eArchives: Memory of the World Bank Project

by Arleen Cannata Seed and Jeanne Kramer-Smyth

The World Bank

I. Introduction

The World Bank's Access to Information Policy, updated in July of 2010, creates an environment for the opening of its records to the public[1]. The World Bank manages a vast Archive which is primarily paper-based and which has formerly been accessible only to researchers who physically visit the Archives in Washington, DC. In order to comply with the policy to make the holdings more accessible to the public, the Bank has embarked on an eArchive Project called the Memory of the World Bank Project.

The objective of the eArchive Project is to transform the Archives from a predominately paper-based collection of materials to a modern, pro-active and innovative archive which actively pushes content out to the public in a manner which engages and informs. This is being accomplished through the digitization of key archival collections, association of appropriate metadata and search tools, and posting of the records to the internet. Once the records are online, users can browse by leveraging finding aids to discover digitized materials; alternatively, they can use search tools to retrieve materials for which the original archival context has been preserved.

II. Methodology

In designing the project some overarching assumptions were made at the outset. First, it was essential to protect the original documents since they are unique and irreplaceable. Second, the digital images must preserve the original order of documents in their folder. Third, the image quality must capture the content of the records faithfully enough that their information is useful to the researcher. Finally, due to resource constraints, the documents would be digitized at the folder level with no further delineation or description of documents.

Once these assumptions were adopted, the next step was to create a *concept note* which explained the basic idea of the eArchive. The World Bank records are 'locked up' in an underground limestone mine and only accessible to the few researchers who visit the World Bank Archives in person. While these materials can be considered 'accessible', they are only accessible to researchers who have the time and money to travel to Washington, D.C. Additionally, because of the lengthy review process and limited space for researchers, a requestor might face a long queue before being scheduled to view the materials. This approach does not hold up against the spirit of the policy. Equally, the World Bank

[1] The basic premise of the Access to Information Policy is that all World Bank documents must be made public, with the exception of those documents which are restricted (the rules of which are found in the policy). See http://documents.worldbank.org/curated/en/2010/07/12368161/world-bank-policy-access-information.

supports transparency and good governance in its funding to beneficiary countries and opening up the Archives would be an excellent way for the World Bank to demonstrate its commitment to Open Data.[2]

The concept note was followed by a *business case* which was written to aid senior management's understanding of the idea and to rally support and funds. The business case outlined the level of effort required, the costs involved, and described the benefits to the institution should the project go ahead. The next step was the *proof of concept*. We wanted to be sure that we could digitize archival records, apply metadata, and associate them with current World Bank documents in a meaningful way. To do this, the group researched best practices of digitization from other organizations. The team undertook site visits to the Smithsonian Archives of American Art and the University of Maryland Digitization Program[3]. A mock up of our proposed approach to posting content online was reviewed by and approved by the Web Programming Office.

The Archives then tested various *scanning methods* to determine the most efficient and effective methodology for the digitization tasks. We scanned the same materials three ways: using the World Bank's copy center, using an in-house contractor in our own offices, and with our vendor for the Mine. We also prepared digitization guidelines so that we would be able to apply an even hand to the scanning, even if it were done at different times and by different people.

It was crucial to *prioritize certain key collections* for digitizing because the Archives' collection is extremely vast.[4] We knew we did not wish to digitize the whole mine[5] because of the expense and much of it may never be referenced. The prioritization was done using the following criteria: relevance to the current work of the World Bank, repeated requests by researchers, or topics which the Archives has judged to be exemplary of the new open agenda. Once we narrowed it down, we decided to start with a pilot on two small sets of records: Robert McNamara's records, former World Bank President, and records on Food Security. Both these topics are currently very much in demand by researchers.

The next task was to *review the records* against the Access to Information Policy to ensure that we followed the guidelines on withholding documents which the policy states must be restricted, and to create a pool of materials which could be made public.

