Livingstone Briefing: The science behind the project

What was achieved on this project?

David Livingstone's original handwriting was completely recovered from the original 1871 Field Diary using a custom multispectral imaging system and custom software developed by a team of four U.S. scientists and engineers. Over time, the handwriting had faded and in many places was not legible. There are several particularly difficult pages; some where Livingstone wrote over the top of pages of the *London Standard* and a few where the ink bled or seeped through the paper. In both of these cases, the unwanted text significantly reduces the legibility of Livingstone's handwriting. Over the course of this project, two new digital algorithms were developed to process the multispectral images – enhancing the handwriting while suppressing the unwanted text. As a result of the application of these new methods, David Livingstone's handwriting has been completely recovered and the entire diary is now available for scholars to study.

In cooperation with the U.S. Library of Congress, the team also conducted an analysis of the inks used in writing the diary. The results of that analysis and the conclusions drawn from them are described in a separate briefing.

Where does the imaging technology for the Livingstone project originate?

The imaging technology used on the Livingstone project was developed over the last decade by the scientific imaging team. This technology was created to recover illegible writing from important cultural heritage artefacts. The team was originally formed to recover the writings of Archimedes in the Archimedes Palimpsest (www.archimedespalimpsest.org). Since then, the team has imaged a Syriac palimpsest believed to be an early medical work by Galen, and is currently embarking on a several year-long project to image a range of palimpsests at St. Catherine's Monastery in Egypt. In addition to parchment manuscripts, the team has worked with the U.S. Library of Congress in spectral imaging studies of paper documents, including the Waldseemüller 1507 World Map (the first to use the term 'America' and show the Western Hemisphere) and the Library's 'Top Treasures' – documents by the United States' Founding Fathers, including drafts of the Declaration of Independence and Gettysburg Address**,** James Madison's Papers and L'Enfant's Plan of Washington DC, as well as treasures from around the world, including Armenian, Chinese and other documents.

How was the imaging of the diary accomplished?

In June 2010, the scientific team brought the imaging equipment from the U.S. to Scotland to image the diary. This imaging hardware included a large format camera, lighting panels, and various computer systems and software. The integrated imaging system is reasonably portable and has been applied to many manuscripts, starting with the Archimedes Palimpsest.

Over the course of a week and a half, both sides of the many pages of the diary were imaged one at a time. Photographed in a dark room, each page was illuminated with LEDs (Light Emitting Diodes) from custom light panels designed and constructed by team member Dr.

William Christens-Barry. Each LED emits light of a very specific color and the set of LEDs covers the visible range of light from blue to red – extending a little beyond the visible region into the ultraviolet and the near infrared. During each illumination, an image is taken using a 39 megapixel camera provided by Kenneth Boydston from MegaVision. Information about each capture (called metadata) is stored within each image, for use by imaging scientists in post-processing the images. The digital images were captured and stored on external drives and brought back to the U.S. for post-processing.

What are multispectral images and what do they reveal?

For each of the over 200 pages of the diary, a multispectral data cube was captured. The data cube consists of 12 to 16 images of a single page taken sequentially over the course of a few minutes. During these exposures, the page rests on a copy stand without any movement and is illuminated sequentially by LEDs of different colors from the light panels. The LEDs cover the spectral range of 365nm to 1050nm – ranging from ultraviolet (UV), through the visible spectra and into the near infrared (IR). Some of the LEDs are used at raking angles, close to the page, to reveal changes in surface texture and topographic features in the document. As each leaf is illuminated sequentially by the sixteen different wavelengths of light, the monochrome large format digital camera captures a separate image for each wavelength of light, or color of illumination.

The resulting images form an image cube of perfectly registered images – where any given point on one image will line up with the same point on all the other images. This data set of multiple images of a page, taken in different colors of light, allows the variation in color of any single character or word on the page to be studied in minute detail. These small variations in detail were used to separate the handwriting from the printed text in the post-processing stage.

How does the post-processing reveal Livingstone's handwriting?

David Livingstone's handwriting could be recovered from the diary, because the inks and paper respond differently to different colors of light. The ink of the handwriting is much darker against the paper when illuminated under blue light, than it is under red. In fact, in the color region beyond red light, the infrared region, much of the handwriting disappears completely. On the other hand, the contrast of the printed text from the newspaper does not change at all anywhere across the whole color range used in the multispectral image cube. As a result, in the infrared region, where our eyes cannot see the light at all, the camera was able to capture images where only the printed newspaper text is present.

