FROM PERFECT TO PREPOSTEROUS: HOW DIGITAL RESTORATION CAN BOTH HELP AND HINDER OUR READING OF DAMAGED SOURCES

JULIA CRAIG-MCFEELY
DIGITAL IMAGE ARCHIVE OF MUSIC MANUSCRIPTS (DIAMM)

ABSTRACT

Digital reproductions of manuscripts are now reasonably easy to obtain from a very wide variety of libraries, although persuading suppliers to meet the standards required for digital restoration - or even high-quality reproduction - is surprisingly difficult. As more researchers tackle image-processing to help them read damaged originals, the need for suppliers to meet a baseline standard is much greater. The paper examines a set of basic guidelines for requesting images for research purposes A demonstration of some 'virtual restoration' techniques on a variety of different types of damage as well as photographic techniques such as ultra-violet, infra-red and multi-spectral imaging as adjuncts to normal color imaging will also be discussed.

Digital restoration involves the transformation of digital objects to enhance our use of them and improve their scholarly yield. Despite the speed with which technology is changing, the techniques of working with digital objects discussed here has not changed substantially since 1998, when Adobe Photoshop was first used to improve a reading of a digitally captured document. The computers are faster, the files are larger, and the quality of the images is sharper and clearer, but not as much as one might imagine given the lapse of twelve years. This is a significant lack of change in a field that is akin to the proverbial river into which one cannot step twice.

Digital restoration is an entirely visual activity, and the techniques used do not lend themselves well to description in the static medium of printed text, particularly as every image requires a different technique to obtain the best results. The following sets forth some general information regarding basic digital restoration processes, and aims to explain what is involved to a sufficient extent that it will allow a better appraisal of restored images presented in support of research. Readers are referred to the Digital Image Archive of Music Manuscripts (DIAMM) Digital Restoration Workbook for a fairly comprehensive introduction to the use of Adobe Photoshop for restoring images.[1], and a representative selection of techniques in use by the author.

As digital restoration of necessity requires a reasonable familiarity with the digital world, the impact of the rapid growth of internet usage and of digital research on digital humanities should be considered in a broader context. In a world where discrimination by age, sex, race, or religion is illegal, businesses and services in the UK now actively and pervasively exclude the non-internet user – discriminating in fact against those who either choose not to use a computer or who are unable to do so. This section of society (most usually the older generation), is frustrated by television advertising repeatedly offering online discounts particularly for insurance or financial services; almost all high-street purchases can be made cheaper on the web, and many companies now will only accept

[1] Downloadable from www.diamm.ac.uk.

communication via email. These types of promotions lead those who cannot afford a computer or the cost of an online connection to become ever more disconnected, losing out on services that they cannot obtain without a computer.

The same is true in a more subtle way of humanities scholarship: increasingly discussions of new findings or other activity relating to a subject area are conducted on social networking sites or email lists, and those not subscribed are left unaware of the content that is circulating among other scholars. What digital scholarship has done is to marginalize a whole generation of scholars, and in the process, their work. Much pre-digital research is losing its currency for no better reason than that it is not available online. This 'non-digital' group does not have access to the things that the internet generation takes for granted. This medium is becoming increasingly important, yet, at the opening of the twenty-first century, we are frustrated by the need to gain credit for online outputs from a community that in large part still resists the online medium and doesn't want to recognize its use as legitimate publication on a par with print and paper.

Digital restoration falls into this unfortunate crevice: restorations allow other academics to work from documents that they could not otherwise read, and the best place to see those restorations is on the web. However, with unmediated and disreputable content rubbing shoulders on the web with genuine research, getting credit for this work is nearly impossible because it is not in print. For example, the creation of the massive database framework that supports the DIAMM online resource is not recognized as a research publication even though it has contributed significantly to the advancement of our knowledge of, and facilitates research into, medieval music. Restoration is a time-consuming process, and it is necessary to balance the gains that might be made in a broad context, against the time spent doing it.

PROCESSES

Almost any mainstream image-processing software can be used for digital restoration; however, what cannot be used are free or cut-down versions of the full software. Cut-down versions will not open images as large as those required for restoration. They also do not provide a lot of the features that are necessary for restoration, even though restoration requires only a tiny proportion of what the full software offers.

