Digitizing the Dream: The King Center Imaging Project-Case Study

William Scott Geffert; ImagingEtc; Glen Rock, NJ / USA
Anand Sethupathy; JPMorgan Chase Technology for Social Good; New York ,NY / USA

Abstract

The King Center Imaging Project is collaboration between JPMorgan Chase and The King Center for Nonviolent Social Change to archive, preserve and share Dr. King's works. JPMorgan Chase, through its Technology for Social Good program, has committed its technology expertise to help digitize over 100,000 of the more than one million pieces of history related to Dr. King. This innovative project has resulted in a global educational website based on a comprehensive collection of Dr. King's papers.

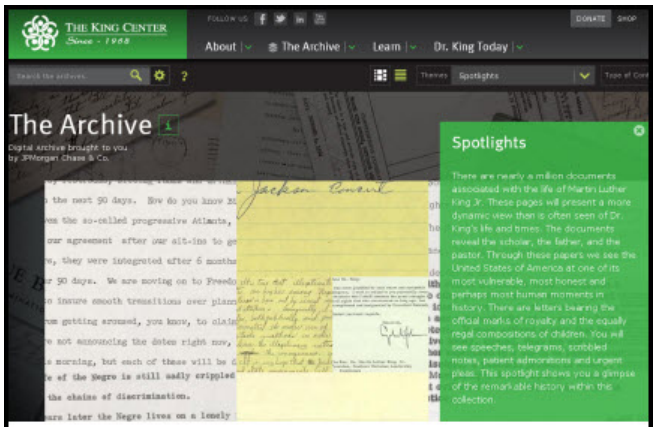
This paper highlights the steps taken from the project's inception in April 2011 to the global launch of the web site in January 2012 to coincide with the anniversary of Dr. King's birthday. The intent of this document is to highlight the tools and technology involved, and to share the experiences of a large digitization archive and web initiative. Anand Sethupathy, the IT lead for the project from JPMorgan Chase Technology for Social Good program has provided insights into the IT infrastructure to help give a more complete picture of this effort.

The IS&T Connection

On my way to the 2011 IS&T Archiving conference in Salt Lake City, Utah I received a phone call from Ali Marano regarding a digitization project. Ali explained how JPMorgan Chase's Technology for Social Good program (that she heads up) had set out to help digitize The King Center Library collection. While the goals that Ali laid out sounded incredibly ambitious, she had a refreshingly clear sense of purpose. From the moment I landed in Salt Lake and at every spare moment during the IS&T conference I was on the phone with Ali laying out various scenarios. Being at the IS&T conference was opportune because I was literally able to confer with colleagues and absorb information from the various presentations. From the outset, the IS&T community and The King Center Imaging Project have been closely intertwined. As I have mentioned before, we are all in the same boat when it comes to digitizing cultural heritage. Whether you are from a local library or one of the world's largest corporations, in the end, it is the shared experiences that drive innovation.

The Plan

The initial project goal was to digitize the entire collection of Dr. King's papers (over 100,000 letters, telegrams, newspaper clippings and handwritten notes) which are considered the core of the collection, and to have a substantial quantity of the material indexed, tagged and available on a new King Center web site by January 2012. The work was to take place onsite at The King Center library that had no existing digital imaging infrastructure and limited computing infrastructure.



The King Center's new website and digital archive was launched on MLK Day (January 16th, 2012). It was visited by over 150,000 people on it's launch day.

The Strategy

After identifying potential third-party vendors, a systems integrator (Micro Strategies, Inc.) specializing in document management solutions and my company ImagingEtc were selected to lead up the respective areas of the project. Ali Marano and Tina Shah of JPMorgan Chase held an initial strategy session on May 23rd at their New York headquarters. All aspects of the project workflow were laid out in one intensive meeting that shaped the core imaging strategy. In this meeting the decisions merits of cameras versus scanners were discussed as well as necessary workstations, image protocols, estimates for storage, capture standards as well as the lab layout and even the type of chairs, electrical service, networking and lighting requirements. My first visit to The King Center in Atlanta was on June 3rd.

