I was thinking deeply of the meaning of it all. I had gone expecting nothing, and no one on the other side in whom I was interested had passed through my mind that day, when suddenly a faint voice in front of me addressed me by my Christian name, and said, "I am your father." The voice became stronger and added, "Smith is here beside me." After this my father referred to an incident which had worried him during life, and, after he had finished, Smith continued the conversation on the same subject. I cannot be more explicit, as what was said was of a private nature; but what impressed me most of all was the fact that when these two were on earth they and I were the only three who knew anything of the incident, and the attitude of each was exactly the attitude I would have expected him to take up had he been still with us. Moreover, I had never referred to this incident to anyone on any occasion or at any time, so that no one in the room that night could possibly have known about it, or understood about what the voices were talking to me.

Example: A 600 DPI two-page spread of the Findlay pamphlet.

Each image -- comprising two pages of the pamphlet -- was between 49 and 53 megabytes (MB) in size, and the image set for the entire pamphlet including verso and recto of front and back covers consumed 3.7 gigabytes (GB) of hard disk storage on the OS X workstation on which it was processed.

3 -- Inspect Each Image

All scanners, periodically, distort images. The cheaper the scanner, the more often it happens. So, to ensure high quality digitization, you need to inspect each image after the imaging run is done, and before you release the physical artifact.



6

conditions of the brain, yet there will be found to remain well-attested facts which will compel science to admit the existence of a soul, and also of a spiritual world, peopled with discarnate intelligent beings, some of whom can occasionally, but more or less imperfectly, get into communication with us."

At the same time, a word of caution is necessary, for the subject is a perilous one for unbalanced minds to dabble with. The messages which purport to come from the discarnate should be critically examined before any credence is attached to them; and the religious or scientific dissertations, often found in these messages, should never be taken at their face value.

Mr. Findlay, I think, has shown a critical, though sympathetic spirit in reciting his experience in this very difficult and puzzling branch of experimental psychology.

Example: Scanner distortion in one of the page images for the Findlay pamphlet.

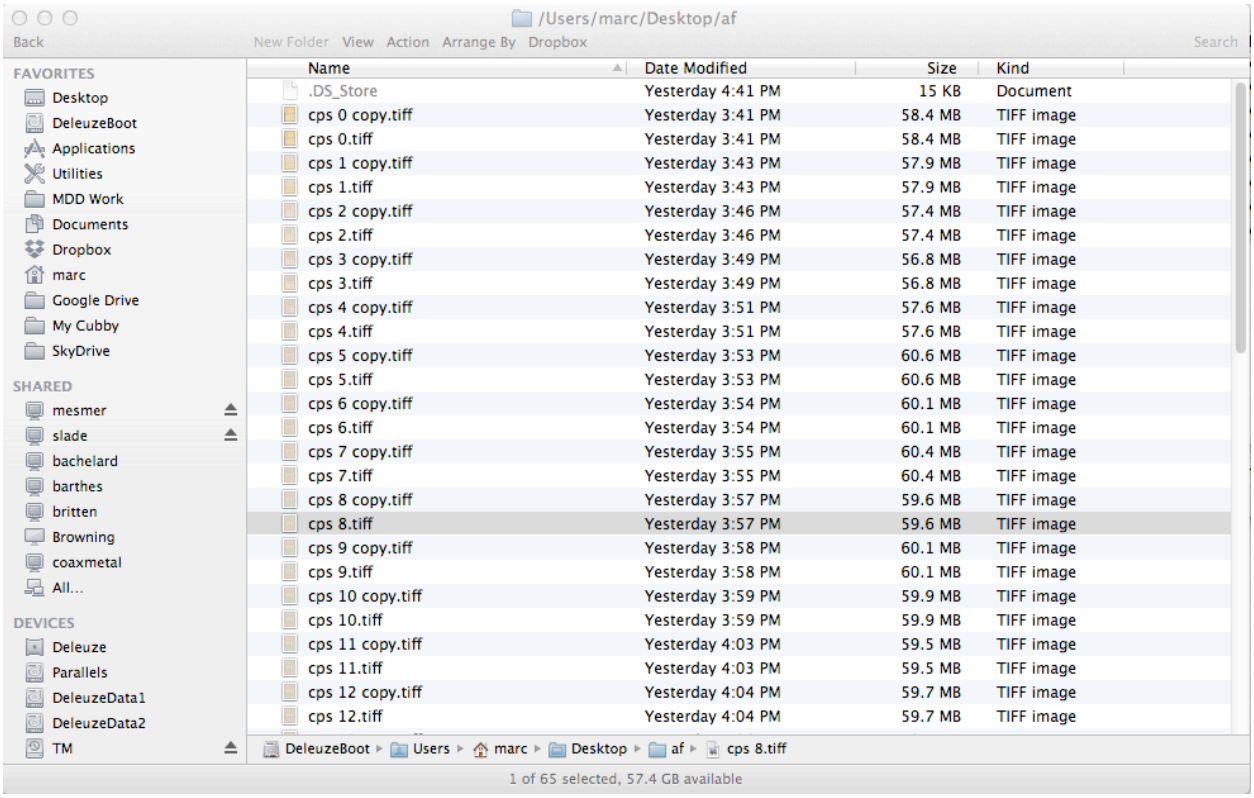
Retake any distorted images before continuing with the workflow.