I have been using Adobe Photoshop since 1998, and one of its most useful features is the facility to create layers that overlay the original, just as a pile of transparent flimsies can be stacked up on an overhead projector. The difference between the virtual and physical in this case is that as well as changing the order of layers it is also possible to change the way in which a layer interacts with those beneath it, with often startling results.

A 12 MB image from a high-quality digital single lens reflex (SLR) camera is widely considered to be 'large', but the sort of image needed for restoration will usually be 10-30 times that size or more and can only be created with specialist imaging equipment. In this context, 'high resolution' means images that will probably nowadays be in the region of 300 MB, and this size will increase to a gigabyte and more during restoration. Restoration therefore requires a computer with a fast processor and more than a standard installation of RAM. It is sometimes possible to get results with lower-quality images, but the worse the damage is to the original, the more quality (i.e. resolution) is necessary in the primary capture.

Scanning surrogates such as photos, slides and microfilms at high resolution does not create a high-resolution image of the original since increasing the scan resolution cannot increase the level of detail on the original photograph. It is only by going back to the original that the best information available can be obtained.

Working from JPG images means that really detailed work is impossible.[2] JPG compression destroys detail deliberately: the compression algorithm saves colors that are nearly similar as the

[2] JPG stands for 'Joint Photographic Experts Group' (sometimes expressed as JPEG. It is a compression format decided on by committee of photographic experts in 1986. More about the JPEG committee can be found here:

same color. Since digital restoration processes rely on very fine differentiation of color – such fine differentiation in fact that it often cannot be seen with the naked eye, only electronically – the use of JPG compression causes the loss of the fine detail essential to the process. Uncompressed TIF images, without any form of compression at any point during the workflow, are a better option, and the format also ensures the image can be stored for many years and still be readable to future software. Once an image has been saved as a JPG it is not possible to recover the lost color or detail information.

Attempting to obtain images of this quality can result in a multitude of problems. Many libraries buy smaller equipment because they cannot afford the high-resolution cameras, but many others make decisions based on their inability to see the difference between one type of image and another. Even staff in imaging departments often cannot see the difference between an image that has been artificially 'improved' and one that was born correct. Ordering images from a leading national institution will not ensure receipt of the best digital image possible. Indeed, the chances are that the quality will be questionable. Many institutions consistently supply out-of-focus, incorrectly calibrated, low-resolution pictures in JPG format and with sharpening applied to make up for the deficiencies in their original capture. This last adjustment falsifies color and detail information that in many cases is the sole reason the images are needed. One process that is commonly used to make poorly-focused images look better to the untrained eye is to apply an "unsharp mask". This filter looks for edges between dark and light colors and exaggerates the contrast by darkening the dark edge and lightening the light edge. This creates a corona around all the material on the page: letters, notes, staff lines, dirt, spots, etc. It will also blur fine detail such as pen-strokes and texture on paper or parchment. Unfortunately, it is in the fine detail, where there is hardly any visible definition, that the information needed for restoration is most often found.

Digital images should be far better than their analog counterparts, but many libraries believe they are poor substitutes because they have not seen a really fine digital image. As end-users we have a responsibility to demand the best quality service from suppliers. That means understanding what we are asking for. Just as libraries have a responsibility to supply excellent images, we in turn have to be informed about digital image media and how to evaluate what we are given.[3]

One extremely valuable feature of digital images is that the original can be kept safe and restoration work only undertaken on copies, so no matter how much the copy is changed during restoration it is always possible to go back to the original and start again. This means that, unlike real-world restoration, the digital restorer can experiment and get it wrong as much as they like : nothing is lost or permanently damaged by experimentation.. However, experimentation takes up space: an issue to consider if setting out to restore a large number of images is that storage of originals and layered restored versions require considerable backup and primary disc space, as well as a fast computer.

Many sources that we cannot read easily now have become problematic because the ink has faded or because the surface is covered with dirt. The easiest restoration process involves simply adjusting the levels of dark and light so that limited contrast is exaggerated. The process is outlined here in figures 1a-e. The original may range in color from pale brown to a paler brown. By adjusting what the software understands as white and black (the value at the extreme ends of the histogram), the difference between dark and light is exaggerated, allowing the eye to differentiate more easily.