This difference in the ink response was exploited to separate the handwriting from the printed text, which was obscuring Livingstone's handwriting making it very difficult to read. One of the team members, Dr. Keith Knox, first observed that the printed text did not vary with the color of the illumination. This led to the creation of a new algorithm for suppressing the printed text. With some simple experimentation, it was discovered that a simple ratio, i.e. dividing the image in visible light (red, green, blue) by the image taken with infrared, significantly suppressed the printed text – revealing Livingstone's handwriting.

In other words, by comparing an image that contains only the printed text with an image that contains both the printed text and the handwriting, the printed text was almost completely suppressed. This spectral ratio method was applied to the complete set of images and it provided the bulk of the images that the scholars used to transcribe the diary.

Another method, which was originally developed for the Archimedes Palimpsest, was also able to separate out the handwriting from the printed text. This method is based on a well-known algorithm called PCA, or principal component analysis. PCA analyzes the variation in the multispectral image and rotates the axes of the image to align with the axes of maximum data variation. Using the PCA algorithm, Dr. Roger Easton, Jr. created a method to separate out the handwriting from the printed text. In his method, a color image is made from the first few significant principal components and the hue axes of the color image are rotated until the printed text disappears and the handwriting appears in the images. This last step requires a human-in-the-loop to choose the proper hue angles that reveal the handwriting. As a result, this method was applied in only a few cases. Dr. Adrian Wisnicki reports that this method did help in several cases where the handwriting revealed by the spectral ratio method was still difficult to read.

The second innovative algorithm that was created for the Livingstone project was applied to the problem of show through. There were only six pages that had this problem. On these pages, the ink with which David Livingstone wrote seeped through the paper showing up on the opposite side from which it was written. The writing from the opposite side shows up as horizontally reversed writing – making the handwriting on that side very difficult to read. In studying this problem, Dr. Knox observed that the handwriting on the side on which it was written had a much higher contrast than its reversed image from the other side. Since both sides of the page were scanned, it was a simple matter to reverse the image from the opposite side, align it with the scan of this side of the paper, giving two scans of the same jumbled handwritings. In these two carefully aligned images, the handwriting from one side is dominant in one image, while the writing from the other side is dominant in the other. By joining these two images together into a single color image, the two handwritings appear in different colors. In the resultant processed image, the handwriting from one side appears in blue and the writing from the other side appears in light yellowish-green. The eye can easily distinguish these two colors and the scholars were able to read the handwriting directly from these processed images.

LEDs can be bright. Can these lights damage a manuscript?

No, illuminating the manuscript with light from LEDs does not damage it. The imaging team has developed this advanced image capture system to maximize the information recorded with minimum impact on the historical documents. An important advantage of this system is that the LEDs do not generate heat, which can damage the fragile pages. Conservators and preservation scientists at the Walters Art Museum and the U.S. Library of Congress have studied the imaging process and concluded that the low levels of UV light used during imaging will not damage the paper or ink. They have concluded that the use of LED lighting is safer than regular incandescent lights and a filtered camera, since the LED's emit no heat,

and therefore help minimize changes in relative humidity that can damage works on paper or parchment.

How big is the Livingstone data and how was all that data managed?

The raw data collection of the over 200 individual leaves of the Livingston 1871 Field Diary consumes about 300 gigabytes of data space. Following the image processing, the images were collected into a complete package of images with documentation and full metadata. The archive data set was based on the archive and metadata model used for the Archimedes Palimpsest, which is documented in the Archimedes Palimpsest Metadata Standards (http://www.archimedespalimpsest.org/programmanage_documents.html). The result is a completely self-documenting and autonomous data set.

Thanks to a grant and additional support from the U.S. National Endowment for the Humanities (http://www.neh.gov/), and also a grant from the British Academy (http://www.britac.ac.uk), the archive has now been published through a collaborative venture between the UCLA Digital Library Program in Los Angeles and Livingstone Online, the leading internet resources for Livingstone's primary manuscripts. This arrangement was facilitated by *The Early Manuscripts Electronic Library*. The project makes all of the images and transcriptions available to the public, with free access, under a Creative Commons License. From November 1 2011 the key websites are as follows:

Livingstone's 1871 Field Diary: http://livingstone.library.ucla.edu/1871diary/

Livingstone Spectral Image Archive: http://livingstone.library.ucla.edu/livingstone_archive

1 November 2011