The ambitious goal was to have the lab operational by just after the July 4th holiday as staff were being interviewed for positions and the lab space needed to be constructed including servers, network, and security systems. The decision was made to utilize digital cameras as opposed to scanners primarily because the camera-based solution offered complete flexibility in terms of addressing the wide range of document types and sizes. There was also interest in considering the ability to re-purpose equipment over time. While the team had a number of equipment quotes in hand, the decision was made to build the solutions from the ground up leveraging best-of-class tools from multiple sources and using only tools that have proven to meet or exceed current capture protocols. Originally five capture workstations were being

©2012 Society for Imaging Science and Technology

discussed, but after more consideration the decision was made to focus on three workstations. Looking back, this strategy was one of the more successful decisions as the systems performed beyond our expectations.

Another early decision was to leverage open source technology and objective imaging standards. The Alfresco document management system was selected as the platform due to the vendor's expertise associated with creating the necessary document organization structure from a collection of materials that had little existing indexing.

The Collection

The King Center Archive is the cumulative body of materials collected by Coretta Scott King up to and beyond the assassination of Dr. King. The work is divided across several different collections comprised of documents, telegrams, handwritten notes, posters, photographs, drawings, pamphlets and books. The King Papers Collection is considered the heart of the Archive as it includes the drafts of important speeches and the direct correspondence between Dr. King and the wider community. The King Papers Collection is housed in boxes each containing multiple folders packed with documents. The Boxes and folders had minimal information on labels and there was no computerized system at The King Center where this information was housed. The primary source of information about this collection resides with the staff of The King Center many of whom were directly involved with the center from it's inception in the 70's as well as the civil rights movement. The digitization initiative was as much about providing a long-term indexing structure as it was about imaging the documents themselves.

The Lab

A portion of The King Center Library reading room was to be reconfigured to be a state of the art digitization lab. The lab and equipment were designed to be as flexible as possible as The King Center may eventually desire to return the room to its original state.

While equipment was being specified and ordered, temporary walls, networking infrastructure, local storage, security systems and related infrastructure were being specified and constructed. The construction also included shoring up the HVAC and electrical systems at The King Center Library to support the anticipated computing and imaging infrastructure.

The plan was to create a flexible workspace with particular attention to ergonomics to provide an environment that would be safe for the original documents and comfortable for the staff. To this end, all workstations were ergonomic adjustable height surfaces, the walls were painted neutral gray, and the overhead lighting was outfitted with high CRI 5000k lamps. While these considerations are often overlooked, these decisions had a direct impact on productivity-especially during the indexing phase where staff needed to read each and every document to create abstract summaries. A nearby storage room was modified to accommodate the servers and networking hardware. The space was built out while the equipment was being ordered.

The Staff

How do you find well-qualified individuals to run highly technical equipment employing the latest standards-based practices? As with most aspects of this project, the JPMorgan Chase team took this challenge in stride by quickly identifying the skills required for the successful execution of this project. Both Micro Strategies and I agreed that the paramount factors beyond the expected criteria were interest and enthusiasm.

The staffing to execute the project was comprised of a mix of resources. JPMorgan Chase's staff (headed by Ali Marano) oversaw the overall operations. Vendors such as myself and Micro Strategies provided subject matter expertise. The actual execution of the imaging and indexing of the documents was mostly completed by an onsite team of 40+ individual contracted to JPMorgan Chase to work directly at The King Center Library.

The onsite staff was primarily comprised of students from the local universities of Morehouse and Spellman Colleges, Kennesaw State and Emory University, as well as members of the Veteran's Curation Program (a military program designed to train veterans in archival technology skills). The onsite staff at the outset of the project totaled just over 40 individuals working across two six-hour shifts five days a week. The staff was nearly equally distributed between the 2 shifts and the project was in operation from 7:30AM to 7PM. The onsite staff was managed by Beverly Dabney, a full-time JPMorgan Chase staff member who directly supervised the team.

Initially the onsite staff was trained on specific tasks such as document preparation or imaging. As the back-end systems came online, portions of the staff moved towards document reconstruction and indexing. Staff members were cross-trained on multiple skills and were able to move between tasks as needed. At the same time, certain team members naturally gravitated to particular tasks or excelled in specific areas. Some of these people eventually came to own specific components of the overall process.

During the equipment specification stage scanner vendors gave dire warnings about the ability of inexperienced staff to handle running a digital camera. While training was necessary, the staff not only became proficient in a short order of time, certain staff members were able to modify and improve the workflow as they encountered a variety of challenging documents.