---

http://www.jpeg.org/. JPG2000 is a new format that is currently not readable by many softwares and browsers, and uses a proprietary compression and storage format that can only be opened by certain software. It is therefore not a safe long-term storage format.

[3] This information is supplied in the DIAMM Digital Restoration Workbook (www.diamm.ac.uk/publications.html.

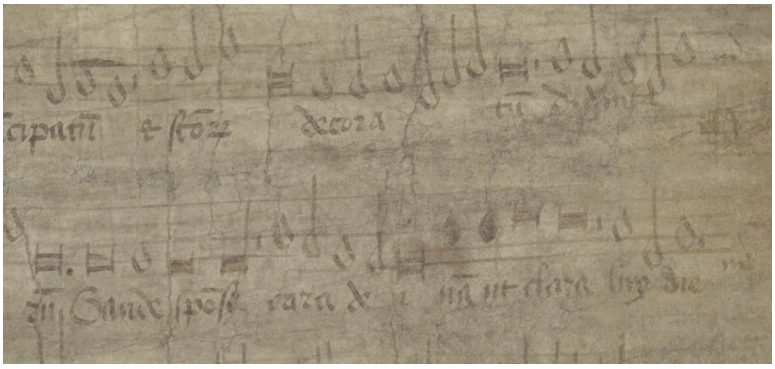

Figure 1a: Original image

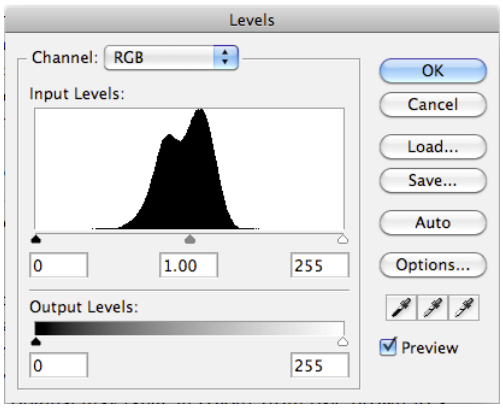

Figure 1b: Original histogram. The left-hand end of the graph represents black values, and the right-hand end white values (note that the available colors are bunched together in the mid-range, showing no true black or white on the image).

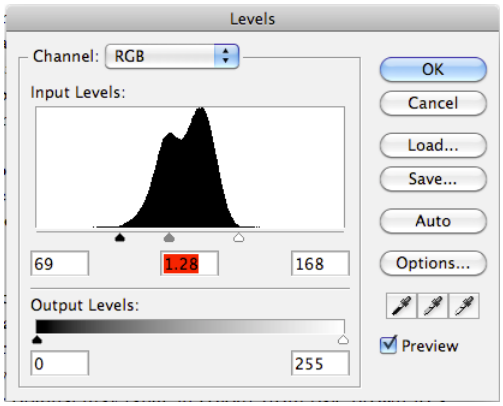

Figure 1c: The black and white sliders have been moved to the edges of the available colors, telling the software that the leftmost part of the spectrum is black and the rightmost is white…