4 -- If Necessary, Duplicate Page Images

The rotation of the pages in the scanned images of the Findlay pamphlet is a non-issue for the preservationist; Acrobat will resolve that problem, in a few seconds.

But, because each image contains two pages, we need two copies of each image, since it is a standard part of our workflow to present one image per page, in the final PDF

output. Thus, we copy each image file -- so that we can crop the images down to single page images) before turning to Acrobat Pro to assemble the PDF.



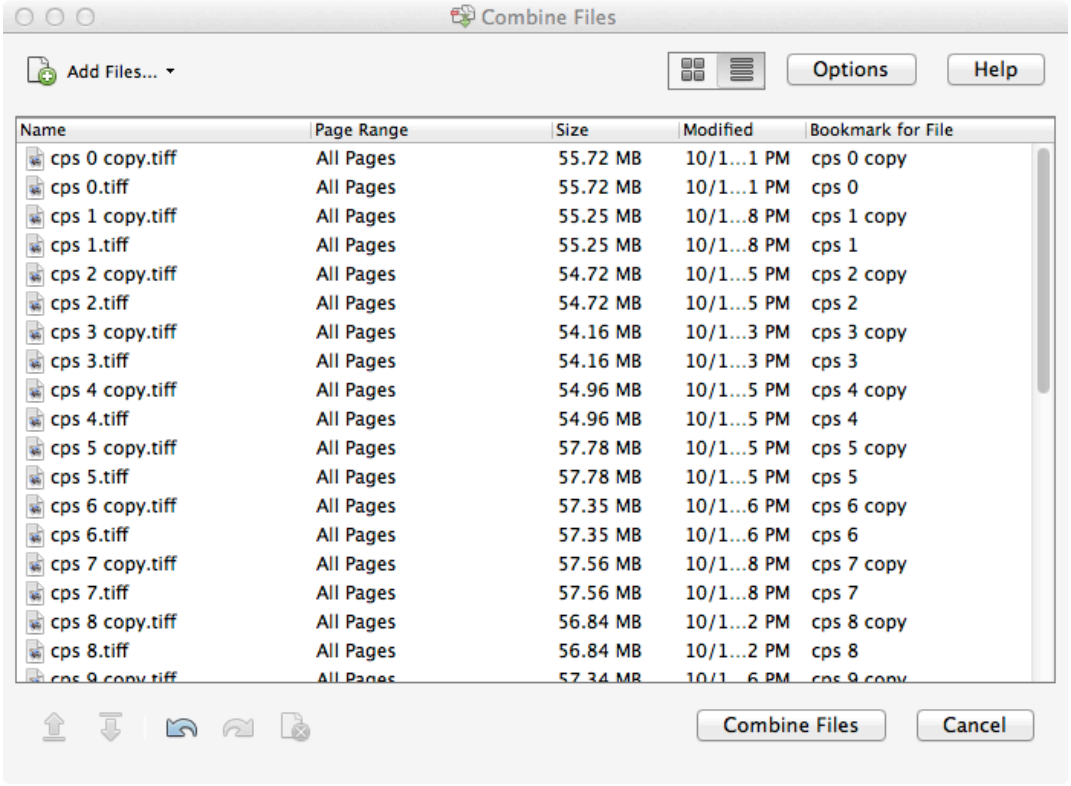
5 -- Combine The TIFF Images in Acrobat

Acrobat is designed to combine image files, and to create a PDF “wrapper” around TIFF images such that those images can be subsequently extracted from the PDF for other image processing purposes.