Beyond this core onsite team there were also over a hundred volunteer employees from JPMorgan Chase as well as leading experts from the civil rights community, including prominent Kingian Scholars. These additional team members leveraged secure remote connections to the document management system as part of the meta-tagging and indexing effort which is by far the most involved in terms of time.

Digitization Specifications

Initial productivity estimates were targeted at a very conservative 600 captures/day per workstation. This was based upon known results of various institutions using digital cameras and unknowns about early audits of the physical collection in terms of document types and sizes. The actual productivity proved to be much higher.

In order to achieve high-productivity while still insuring quality and consistency the workflow was configured around the Metamorfoze Preservation imaging guidelines. Tools and targets of the FADGI federal agencies digitization guidelines initiative were also employed to insure that the workflow satisfied both of the protocols that were on track to be unified under current ISO

efforts. For example: Both the UTT charts AND Golden Thread charts were utilized in the creation and ongoing workflow.

The specific imaging parameters are 400PPI Tiff format (verified via UTT/Golden Thread Analysis) 8Bits per pixel and eciRGBv2 color encoding. Three Hasselblad H4D 50MS cameras were utilized primarily because the multi-shot option eliminates potential color moiré issues with halftoned images. The cameras were profiled using BasICColor Input software and the X Rite DCSG chart. The IQA Color Module from Image Engineering was utilized to verify each camera station on a daily basis. If the validation fell outside the Metamorfoze tolerances for color and tone, the production would not continue until the camera passed this check. Because the camera stations were verified and fixed at a 400PPI field, the UTT and Golden Thread were not necessary to be used on a daily basis, but were used in the event that the camera station was reconfigured for oversized works.

To further streamline the ongoing QC, a special grayscale chart was created for the project by the Munsell Color division of X Rite comprised of the exact same semi-gloss patch material as used in the DCSG color chart. This way, the values of the grayscale are perfectly compatible with the DCSG chart that the workflow is validated to. This chart was included in every capture in addition to the Golden Thread small object level target.

I could not imagine a better case study to illustrate the benefits of standardized objective capture practices. Out of 40 team members only one had a formal photographic background and none had ever used technical charts, ICC profiling or validation software, yet in a matter of days they were imaging and running the daily QC checks. During the course of 6 months there were only a handful of times the production was held up over failed validation and in each case while the technical issues were minor and easily resolved.

All additional MacPro workstations in the lab were configured with NEC Spectraview displays calibrated (L*, D50,120cd/m2) using BasICColor Display software and the BasICColor Discus hardware. Between the L* eciRGBv2 encoding and L* display calibration, this project was purpose built to be media neutral as opposed to encoding for a particular output. This hardware was supported by Fotocare in New York.

Productivity Enhancements

With the solid foundations of capture in place, there were several key strategies to reach the desired productivity. The approach was built upon the concept of addressing every potential bottleneck in the imaging pipeline via extensive up front planning and testing.

For example: by fixing imaging stations to a verified 400PPI and using proactive daily system verification, we eliminated the need to adjust focus or camera height. If the artwork fit within the camera's field of view it was imaged consistently. It is important to note that in order to guarantee this level of stability; we needed to outfit the cameras with special precision Lens/Bellows units from Michael Ulsaker and heavy-duty copy stands from TTI to eliminate the possibility of cameras drifting out of specification. Using this large field of view presented challenges related to positioning original documents. While we could have used rulers or other physical framing guides, these would either have to be in the field of view or would have to be removed before each capture. The solution to this challenge was to employ specially designed low power cross-hair lasers. Positioned near the camera lens and aimed down to the copy surface the lasers were aligned to create a pre-defined crop area. As the lasers are extremely low powered, the line visible to the eye is overpowered by the electronic flash illumination rendering it invisible to the camera sensor. In the end, the time spent up front carefully measuring out each station and building specific templates significantly improved the capture throughput.

Another productivity enhancement was the use of bar codes and bar coded separator sheets that were inserted with the documents to identify boxes, folders, multi page documents, duplicate documents etc. Imaging operators would use a low cost USB bar code reader to create the file naming sequences, and the document management software took care of interpreting the barcode data to recreate all document relationship logic. In addition, the extensive use of barcodes throughout the workflow allowed the tracking of the locations and workflow stage of every item as it moved from storage to imaging and back to storage.