Once a positive list of public records was confirmed, the team implemented a *scanning plan* which preserves both metadata and original order, and forms the basis for associating the records with other online offerings. Some of the issues which required consideration were the presentation layer, the organization of the records which metadata to display, and how finding aids should be prepared. Furthermore, we considered the need for a repository and its technical and size requirements. For the

[2] For more information on the World Bank's Open Data program, see http://data.worldbank.org/

[3] Site visit summaries: Archives of American Art: http://www.spellboundblog.com/2012/02/03/digitization-program-site-visit-archives-of-american-art/; University of Maryland: http://www.spellboundblog.com/2011/12/12/digitization-program-site-visit-university-of-maryland/

[4] If one were to stack all the boxes which the World Bank stores in the mine, they would be 7.5 times the height of Mount Everest.

[5] Our back of the envelope calculation put the price tag on digitizing the whole mine to be nearly USD 100 million. And that's just for digitizing, and doesn't include adding the metadata, setting up a repository, or the web sites.

McNamara papers, we attached the scanned images to our online finding aids which are on the external website of the Bank and store them on an internal serve. We contrasted this with a different approach for Food Security; we designed a website to link to an existing platform of information hosted by the World Bank and its partners. Once posted, we tested access and confirmed that the records were accessible online.

The last step was the preparation of a *marketing and awareness campaign* so that we could inform both World Bank internal staff and the general public that this electronic archive was available.

III. Resource Challenges

a. Human Resources

Staff in the World Bank Archives manage the contract with the vendor for the underground repository, use sophisticated software to run the logistics and warehousing of millions of boxes, accept transfers of boxes from over 100 country offices, prepare knowledge products and finding aids, provide records management advice to the entire organization, and manage the Access to Information implementation. On top of this, the eArchive project has placed additional burdens on the staff to review materials, scan the ones which could be made public, associate metadata, and store them such that they can be retrieved. With only one part-time electronic archivist, the team has leveraged consultants and contractors who were deployed for other information management work. We set up the scanner and the settings and kept a box of materials handy. Anyone with a few hours to spare was put to work scanning. Ultimately, having an eArchive should reduce the number of researchers who come into our offices, but the initial outlay of time is considerable.

b. Funding

Funds are needed for the scanning whether this is done in-house or contracted, and for equipment for both scanning and storage. Funds are also needed for design of the interfaces and the pre-automation phase, the final quality assurance of the images, and for associating the file with the correct fonds description or metadata. Additional budgeting considerations include: review for declassification against the Access to Information Policy requires 14 hours per linear feet; a condition assessment review[6] which requires additional staff effort, both before and after the digitization process; and, document preparation and reassembly for the paper records. Additional funds are needed for the awareness campaign. The Archives requested an initial USD100,000 to undertake a pilot; this pilot would be of one series of records, with the understanding that additional funds would be sought to scale up the project.