All our preservationist needs to do, to turn the 65 TIFF images that form the basis for our digitized version of the Findlay pamphlet into a single PDF file, is: invoke the Adobe Acrobat Pro “combine” function.

In both Windows and OS X, that function is invoked by: File --> Create --> Combine Files Into A Single PDF (in the current version of Acrobat Pro). Because of the way Acrobat is coded by Adobe, the functionality of the application is mostly identical across operating systems, so we’ll use OS X versions of the Acrobat functionality from here forward in this walk-through.

The TIFF files are loaded into the Combine function’s interface.



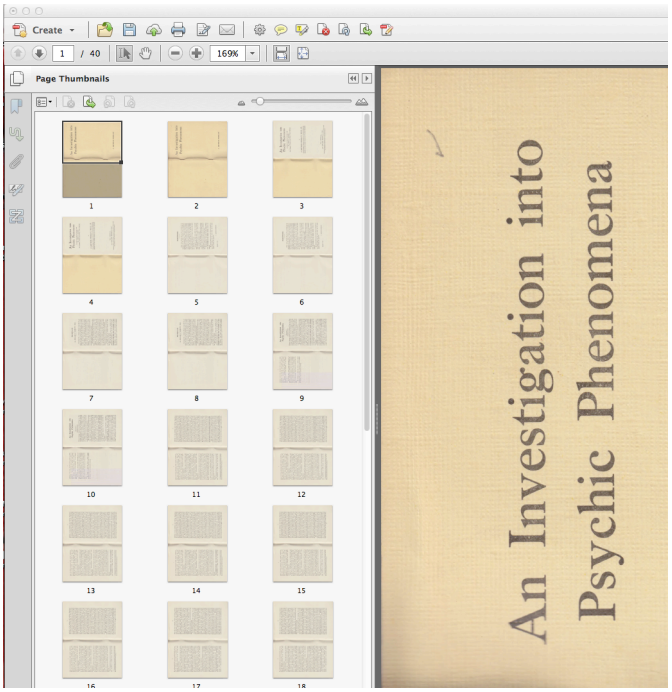
Press “Combine Files” and the initial PDF is created and displayed. This combination process can take quite a long time when the TIFF images are large, when the number of TIFF images are large, or when the workstation on which the work is being done is underpowered (too little CPU or memory). In some cases, an underpowered workstation can result in corrupt combinations, or abrupt termination of Acrobat Pro.

6 -- Rotate The Pages In Acrobat As Necessary

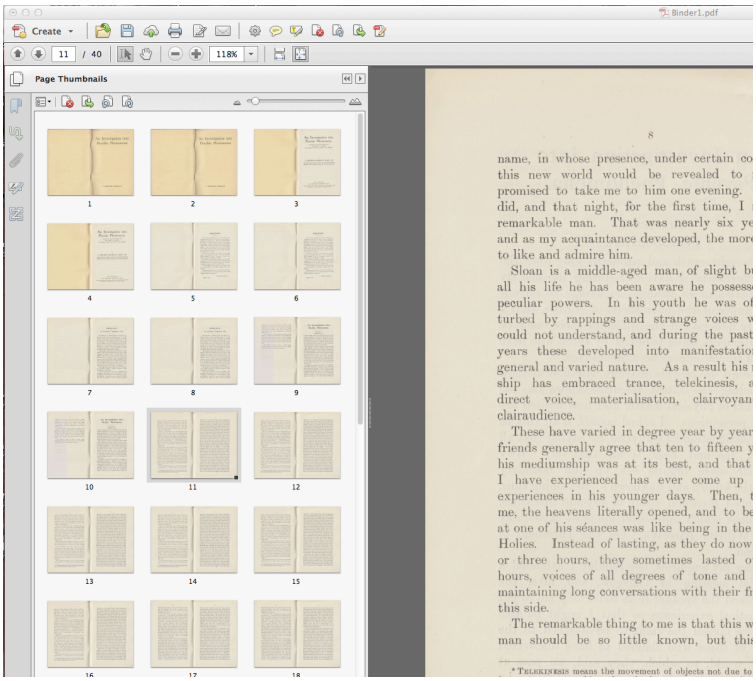
In this example, we have two-page spread images, in portrait mode, and we need those images in landscape mode in order to get on with cropping the two-page images down to single-page images.[14]