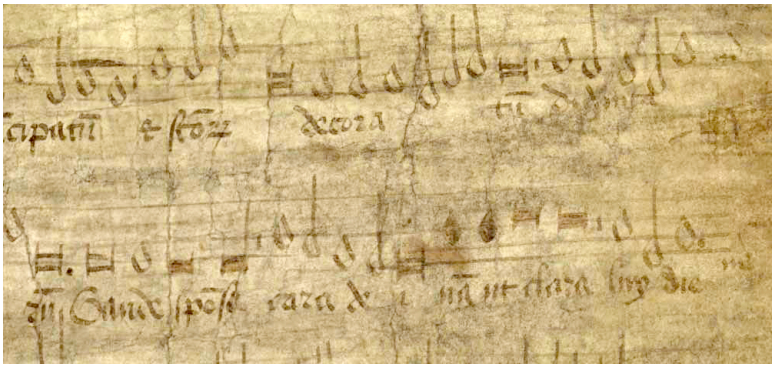

Figure 1d: Adjusted image

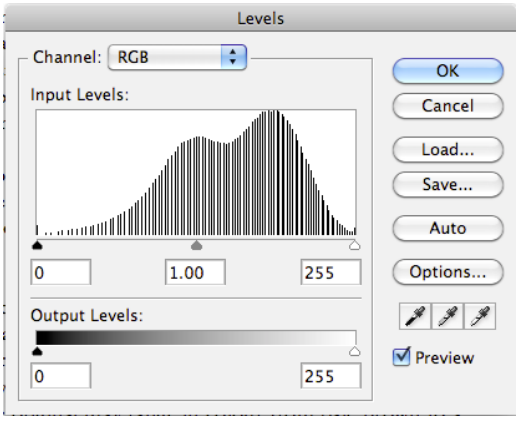

Figure 1e: New histogram for the adjusted image showing that the color values available have been spread out across the spectrum from black to white (leaving some gaps), rather than being bunched together in the middle of the spectrum.

The result may not look very attractive, but that is not the aim: the goal is readability. For about 75% of damaged sources, this process is all that may be needed to read a faded or dirt-covered text.

Any restoration process should ideally be applied over a *complete* image, not just portions: as soon as a particular area of a document is selected to the exclusion of other areas an editorial decision has been made both about what is on the page and about what the restorer expects to see: this is analogous to deciding an outcome in advance and then making the evidence support that conclusion. If a process is applied over a whole page then that minimizes the danger of predetermining the outcome. Of course there will almost certainly come a point where it is necessary to select only a part of the page, but that choice is usually obvious rather than editorial.

In a significant proportion of cases the simple level-adjust method is not enough to lift the hidden text sufficiently from the background color, and in this case there are two methods that can be used: the first is color selection, and the second is to use a high-pass filter in Adobe Photoshop and change the overlay effect on the material beneath it.

A fair amount of time is spent looking at the image, attempting to work out what is there, and trying to ignore the brain's attempts to make sense out of the shapes it perceives. It is often necessary to resist the impulse to interpret shapes and colors in order to avoid making advance editorial decisions that may be wrong. One useful skill in digital restoration is the ability to ignore the content of the image and simply view it as a collection of pixels of different tones and colors.

Color selection relies on uncorrupted color information in the image, but also on resolution since in order to differentiate between fine variations in tone it is necessary to magnify the image considerably. A pixel can only ever be a single color. Since a digital image is composed of square pixels, if a color change comes in the middle of a pixel, the software governing the image will make a decision on what the color is for the whole pixel based on the majority of the area covered by a single color. If two colors are equally represented in a pixel, the software will choose a color mid-

way between the two (so a half black-half white pixel will be represented as mid-gray). The higher the resolution of the capture device, the more pixels are crammed into the same area of the original, which means that there is a wider gamut of colors available in the final image.

Each color that appears to represent 'text' (as opposed to dirt) is selected, and that selection can be highlighted with the color selection tool. The best way to highlight is to use a color like bright green. This lurid color means that nobody looking at the restoration could possibly mistake it for the original. By distancing the color from the tone of the background, however, it stands out more than it would if it were a variation on the existing tones, such as black or a darker brown.

Color selection can be rather random and requires patience: the more individual selections made, the better the result can be. There is a great deal that the software can see that our eyes cannot. Finding the right color to select is often a question of trial and error. More often than not a good result can be a matter of luck, though experience and practice can add a bit of an edge to luck.

Photoshop includes a large number of filters, most of which would never be used in this sort of work. However, some quite unlikely-seeming ones can be useful and may provide unexpected insights. The third process that is worth applying (after "level adjust" and "color select") is the "high-pass filter", though some images may respond only with the filter and nothing else. A high-pass filter adjusts the picture based on frequency changes – or color variants. Because filters can't automatically be added as layers the way level adjusts can, the best way of using one is to create a new layer duplicating the master image, then apply the filter to that new layer. The way this new layer interacts with those below it can be changed, which is helpful not for what the filter does itself but for the way a filtered version of the master image can alter the underlying master image.