c. Policy

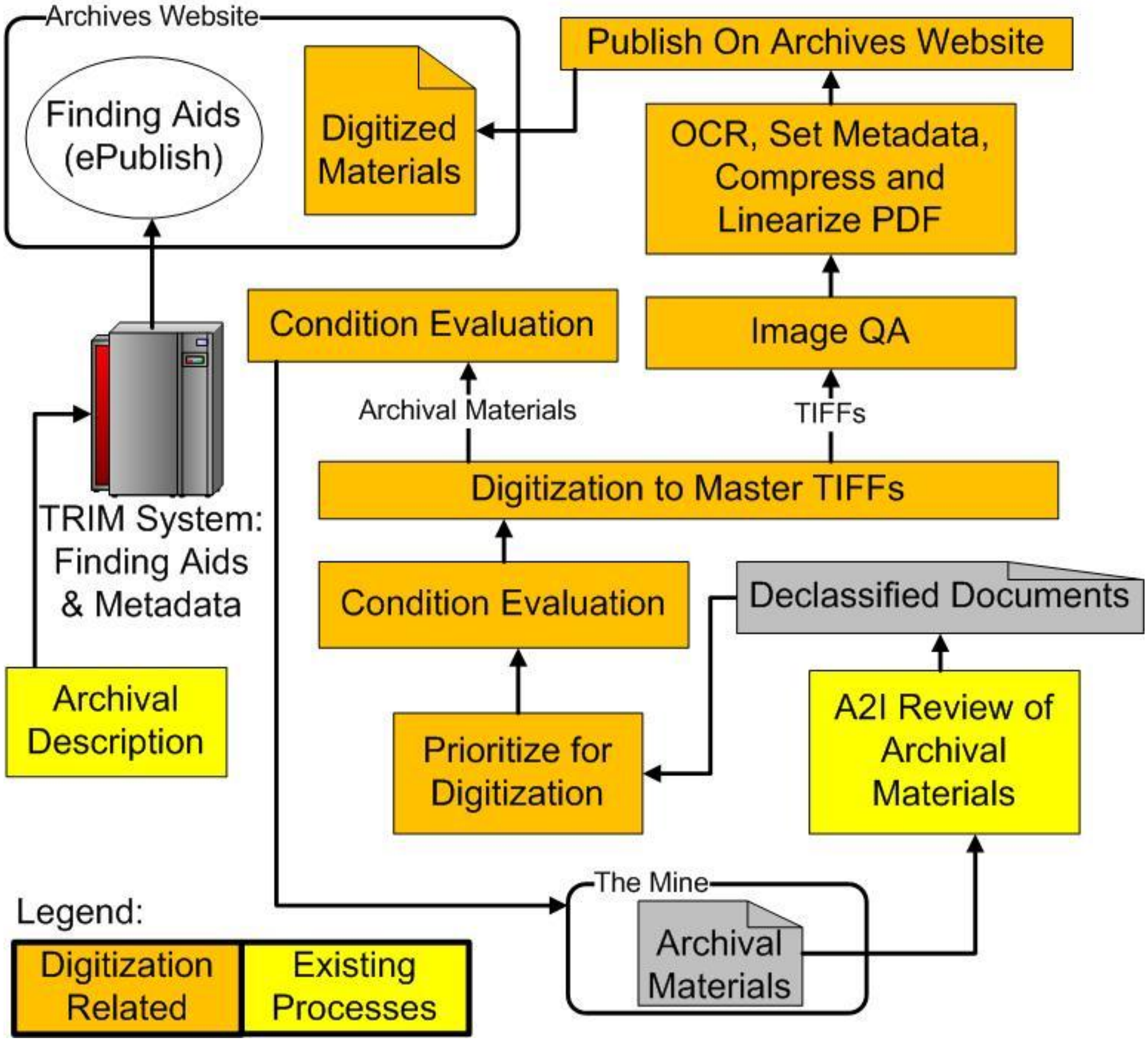
The archivists collaborated with the legal division at the World Bank to define a standard copyright and disclaimer clause which would be referred to from the cover sheet of each PDF of digitized materials.

[6] A condition assessment determines and describes the physical state of the materials.

Appropriate time for negotiations of this type must be budgeted into any timeline for publication of archival materials online.

d. Processes

The team prepared a complete process workflow to ensure that all the requisite steps would be covered and that the required resources would be available at the right time. This workflow is summarized in the chart below:



We considered leveraging the work of researchers who come to the Archives with handheld scanners or cameras and take digital images of the records. Sadly, this approach was not successful; the images were either not the quality we required, were incomplete or in the wrong format. Another innovative idea is to establish a service whereby we would scan records on demand for a researcher if they were not able to travel to DC. We could charge a nominal fee for the scanning, but it would be less expensive than having the researcher travel to Washington. Another option we are considering is to have researchers on retainer who could be hired by someone to do the research for them; in the process of making scans for their client, they could provide us with a copy.

IV. Evaluation of the Scanning Methods

In the course of our evaluation of the three scanning options described above (in-house archivists, copy center, vendor), the eArchives team performed extensive research into best practices of digitization. We consulted published guidelines on technical specifications, digital formats, metadata, and document handling.

A pilot was done to digitize five folders of archival materials, which included:

- Handwritten correspondence
- Typewritten materials on onion skin, both with and without watermarks
- A black and white photograph
- Diagrams and maps
- Mimeograph pages
- Documents with embossed stamps

The selected materials were scanned three ways: by the World Bank Copy Center (using their standard flatbed scanner and available staff familiar with standard scanning practices), by a member of the World Bank Group Archives existing staff (using an flatbed scanner already in possession of the Archives), and by the vendor at the secure underground repository (using their top of the line equipment and staff trained in digitization of archival materials).