To do that in a single operation, open the Acrobat left-side panel in “page thumbnail” model by clicking on the left margin of the Acrobat form with the mouse button you do not normally use, and select “Page Thumbnails” from the pop-up menu. Select the first image in the thumbnail view, and then click Edit -> Select All to select ALL the thumbnail images.

[14] Note that plenty of amateur preservationists don’t crop two-page spreads down to single-page images. They release items as two-page spread PDFs. That’s a legitimate choice, but it can make for poor OCR/indexing results, since Acrobat’s in-built OCR engine is (a) crappy for everything but high-quality black-and-white images, and (b) unable to recognize that it’s dealing with two-page spreads when it rasterizes images for OCR.



Then, click anywhere in the thumbnail view with the mouse button you don't normally use, select "Rotate Pages" from the pop-up menu, and rotate the pages appropriately (in our example, clockwise 90 degrees).



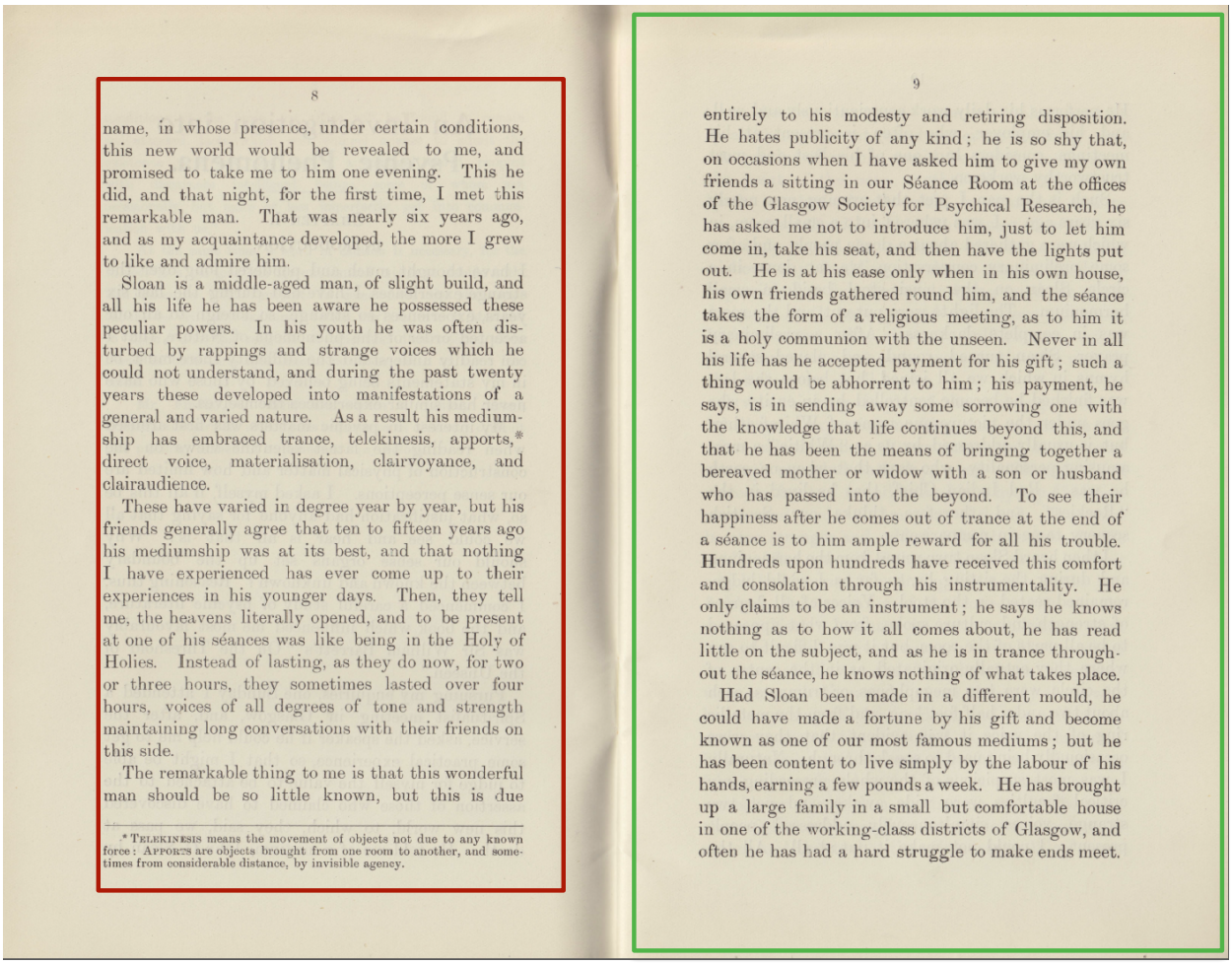
The spreads are now oriented for cropping.