Document Workflow

Documents moved through a process which started with document preparation and ended with the documents being posted to the website. Quality assurance checks had to be embedded into every step of the overall process.

Documents went through a workflow which culminated in posting to the website

Document Preparation

The first stage of the workflow after removing each document from the shelf is document preparation. In this process, the already barcoded box and barcoded folders are opened and a staff team inspects each document, and performs several preparatory steps including removal of staples, checking the physical state of the document and if the document is able to be digitized. Another important part of the process is the insertion of separator sheets. The separator sheets tell the document management system if the document is a new document, an associated document, a copy of a document, a related document (such as an envelope), an excluded document (documents with personal checking accounts etc.), and also the beginning and end of each box, folder and document in the box. This upfront work allows the document management system

©2012 Society for Imaging Science and Technology

to create important relationships automatically upon ingestion. For example: the system knows via bar code separator sheets when to generate a multi-page PDF of a multi-page letter. Of course, after digitization, these separator sheets are removed as well as the associated separator sheet images. The downsides of this process are that it is labor intensive and needs to be performed carefully. It also generates significant overhead as each barcode needs to be digitized along with the original documents. If a folder contains 100 unique single page documents, 50% of the digital captures would be of separator sheets. As the imaging program got underway, we began cropping these images down via capture presets to just the area of the barcode to reduce capture and processing time. The use of different colored separator sheets helped the imaging team with QC as they could easily identify the red "End Folder" separator sheets and the other color-coded sheets in the image thumbnail views.

Separator sheets were inserted at the doc prep stage to help automate later stages

Imaging

After document preparation, the images are transferred to an imaging workstation where the operators would work individually or in teams of two to place and capture each document. While teams of two were not absolutely necessary for digitization, it allowed the staff to better manage necessary break time. Imagers would switch roles between camera operation and QC. The extensive use of barcodes definitely helped productivity and reduced file naming errors and also helped track the original materials as they moved through the workflow.

The net result of the imaging strategy is that after a brief ramp up and training period, the team was able to digitize up to 6,000 images/day (2,000 captures per station across the morning and afternoon shifts). While this was the peak productivity with similar sized documents, the average productivity was closer to 4,000-5,000 captures/day, still well beyond our very conservative projections. It is important to note that while there were some bound materials and oversized documents in the collection, the majority of the original material was within legal size, contributing to the high productivity. If there were bound materials involved, the productivity would obviously be much lower. When odd-sized documents were encountered we found that the staff enjoyed the occasional opportunity to adapt to the material. Of course this is the true benefit of digital cameras over scanners.

Import to Document Management System and Box Reconstruction

After imaging files are moved on a daily basis to the central server from local workstations an import process is triggered. In this process, the Alfresco system analyzes the incoming data and parses the files based upon the barcode information. It also rectifies the physical box, and folder barcodes to the capture directories. The process is simple and logical, but in order to work effectively there is a significant level of document management system experience required. The engineering, testing and deployment of this system was a major challenge for all involved given the timeframe. When successfully imported and verified, the separator sheets are removed from the collection and the original documents are placed into brand new archival folders and boxes. The boxes are then replaced to the archive where a new fire suppression system has been installed as part of the project.

Due to the very high production volume, an area that proved to be particularly challenging was that relatively simple errors in imaging like an operator failing to capture an image of a "new folder separator sheet" at the head of a folder would jam up the document management system import process as the entire folder would be rejected (because the system knows in advance the correct structure of folders). These errors would require the box to be manually audited for accuracy and proved to be a workflow bottleneck.

Indexing, Tagging

With documents now managed via the system, teams of staff are deployed to create the necessary related data beyond the foundation indexing of box, document, or folder relationships. For example: each physical folder was digitized because they include handwritten notes or typed labels. These images are opened and individually transcribed. This work was performed using ICC calibrated 30" displays that allowed a full 1:1 page view of the folders and documents.

Beyond the folder information, each handwritten document was read and transcribed into an abstract summary and important pre-defined keywords and other metadata were added. The work in this phase was triple checked across multiple layers of staff and volunteers from staff in the imaging lab, to secure remote web-based access by teams of scholars, to JPMorgan Chase volunteer employees worldwide. The monumental task of launching a vast collection to the web with meaningful free-form search capabilities required sheer manpower to achieve. Of course the work required a significant amount of cross-checking to insure accuracy.