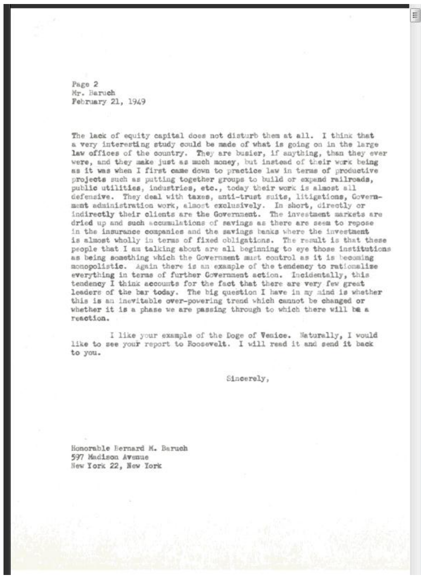
In each case, those performing the scans were provided with the same digitization requirements document. The archival materials were evaluated before and after each scan to ensure that the materials were returned in their original order and suffered no harm in the process.

After the digitization process was completed, we compared the final deliverables from each of the sources and considered all of the following:

- Quality of the images
- Accuracy of scans
- Adherence to the requirements
- Service and equipment availability
- Cost
- Workflow and QA processes in place
- OCR quality

The following figure provides details of the digitization sample from the pilot, showing the same page as scanned from each of the three digitization sources:

World Bank's Copy Center



Page 2
Mr. Baruch
February 21, 1949

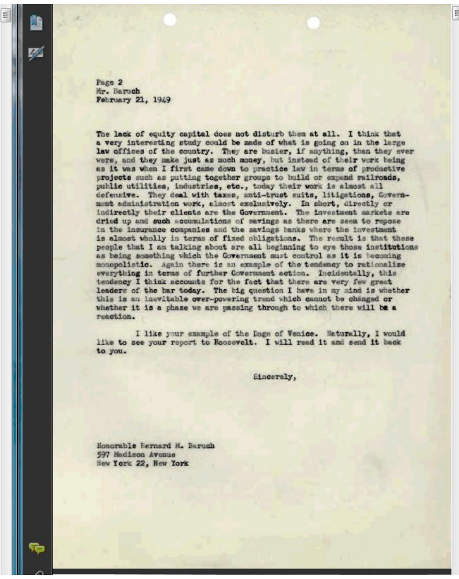
The lack of equity capital does not disturb them at all. I think that a very interesting study could be made of what is going on in the large law offices of the country. They are busier, if anything, than they ever were, and they make just as much money, but instead of their work being as it was when I first came down to practice law in terms of productive projects such as putting together groups to build or expand railroads, public utilities, industries, etc., today their work is almost all defensive. They deal with taxes, anti-trust suits, litigations, Government administration work, almost exclusively. In short, directly or indirectly their clients are the Government. The investment markets are dried up and such accumulations of savings as there are seem to repose in the insurance companies and the savings banks where the investment is almost wholly in terms of fixed obligations. The result is that these people that I am talking about are all beginning to eye those institutions as being something which the Government must control as it is becoming monopolistic. Again there is an example of the tendency to rationalize everything in terms of further Government action. Incidentally, this tendency I think accounts for the fact that there are very few great leaders of the bar today. The big question I have in my mind is whether this is an inevitable over-powering trend which cannot be changed or whether it is a phase we are passing through to which there will be a reaction.

I like your example of the Doge of Venice. Naturally, I would like to see your report to Roosevelt. I will read it and send it back to you.