7 -- Crop The Pages In Acrobat

Cropping is necessary if you want to present single-page images in PDF, and optimize the chances that Acrobat's in-built OCR engine will do a decent job OCRing and indexing the PDF file. It may also be necessary to remove gutter shadows from the final version of the document.

Cropping is also tricky. You want to do the absolute minimum crop for each page -- the crop that retains as much of the negative space around the text block as possible. The example below attempts to illustrate this concept.

OVERLY AGGRESSIVE CROPPING YOU WILL PAY FOR LATER **CONSERVATIVE CROPPING YOU WILL BENEFIT FROM**



8

name, in whose presence, under certain conditions, this new world would be revealed to me, and promised to take me to him one evening. This he did, and that night, for the first time, I met this remarkable man. That was nearly six years ago, and as my acquaintance developed, the more I grew to like and admire him.

Sloan is a middle-aged man, of slight build, and all his life he has been aware he possessed these peculiar powers. In his youth he was often disturbed by rappings and strange voices which he could not understand, and during the past twenty years these developed into manifestations of a general and varied nature. As a result his mediumship has embraced trance, telekinesis, apports,* direct voice, materialisation, clairvoyance, and clairaudience.

These have varied in degree year by year, but his friends generally agree that ten to fifteen years ago his mediumship was at its best, and that nothing I have experienced has ever come up to their experiences in his younger days. Then, they tell me, the heavens literally opened, and to be present at one of his séances was like being in the Holy of Holies. Instead of lasting, as they do now, for two or three hours, they sometimes lasted over four hours, voices of all degrees of tone and strength maintaining long conversations with their friends on this side.

The remarkable thing to me is that this wonderful man should be so little known, but this is due