ETHICS

We are faced with an ethical dilemma in digitally editing images of manuscripts. Firstly, we have to be sure that we do not manufacture evidence to suit our expectations of an outcome; secondly, we have to be extremely careful both to avoid misrepresentation of the original source and also to ensure that our restoration work can be examined alongside the original, so that those using our restorations can make their own decisions about the veracity of the outcome we have created.

Where is the line between restoration and editing drawn? I believe that it should be drawn the minute we adjust an image at all. Assuming the image is an accurate representation of the original, then anything done to change that is editing. Where the line is drawn between simple enhancement to restore readability and something more questionable must depend on common sense. Adjusting an image for private use is one matter, but if that image is then presented to others to support an argument, the audience must be given the chance to evaluate the work for themselves and not to have to take on trust what has been found. The best way of ensuring this is to publish restorations alongside the original image in the digital medium, since printing two copies of every image – particularly in color – in a book or journal is still expensive.

Some heavily edited images were published in an appendix to the facsimile of the manuscript commonly known as Bologna Q15, dealing with severe burn-through on several paper leaves.[4] These were not digital restorations. They were edited images, using a cloning process, appropriately nicknamed 'fakesimiles', and required quite a heavy editorial hand. About 90% of the restoration is justifiable: burn-through that clearly does not belong on the face of the leaf is hidden by covering it with cloned material from a clean page (or part of the same page). For the remaining 10% the editor has decided from context or from a knowledge of the style of the scribe or period what the missing notes must be, and the damage has been restored to show that result.

Having insisted that only the best quality of color image is possible for restoration purposes it would be cavalier not to look, at least briefly, at the situation where only a poor surrogate survives

[4] Bologna, Museo Internazionale e Biblioteca della Musica, olim Civico museo bibliografico musicale, ms. Q15. See Margaret Bent, *Bologna Q15: The Making and Remaking of a Musical Manuscript. Introductory Study and Facsimile Edition*, 2 vols (Lucca, 2009).

and thus is the only material available for restoration. A black and white image is shown below that in theory should not be of any use for restoration. The original document is a pastedown that has not been lifted from the binding boards. In order to try and read what was on the reverse (stuck-down) face, the leaf was treated with a chemical that made it go partially transparent for a short time. The leaf was then photographed, but only in black-and-white. This process will never be repeated, so the only record we have is a grayscale glossy photo. (figure 2)

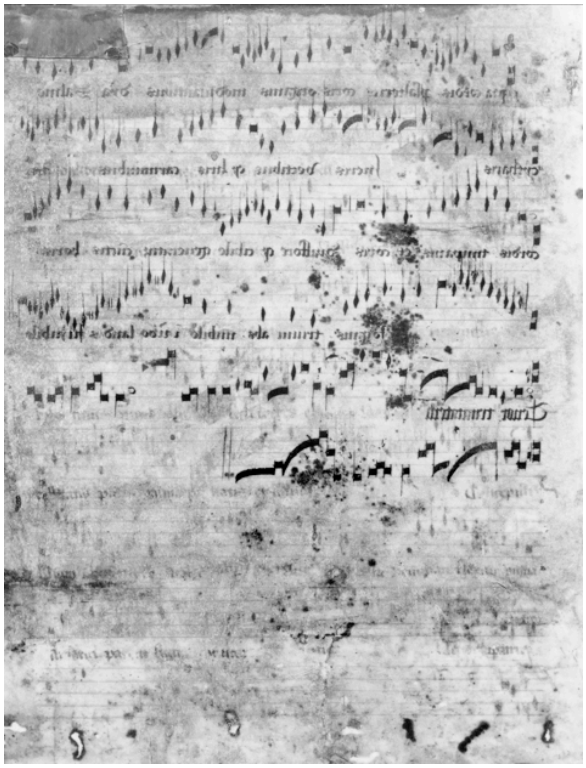

Figure 2a. Harvard, Houghton Library Typ. 122, rear pastedown, original grayscale.