Sincerely,

Honorable Bernard M. Baruch
597 Madison Avenue
New York 22, New York

World Bank Group Archives Staff



Page 2
Mr. Baruch
February 21, 1949

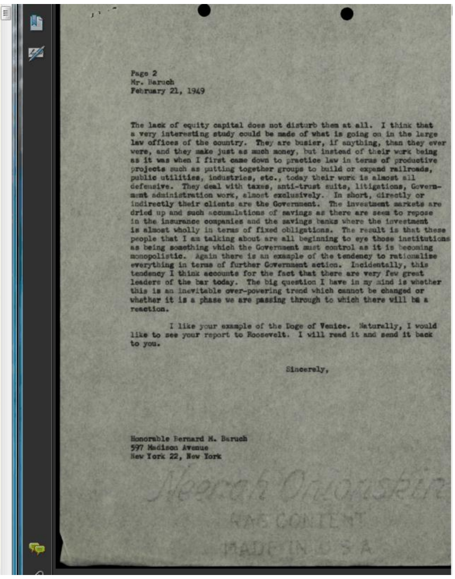
The lack of equity capital does not disturb them at all. I think that a very interesting study could be made of what is going on in the large law offices of the country. They are busier, if anything, than they ever were, and they make just as much money, but instead of their work being as it was when I first came down to practice law in terms of productive projects such as putting together groups to build or expand railroads, public utilities, industries, etc., today their work is almost all defensive. They deal with taxes, anti-trust suits, litigations, Government administration work, almost exclusively. In short, directly or indirectly their clients are the Government. The investment markets are dried up and such accumulations of savings as there are seem to repose in the insurance companies and the savings banks where the investment is almost wholly in terms of fixed obligations. The result is that these people that I am talking about are all beginning to eye those institutions as being something which the Government must control as it is becoming monopolistic. Again there is an example of the tendency to rationalize everything in terms of further Government action. Incidentally, this tendency I think accounts for the fact that there are very few great leaders of the bar today. The big question I have in my mind is whether this is an inevitable over-powering trend which cannot be changed or whether it is a phase we are passing through to which there will be a reaction.

I like your example of the Doge of Venice. Naturally, I would like to see your report to Roosevelt. I will read it and send it back to you.

Sincerely,

Honorable Bernard M. Baruch
597 Madison Avenue
New York 22, New York

Vendor at the Mine



Page 2
Mr. Baruch
February 21, 1949

The lack of equity capital does not disturb them at all. I think that a very interesting study could be made of what is going on in the large law offices of the country. They are busier, if anything, than they ever were, and they make just as much money, but instead of their work being as it was when I first came down to practice law in terms of productive projects such as putting together groups to build or expand railroads, public utilities, industries, etc., today their work is almost all defensive. They deal with taxes, anti-trust suits, litigations, Government administration work, almost exclusively. In short, directly or indirectly their clients are the Government. The investment markets are dried up and such accumulations of savings as there are seem to repose in the insurance companies and the savings banks where the investment is almost wholly in terms of fixed obligations. The result is that these people that I am talking about are all beginning to eye those institutions as being something which the Government must control as it is becoming monopolistic. Again there is an example of the tendency to rationalize everything in terms of further Government action. Incidentally, this tendency I think accounts for the fact that there are very few great leaders of the bar today. The big question I have in my mind is whether this is an inevitable over-powering trend which cannot be changed or whether it is a phase we are passing through to which there will be a reaction.

I like your example of the Doge of Venice. Naturally, I would like to see your report to Roosevelt. I will read it and send it back to you.

Sincerely,

Honorable Bernard M. Baruch
597 Madison Avenue
New York 22, New York

Based on this evaluation, we identified the vendor at the Mine as our first choice for digitization of our archival materials for the following reasons:

- Highest quality images and OCR
- Full digitization workflow, with complete handling of equipment, QA, and large support staff
- Ongoing digitization operation which can support our needs at a large or small scale on demand
- Solid experience handling archival materials and ensuring safety and protection of assets in their control
- Great willingness to work with us and ensure that they customize their process to suit our needs

The second best option, which would permit us to reduce cost and take advantage of periodically available personell resources, was to digitize the materials using our existing hardware within our team.

We rejected the World Bank Copy Center option because it would require purchase or rental of equipment. The available staff also appear more versed in bulk duplication of standard documents rather than archival materials.

Due to resource constraints the final decision was to use the internal Archives staff to perform the scanning.