Posting to the Website

After several layers of quality control in our indexing and tagging process, documents could be posted to the website. If a document had reached the end of the Indexing and Quality Control stages and was marked as "Display on Web", it would get imported into the website. At the outset of the project, we were trying to push all documents to the website; however, after seeing end users get overwhelmed by thousands of form letters, we decided to be more deliberate about what content was pushed to the website. A team of Kingian Scholars helped us define a set of protocols that identified interesting content. These guidelines were followed to identify the prioritization of documents pushed to the website. If The King Center ultimately decides that everything should be online, they would need to set the "Display on Web" attribute for all documents.

On MLK Day 2012, the site launched with several thousand documents. Up to several hundred new ones are added each day to the website. In order to ease end users into the archives, "themes" were created on the website, which represented curated collections

of documents that would be of greatest interest to casual visitors. A robust advanced search function allows scholars to find more specific artifacts they are interested in.

Storage Infrastructure

The King Center has a long-term vision to digitize all of the assets stored in its Archives which comprises over 1 million documents and over 4,000 hours of audio/video materials. Once fully digitized, The King Center Archives are expected to be on the order of 400TB of data. To meet the near and long term needs of The King Center a multi-tiered strategy was put in place. At the outset of the project, local servers and a NAS array were deployed onsite along with a tape backup solution. In addition to this local storage, a strategy was developed to mirror the archive to cloud storage provided by EMC & AT&T. The storage strategy can evolve over time as the ongoing maintenance of the digital archive requires.

Digital Derivatives

In almost every single cultural heritage and library effort I have seen that involves web deployment, sites struggle with the practice of inclusion of technical targets in master archival image files and the need for cleanly cropped images for the web. Many employ tedious manual derivative generation methods that may include a series of various sized image files such as thumbnails, previews, etc. These derivatives take up space and serve little long-term purpose other than to leave the Master image files untouched. The fact is that with today's wide range of potential delivery devices from HDTV's to Mobile Phones and Tablets, prepress, print-on-demand etc. there is no single derivative specification or color space that satisfies all potential image usage.

In order to achieve the goal of creating the ability to crop and transform images via logic within the document management system the JPMorgan Chase team, Micro Strategies and I discussed the various options to arrive at the following specification:

- A web based user interface for cropping images integrated into the existing Alfresco document profiling process (Metadata/QC)
- The ability to store crops as editable metadata tags
- Integrated ICC color transformations from the master image to the web derivative (eciRGBv2 to sRGB)
- Downsampling and format conversion (Tiff to PDF/JPEG etc.)
- Automated cropping and squaring
- Automated generation of web handoff (derivative) files

Micro Strategies was able to leverage ImageMagick (another open source tool) to create the web-based cropping solution. Fully automated cropping/squaring routines were tested successfully, thanks to the extremely precise capture protocols, but the decision was made to not build out this functionality in order to meet other time-sensitive project priorities. Palantir (the web development partner on the project) accesses the cropped 1MB handoff derivative files to generate all downstream web derivatives required to support the website design built using Drupal, another open source technology.

The successful integration of non-destructive metadata-driven cropping is one of the most valuable imaging related aspects of this project. This is not the first use of this workflow approach for archiving, but it is highly recommended because the use of external image editing applications (Adobe® Photoshop™) is completely avoided. With relatively minor effort, master image files can be programmatically rendered to any potential size, file format, crop treatment or color space.