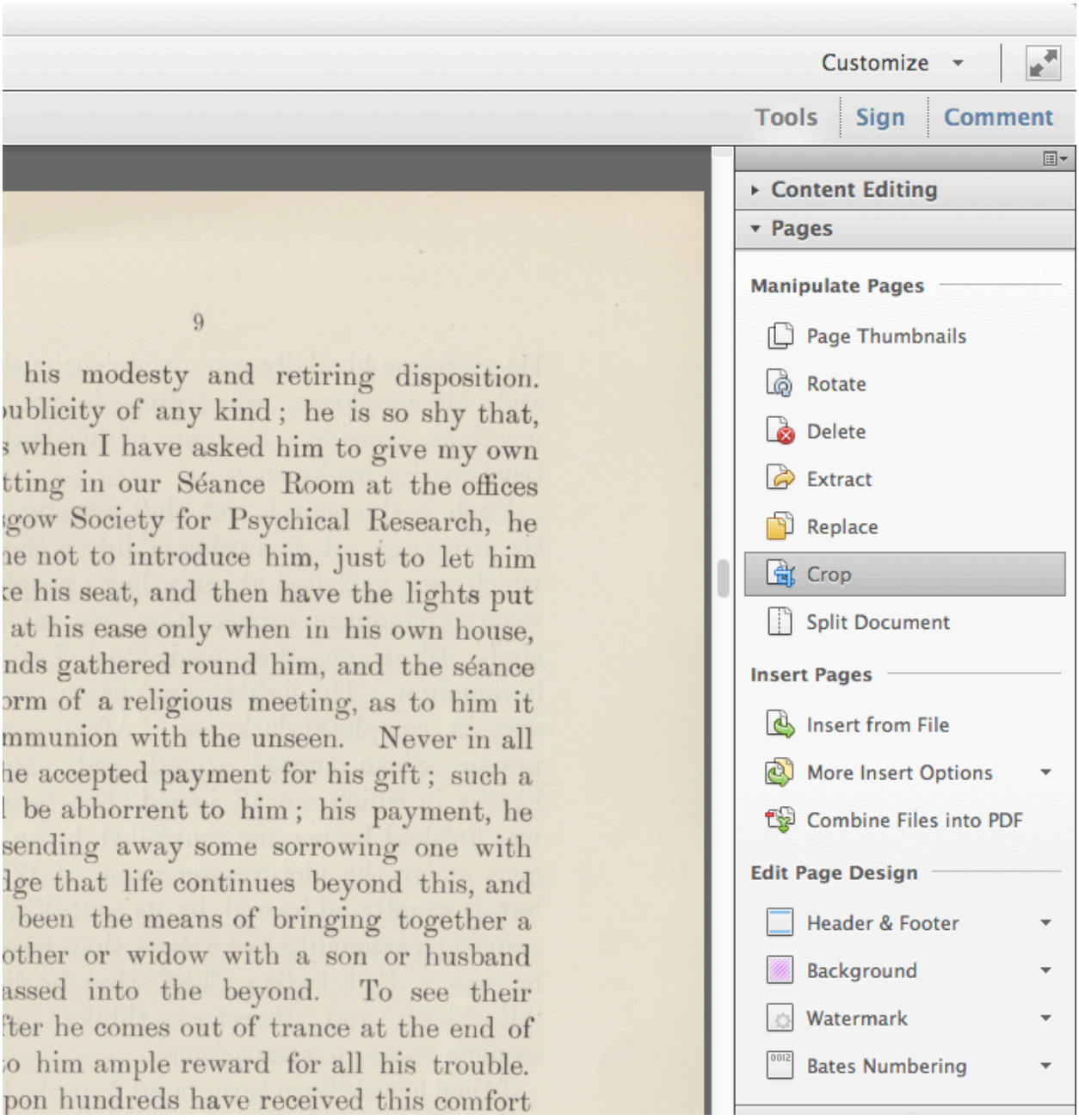
* Telekinesis means the movement of objects not due to any known force: Apports are objects brought from one room to another, and sometimes from considerable distance, by invisible agency.

9

entirely to his modesty and retiring disposition. He hates publicity of any kind; he is so shy that, on occasions when I have asked him to give my own friends a sitting in our Séance Room at the offices of the Glasgow Society for Psychical Research, he has asked me not to introduce him, just to let him come in, take his seat, and then have the lights put out. He is at his ease only when in his own house, his own friends gathered round him, and the séance takes the form of a religious meeting, as to him it is a holy communion with the unseen. Never in all his life has he accepted payment for his gift; such a thing would be abhorrent to him; his payment, he says, is in sending away some sorrowing one with the knowledge that life continues beyond this, and that he has been the means of bringing together a bereaved mother or widow with a son or husband who has passed into the beyond. To see their happiness after he comes out of trance at the end of a séance is to him ample reward for all his trouble. Hundreds upon hundreds have received this comfort and consolation through his instrumentality. He only claims to be an instrument; he says he knows nothing as to how it all comes about, he has read little on the subject, and as he is in trance throughout the séance, he knows nothing of what takes place.

Had Sloan been made in a different mould, he could have made a fortune by his gift and become known as one of our most famous mediums; but he has been content to live simply by the labour of his hands, earning a few pounds a week. He has brought up a large family in a small but comfortable house in one of the working-class districts of Glasgow, and often he has had a hard struggle to make ends meet.