V. Technology Choices and Challenges

An integral part of any digitization project is the actual activity of digitizing, processing, and publishing of archival materials. Successful digitization requires smart choices of hardware, software, standards, and training. This section describes the choices and challenges we faced after choosing to digitize the archival materials with our existing team using hardware already in place.

The materials we have selected to digitize are predominately from the 1950s, 1960s, and 1970s from around the world. This means that many of the pages we need to scan are larger than traditional letter size or A4 sizes. We quickly realized that the Fujitsu fi-6240 with a flatbed size of eight and a half inches by eleven and a half inches was too small for almost all of our materials. Instead, a Fujitsu fi-6770A with a flatbed size of twelve inches by eighteen inches was able to handle the many oversized documents. We do not use the automatic sheet feeder which is part of this model, rather we place pages individually on the flatbed for scanning.

While we preferred an overhead style scanner, this option was ruled out due to cost. We use Fujitsu's ScandAll Pro software to perform the scans, which is the default software shipped with our scanner. The images are captured as 400 dpi, 24-Bit color, TIFFs with no compression. We consider the TIFFs our master images. On average, we found that we are able to scan two pages a minute with these settings, software, and hardware. Another scanner we are considering is the Epson Expression 10000XL Wide-Format scanner. This scanner has a lower price point than the Fujitsu scanner, but can still support our requirements (large flatbed, 400 dpi, 24-big color). We have restricted our evaluations to those scanners rated as professional and rated to maintain their functionality through 100,000 manual scans.

We have chosen to provide access to our digitized materials as PDFs, with each folder of archival materials corresponding with a single PDF. This ensures that all the documents within a single archival file unit are kept together in their original order. We have created a standard cover sheet which is the first page in every PDF and includes contextual information about the Fonds, Sub-Fonds, Series, and Sub-Series as applicable as well as basic citation format and a link to our standard copyright information. This cover sheet is created in Microsoft Word. Additional basic metadata is added in the properties of the PDF.