In my opinion, the next logical application of this strategy is to move towards a raw file based archive integration. There is technically no reason why an archive could not be comprised of master DNG raw files where all downstream translations of the raw data could be metadata-driven routines applied within the management system. Consumer level tools like Adobe® Lightroom™ and other desktop imaging software tools already leverage this functionality, yet high-end DAM systems and document management systems have not matured to this level of integration. The DNG file format as it is used today is still a bit immature to use as the ONLY master asset, but in time this approach could become the ultimate archival solution. What we need is a greater focus on the standardization of how raw data is being created and utilized. (see my IS&T "DNG Dilemma" paper)

Public Facing Efforts

Development of the public-facing aspects of the project were tightly coordinated and developed in parallel with the digitization and document management effort. To raise awareness of the of The King Center Archive collection, JPMorgan Chase set out to create a formal outreach program that included the creation of a modular expo booth to tour around the country. The full expo booth incorporated an innovative "Dream Wall" element. Along with information about the Dr. King, The King Center and the Imaging Project, visitors are able to write a personal "My Dream Is" note. The note is digitized using a DSLR camera copy station, and is then placed by the visitor onto a large backlit plexiglass wall. By the end of each event, this wall is literally covered with dream cards. The Dream Wall is especially popular with families with children. The most interesting part is that these digitized dream cards are also permanently incorporated into The King Center archive and visitors can see them on the web as they are transcribed and indexed by geographic location.

The incorporation of new content from the general public into the collection helps establish The King Center archive as a vital part of the global community as opposed to a center only for formal academic research into nonviolent social change. Within months of the web site launch, schools worldwide started to leverage the collection via the online resources to allow students direct access to documents of the civil rights movement. The project has been the subject of many magazine and newspaper articles, and has been covered by local and national television news programs as well as via community events.

Lessons Learned

It is important to note that this project began with The King Center reaching out to JPMorgan Chase to help archive and share the collection, an effort that the Center had been trying to accomplish for many years.

It is not very often that a single archiving project encompasses almost every single aspect of the challenges facing today's archiving community in such a short time period. Beyond accomplishing the original project goals it is clear to all involved that bringing a large collection from storage shelves to the world at

©2012 Society for Imaging Science and Technology

large presents tremendous challenges.

Aside from scale, the challenges of archiving are the same for a large multinational bank or a small not for profit organization. While the tools and technology used to achieve the goals of this project may be unique, there are elements that can benefit any archive project:

- Utilization of standardized capture methods, chart-based validation and QC
- The use of ICC color management (with validation) across the entire workflow (including web derivatives)
- The use of L* based eciRGBv2 encoding and L* display calibration across the entire lab
- Attention to workstation ergonomics and proper lab illumination
- The use of digital cameras as opposed to scanners can be both flexible and highly productive
- Fixing cameras to chart-validated PPI can dramatically improve throughput when appropriate for the collection
- The use of bar codes along the workflow to help minimize errors and track of materials.
- Cross-training of staff to keep the workforce flexible, fresh, and to identify team leaders
- The use of 100% open source tools and modern cloud resources for a flexible, sustainable IT infrastructure
- The use of open-source tools to create web-based post-production interfaces and automated imaging-related tasks
- Incorporating public-facing efforts to encourage interaction and access
- Distributing metadata, and QC, and other project resources across secure web based workspaces.

It is important for me to take a moment to look beyond the necessary hardware and technical specifications of a project to consider the human side of the project and why archiving is so important. The opportunity to work directly with The King Center staff, the family of Dr King, the incredible team of students and military veterans, the JPMorgan Chase Technology for Social Good team, and the numerous vendors has been an incredibly powerful experience. Everyone involved in this project has shared a singular sense of teamwork and accomplishment. These feelings are most apparent as I see the reactions of families as they visit the various Expo booth events that are often manned by the same people that have performed the archiving work.

This project began for me at the 2011 IS&T Archiving conference in Salt Lake, and has employed many of the best practices and protocols that have emerged through this community. Sharing this project experience at the 2012 Archiving conference in Copenhagen underscores the importance of this ongoing global dialog. Regardless of funding and scale, the true value of this and any other archiving effort is that people worldwide have been empowered to rediscover important windows to the past.

The King Center Archive URL: http://www.thekingcenter.org/archive

Author Biography

Scott Geffert is President of ImagingEtc Inc. www.imagingetc.com. A New Jersey based consulting firm specializing in digital imaging workflow and standards compliance. ImagingEtc Inc consults museums and corporations worldwide. Scott has been involved in photography since 1975 and has been active in digital imaging since 1984.

Anand Sethupathy works in JPMorgan Chase's Technology for Social Good group where he works to leverage the company's technology resources and expertise to assist NGO's and Social Good organizations around the world. Prior to his work at JPMorgan Chase, Anand worked in the nonprofit technology sector for 6 years and in the IT industry for over 10 years.