Cropping in Acrobat is simple: you select Tools (top right bar) --> Pages --> Crop.



Then, for each half of the two page spread, page by page, draw a bounding box around the page you wish to retain, and press Return/Enter twice. That will crop the page, retaining what is inside the bounding box.

8 -- Check Pagination, Flow and Image Quality

When you are done cropping, save the PDF file, using a different file name than your previous version.[15]

Then, starting at the beginning of the PDF file, check every page, with three questions in mind:

1. are all pages present?
2. are all pages of equal, and high, quality?
3. does the test scan properly across page breaks?

This last question gets at a common problem in preserving texts from the period of typesetting-by-hand: page numbers are, more often than you might think, garbled. For example: page 23 appears twice (two unique text blocks are numbered 23) and page 217 is absent (the physical artifact's pagination is 215, 216, 218). If you're doing bibliographical description, you way to note this in your bibliographical description, but in every case, you want to make sure that all pages in the physical text are present in the digital version of the text, and that all sentences scan properly across page breaks.

9 -- OCR In Acrobat

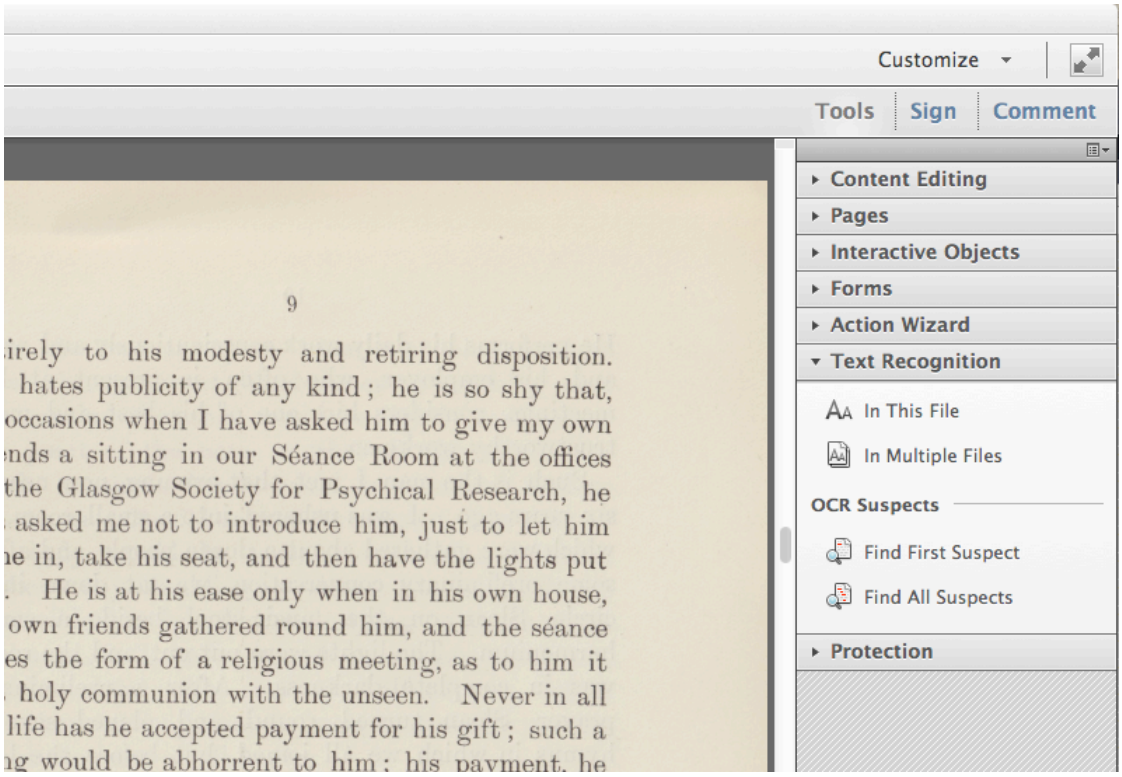
Acrobat Pro has a built-in OCR engine of third-rate quality that embeds what it OCRs in an "index" within the PDF, making the PDF searchable in Acrobat, and making the PDF indexable by search engines (most of which will index the embedded index in PDF files when they find them, and none of which, to my knowledge, re-index PDF files when their crawlers find those PDF files on the Web). Additionally, most modern operating systems have in-built search facilities that "know" about Acrobat indexes, and exploit them when they find them -- making indexed PDFs searchable on your local machine.