Having removed in stages what was clearly on the visible (upper) surface, the image was converted to RGB color, mirror-flipped so that what appeared in reverse from the back face was the right way around, and saturation techniques were used to change the color of shapes that were apparently notes on the reverse: different colors represent different levels of certainty on the part of the editor. (figure 2b)

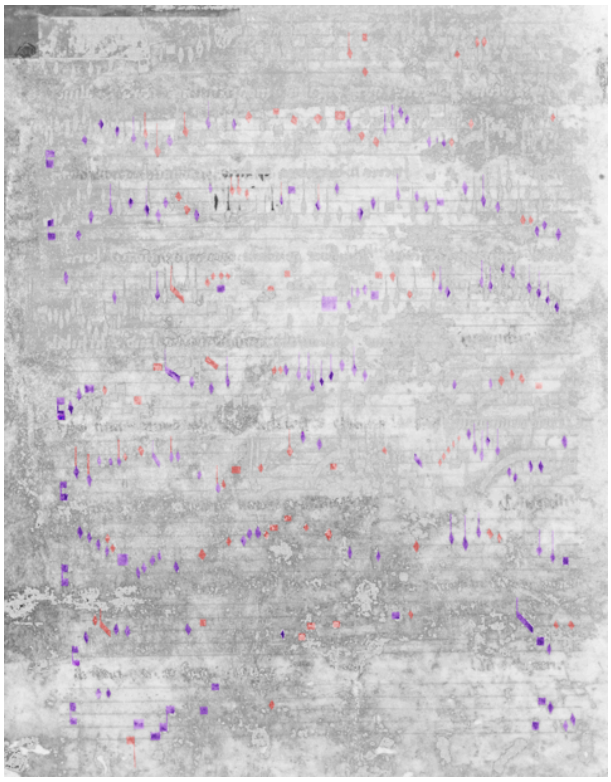

Figure 2b: Harvard, Houghton Library Typ. 122, rear pastedown, edited version.

This process was, however, unquestionably editorial at the very earliest stages, so again these results should not be called a restoration. It is clear that there is some confusion about the reading, but by presenting this version alongside the original, the user can make assisted decisions about the content.

ULTRAVIOLET, INFRA-RED AND INTERFERENCE BANDPASS IMAGING

In the digital medium, as in analog photography, there are even more specialized techniques than simple RGB imaging. There are few studios in libraries that will undertake this sort of work as it requires a specially-built camera without the standard built-in IR filter. However, even if it is possible to obtain images of this sort, the basic color image – if it was created properly – often yields more results in restoration. Ultraviolet (UV), photography only works if there is something on the page that will fluoresce under that light. Infrared photography (IR) is similarly quite limited and depends on a particular type of damage such as heat or the sort of massive deterioration suffered by things like Roman papyri or the Dead Sea Scrolls. Interference bandpass imaging uses a filter to exclude all but limited bandwidths of light – usually 40-50 nanometer segments – and records what is visible on the document in that spectrum. More interesting can be the use of raking light, but again it is unusual to find a studio with the time or interest to try any of these options.

In conclusion, for effective digital restoration, the best starting point is as near a perfect color digital image as possible. With this in mind, the following specifications are provided with the suggestion that any order for digital images should require the supplier to meet the following basic requirements:

File Format	24-bit RGB TIF (JPEG or other compressed formats should *not* be used at any point during the workflow).
Size	Minimum 400 dpi at actual size. 600 dpi (or higher) preferred.
Unsharp mask	OFF/Not Used.
Color	White and black points adjusted pre-capture by 5-8 points to avoid color clipping.
Gray balance/exposure	Gray balance and exposure calibrated using graduated grayscale.

Lighting	Even distribution of light across the whole area to be photographed with no darker or lighter areas.
Dublin Core Metadata	Embedded at capture in 'file info'. This should preferably include information about the capture equipment and software in addition to Dublin Core data (*optional if not included in capture software tools).
Color profile	Correct profile for capture device embedded at capture.
Color scale	Industry-standard scale photographed alongside each leaf as part of the image, preferably with both color and graduated grayscale targets as in the Kodak Q13 package. Gretag-Macbeth targets are equally acceptable (Graduated grayscale minimum requirement). The color scale must not touch the document; thus if the images are later required for facsimile publication they can be easily cropped to remove the scale.
Size scale	Photographed alongside each leaf as part of the image; cm and/or inch.
Focus	All elements sharp at 100% magnification on screen.
Post-processing	No post-processing whatever.