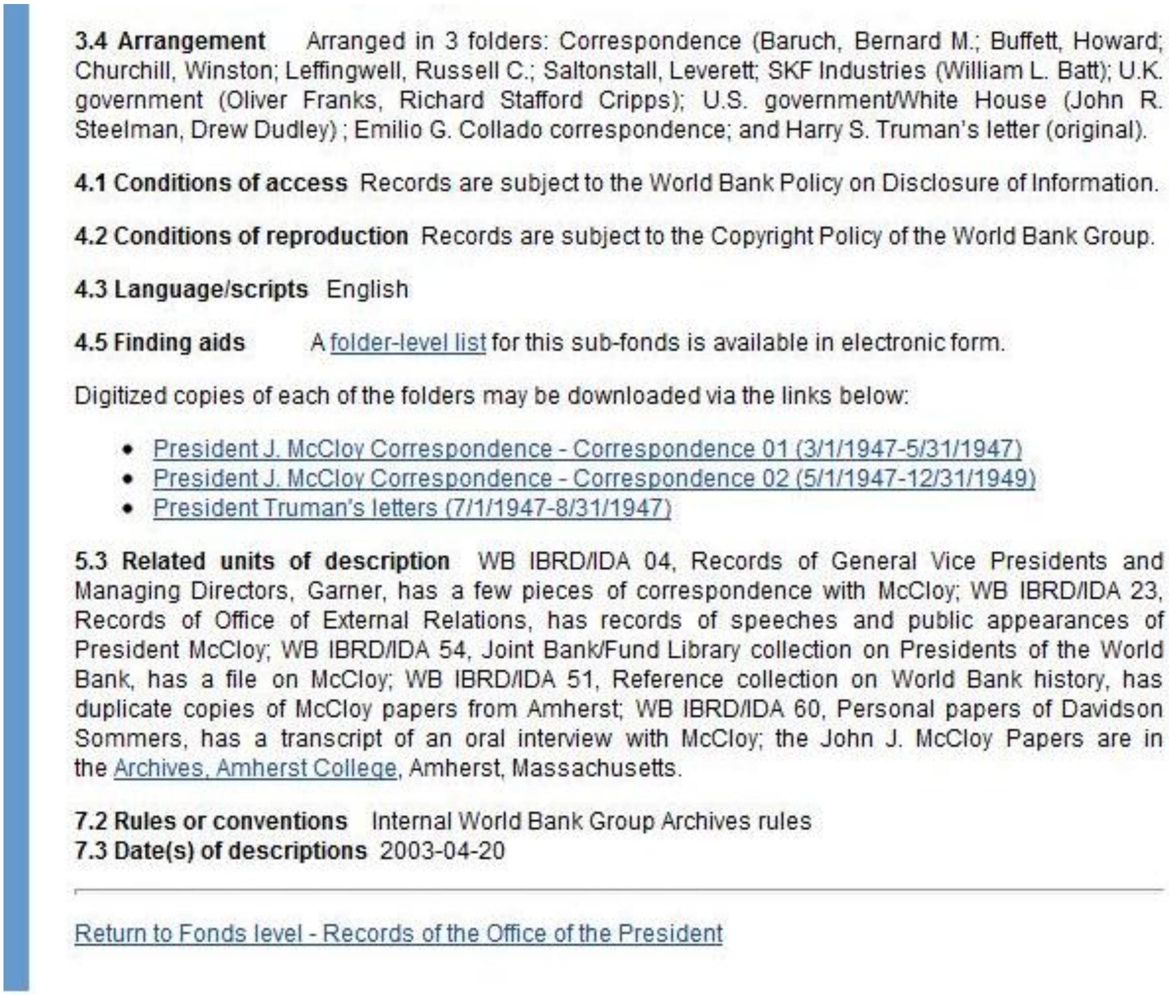
Our internal tracking system ("TRIM" in diagram above) already assigns every archival folder a unique identification number. This number is used as the prefix for every TIFF generated, followed by a sequential number to preserve the order of the pages; a standard TIFF file name would be 123456-0001.pdf. This ensures that any image can be easily tracked back to the exact location in the collection from the file name. This unique prefix number also permits us to tie the PDF to the correct archival finding aid.

High resolution 400 DPI full color images are large. Each one of our images averaged 50 megabytes. In order to hold the TIFFs for a single digitized folder we needed approximately 3 GB of hard drive space. This adds up very quickly. We made sure that our TIFF master files are stored in two locations to protect against any data loss.

Once the master TIFFs were created, we used Adobe Acrobat 9 Pro to combine the Microsoft Word cover sheet and all the TIFF images belonging to a single folder into a PDF. Then the contents were run through optical character recognition and finally compressed to reduce the size of the file. We found that compressing the PDF did not dramatically impact the final appearance of the pages. We used the standard 'Reduce File Size' option available in Adobe Acrobat 9 Pro to compress the document to one third of its original size. Basic metadata is populated in the properties of the PDF, including Title, Author

(World Bank Group Archives), and standard copyright information. Finally, the PDF was linearized to ensure that the time before the end user views the first page of the PDF is optimized.

Given that we have not implemented a formal platform for publishing our digitized materials, we chose to upload the PDFs to our standard web publishing platform. We then linked to each folder from within the appropriate section of the archival finding aid which describes the portion of the collection the folder belongs (as shown in screenshot below), permitting users to discover and download folders within the context of the finding aid.



3.4 Arrangement Arranged in 3 folders: Correspondence (Baruch, Bernard M.; Buffett, Howard; Churchill, Winston; Leffingwell, Russell C.; Saltonstall, Leverett; SKF Industries (William L. Batt); U.K. government (Oliver Franks, Richard Stafford Cripps); U.S. government/White House (John R. Steelman, Drew Dudley) ; Emilio G. Collado correspondence; and Harry S. Truman's letter (original).