The OCR function in Acrobat is in the same panel as the Crop function: it's called "Text Recognition".

---

[15] We usually save the PDF file at the end of each step, using a different file name (filename_stage1.pdf, filename_stage2.pdf, etc.), so that if we make a terrible mistake in one step, we can go back to the file as it existed at the end of the prior step, and don't have to recreate the file from scratch again. At the end of the process, we delete all versions of the file except (a) the initial build and (b) the final released version, and retain both of those versions with the TIFF images that we used to create the PDF file. That artifact set -- two PDFs and a set of TIFF images -- are retained in our internal archive in case (a) we discover we've made a mistake after release and need to rebuild the PDF and (b) some unforeseen circumstance arises. Disk space is cheap. Keep your final product and its raw input, to be safe.

Note: You may be asked if you want to “downsample” your high-resolution images during this process. We do not recommend downsampling to any resolution below 300 DPI. Downsampling will reduce the size of the PDF, sometimes significantly, and that has value, particularly for the millions of Internet users who have to download objects over slow and/or expensive Internet connections.



The function allows for OCR corrections -- “find suspects” -- if you’re inclined to do that. We generally don’t do that; we just let the OCR engine do its thing, and then dump the index as a flat text file, and inspect that to see how well the OCR engine did at indexing the file.

10 -- Dump The OCR’d Index And Check It

After the Acrobat OCR engine has recognize the text in the file -- or tried to -- save the file with a different file name, and then dump the index to a flat text file, to inspect the quality of the OCR job.

To do that, click File --> Save As Other --> More Options --> Text (plain)

You can open the resulting file in any text editor and see how well the OCR engine did with the file.

Unfortunately, there are no good guidelines we can give you, if the OCR engine did a poor job with the file, to fix that problem. We have found files, assembled by different methods (for example, from medium-resolution JPG files by third-party PDF creators) that Acrobat's in-built engine refuses to recognize at all (not a single character is recognized), and nothing we've tried has remedied that problem. The in-built OCR functionality is a bit of a mystery, unfortunately. One thing that does often work to remedy poor OCR in Acrobat, with TIFF -based documents that are true-color (like the example we are using here) is: converting those TIFF images to black-and-white before creating the initial PDF file.

Really, you're not checking the OCR'd index results to produce a perfect index (or, rather, we don't think that's part of the preservation problem); you're checking the OCR'd index results so you can warn potential users if the index is poor or non-existent, so that they don't assume the index is good, and miss things they are searching for.

11 -- Have Someone Else Review Your Work

By this time in the workflow, you have spent hours or perhaps days, in the case of a book, with this digital representation, and you are likely to miss more-or-less obvious problems that a fresh set of eyes will pick up immediately.

So, let someone else look at your work before you release the PDF into the wild.

12 -- Release The PDF Into The Wild

We believe this is the most important step in the workflow. Making yourself a digital copy of a physical printed artifact does not in fact preserve it, except for you. If your disk dies, or your PC is stolen, the digital representation is lost.

What does optimize that object's chances of survival is: **making many copies of it**. And the Internet does that better than any other mechanism in human history.

The good news is: you don't need to run a web site, or be technically sophisticated, to give up a copy of your digital object to the Internet. Just go to www.archive.org (the Internet Archive), and upload a copy of your PDF into the Internet Archive's documents collection. Become a little famous if you like, and identify yourself as the source.

That collection is indexed by the search engines, and your document will be found, downloaded and used by people who are searching for it, or for content it contains. They will make copies, as will sites that mirror the Internet Archive's contents.

The document will survive. It will outlast all of us.

Conclusion

This is, of necessity, a quick-and-dirty guide. The only way to learn how to do this kind of work is: practice.

Many preservationists will have issues with the contents of this document. Please feel free to express your concerns, to marc@ehbritten.org. If the document can or should be modified to reflect your concerns or issues, it will be. If not, feel free to write your own guide. The more guides there are, the better job we'll all do preserving printed material for posterity.