4.1 Conditions of access Records are subject to the World Bank Policy on Disclosure of Information.

4.2 Conditions of reproduction Records are subject to the Copyright Policy of the World Bank Group.

4.3 Language/scripts English

4.5 Finding aids A folder-level list for this sub-fonds is available in electronic form.

Digitized copies of each of the folders may be downloaded via the links below:

- President J. McCloy Correspondence - Correspondence 01 (3/1/1947-5/31/1947)
- President J. McCloy Correspondence - Correspondence 02 (5/1/1947-12/31/1949)
- President Truman's letters (7/1/1947-8/31/1947)

5.3 Related units of description WB IBRD/IDA 04, Records of General Vice Presidents and Managing Directors, Garner, has a few pieces of correspondence with McCloy; WB IBRD/IDA 23, Records of Office of External Relations, has records of speeches and public appearances of President McCloy; WB IBRD/IDA 54, Joint Bank/Fund Library collection on Presidents of the World Bank, has a file on McCloy; WB IBRD/IDA 51, Reference collection on World Bank history, has duplicate copies of McCloy papers from Amherst; WB IBRD/IDA 60, Personal papers of Davidson Sommers, has a transcript of an oral interview with McCloy; the John J. McCloy Papers are in the Archives, Amherst College, Amherst, Massachusetts.

7.2 Rules or conventions Internal World Bank Group Archives rules
7.3 Date(s) of descriptions 2003-04-20

Return to Fonds level - Records of the Office of the President

In order to be able to enhance the eArchives interface in the future, we are storing our master TIFFs for future use. We expect that we may need them to generate a different access format from the PDFs we are choosing to use now. As mentioned earlier, the TIFF files are quite large – so it is important to consider where these files can be kept safely for long term reference. We currently have two mirrored low access hard drives which will make sure the few of us who need access have it, but do not require the same level of ongoing support as business critical or content management systems would require.

VI. Infrastructure Recommendations

The following table compares two approaches to the infrastructure requirements for supporting a digitization project. The ‘Lightweight Approach’ permits the digitization program to move forward with digitization without the required investment for the ‘Full Implementation’. While the efficiency will may be lower with this approach, given the other bottlenecks related to A2I review and condition assessment, it is important to begin digitization as soon as possible. This ‘Lightweight Approach’ will permit the project to move forward. The repository, integrated web publishing, and workflow management can be incorporated later. The ‘Lightweight Approach’ has been adopted for our project, with hopes to transition to the ‘Full Implementation’ after we secure funding for a greater investment in infrastructure.

Function	Lightweight Approach	Full Implementation
Storage of master TIFFs and PDFs	Hard drive, with mirrored backup, with folder ID number in all names	Document repository based on Documentum. See Repository Business Requirements Document
Image QA	Spot checking manual QA via review of PDF	Simple workflow implemented within document repository
Extraction of folder metadata from TRIM	Manual as needed. Possible use of script to generate skeleton folder structure	Automated synchronization of folder metadata between TRIM and document repository
Web Publication	Manual upload of PDFs to web CMS. Manual creation of links to PDF from folder lists within finding aid.	Integrated publishing support with push from document repository to web platform. See ‘UI Business Requirements Document’.
Web QA	Validation of published content	
A2I Review	Required	
Condition assessment before and after	Required	

digitization		
Basic Metadata	• PDF cover sheet including all archival context • TRIM folder ID • Image capture admin metadata • File names including page order numbers	
Extended Metadata	• Master list of staff performing scanning, tracked at box level	• Increased integration of descriptive archival metadata into PDF properties • XML representation of metadata based on standard conventions

VII. Benefits of the Project

The following benefits are being realized by this project: increased openness of a rich store of development information which is unavailable anywhere else; provision of a best practice example in the freedom of information which encourages countries to adopt similar policies to promote transparency and demonstrate good governance; a better informed world citizenry who has direct access to important information about development; provision of easy access to public records from the World Bank Archives to interested parties around the world; management of appropriate adherence to Access to Information Policy in the dissemination of records online; and, our best efforts to present the materials in a manner which communicates the context and original order.

VIII. Conclusion

The eArchives project supports the spirit of the Access to Information policy through the digitization and online publication of public archival materials. Our dream is to make accessible and public many high value records which the World Bank has amassed over the years – in every sector, every network, every country – and put this great, deep, and broad collection of development knowledge online for the world to access free of charge.