# Technical and Archival Evaluation of Test Transfer from State Department SMART System

On April 13, 2004, the National Archives and Records Administration signed a Memorandum of Understanding with the Department of State. The subject of the memorandum was to demonstrate the electronic transfer of e-documents to NARA and to explore knowledge management technologies related to the analysis of large quantities of data. The documents referred to in the MOU were products of the State Messaging and Archive Retrieval Toolset (SMART) system. Over the last seven years the State Department has developed and implemented the SMART system and the system now contains records which can be used to fulfill the MOU.

Section A of the MOU relates to the identification of the target collections. The State Department has identified these collections within the SMART system. The test transfer consists of approximately 24,400 messages. 9,000 of these messages are record emails and the remainder are traditional cables. The date range of the test data is January 1, 2009 through March 31, 2011.

Section B of the MOU indicates that the test records will be transferred to NARA using simplified delivery processes. NARA and the State Department decided to transfer the test data on optical media. The test transfer was received and has been successfully evaluated.

Section C of the MOU details an evaluation process related to the transfer of test records. NARA has completed the evaluation the test data. The results of that evaluation appear below.

Section D of the MOU describes various knowledge management techniques which NARA will utilize to facilitate analysis of the test data. This analysis has been performed and the results are below.

The test records are being handled according to Section E of the MOU, which indicates that the test records will only be used for the purposes indicated by the MOU, and that NARA will exercise due diligence in its handling of the data.

The SMART test transfer arrived at NARA on one DVD in a compressed format. The messages were uncompressed into 24,458 folders, comprising approximately 7 GB of data. Each folder's name comprises 36 characters (i.e. ffb229d1-ea1a-43e0-9509-9eb2badf60cb). Each folder represents one message, and any attachments. Each message exists in two formats, XML and PDF. The XML files were all named manifest.xml, while the PDF files were named according to an unknown specification, which appears to utilize the location of the consulate sending the message, as well as other information. Attachments were named similarly to the PDF files. Accompanying the test transfer were a cover letter and the XML Schema Definition, which defines the fields in the XML file.

A technical evaluation of the data was performed, examining the records for issues which may affect access, authenticity, or comprehension. Where applicable, current NARA Transfer Guidelines were used to evaluate the data. The technical evaluation revealed several major issues, as well as several minor issues. Those issues are:

1. Major issue - Text is missing from PDF (i.e. "10-SAN JOSE-416.eml.pdf"). At least one PDF record had entire sentences missing from the file. This was confirmed by comparing the text in the PDF file to the text in the XML file. This issue is very serious and affects the authenticity of the record.

2. Major issue - Scan resolution is too low for NARA standards in PDF (i.e. "1-Bouterse 1-27-11.PDF.pdf.pdf"). In some cases attachments to emails were scanned from existing paper. Several scans were identified which do not meet current NARA Transfer Guidelines for scanned textual records. The scanning resolution of these images was below the NARA minimum of 300dpi.

3. Major issue - Scans in PDF use lossy compression (i.e. "1-Bouterse 1-27-11.PDF.pdf.pdf"). According to current NARA Transfer Guidelines, records created from scanned text may not be saved using a lossy compression format, which throws away data to reduce file size. Several scans were identified that utilize a lossy compression format.

4. Minor issue – There are possible text encoding issues in PDF (i.e. " 09-FTR-96.eml.pdf"). At least one PDF file, and the accompanying XML file, had question marks replacing letters which contained accent marks. This led to sentences like "Con relaci????n a lo conversado el d????a de ayer en vuestra Agencia." in the XML file. The PDF version of the file was similar, but with only two question marks instead of four.

5. Minor issue – There are possible code snippets in PDF (i.e. " 11-ISLAMABAD-506.eml.pdf.pdf"). Several files were identified which had apparent snippets of code (i.e. <![endif]->) at the beginning of the PDF. The code snippets do not occur in the XML version of the messages.

6. Minor issue – There are multiple file format extensions in PDF file name. As seen above, many of the files have multiple file format extensions in the PDF file names. This may lead to confusion when searching or attempting to identify specific files.
7. Minor issue – There are attachments referenced in many XML files called metadata.dat that do not appear in the record's directory (i.e. "10-FTR-14876.eml.pdf.pdf").
8. Minor issue – PDF versions of several emails indicated the attachment of files which do not appear in the record's directory (i.e. "10-FTR-14876.eml.pdf.pdf").
9. Minor issue – At least one PDF record contained images which were not viewable (i.e. " 11-ISLAMABAD-506.eml.pdf.pdf").

In addition to the above technical evaluation, a brief archival evaluation was also performed which raised several questions.

Archival Evaluation

1. Issue - All XML files have same name. All 24,000 messages were named manifest.xml. This can cause considerable confusion when attempting to provide reference access to the records. It also makes it very difficult to properly replace a file which has been removed from its directory structure. In addition, the naming of the folders is not intuitive, nor did State provide any finding aid which links a folder name to a specific message.
2. Question – Why are there both PDF and XML versions of the records? Which version is considered the record? In the small sample reviewed, it appears a user needs both the PDF and the XML file to understand the record. The XML files include additional record management and other metadata that is not part of the record material of the record (such as MessageID or hash codes) so it makes sense that such metadata would not be included in a "user friendly" PDF version of the record material of the record. However, it is not clear what information is used to create the "user friendly" PDF version of the record. Are the PDF files generated from the XML files or are both files generated from the message as stored in SMART? Is there a crosswalk for the fields in the PDF files vis-à-vis the fields in the XML files with an explanation for any differences?
3. Question – How does the user identify what records are emails versus telegrams versus memos? It is unclear if the XML field MessageType provides this information and it appears there is nothing in the PDF to indicate this.
4. Question - How does one identify or maintain the link between the two versions of the message and any attachments? This is especially problematic if all the XML files are names manifest.xml and the attachments do not contain the MRN. If the plan is to

transfer the records with a folder for each record containing both versions (formats) of the record and any attachments, that would require maintaining the directory structure for preservation and access.

5. Question - Is the MRN the only unique number that appears on both the PDF and XML that can be used to link the two versions?

As part of the evaluation of the data, NARA has explored various knowledge management techniques designed to facilitate access to the data. Although no proper finding aid was transferred along with the data, it is possible to work with the XML files to create a finding aid. A simplified script was written which extracted various pieces of data from the XML files (for example, file name, message type, author, subject, date, attachments names, etc.). This script created an Excel spreadsheet which could be utilized as a finding aid. In addition, since the XML files are ASCII text, a full-text search engine could also be used for access purposes.

The NARA evaluation of the test data indicates several serious problems, both technical and archival, which need to be resolved before an actual transfer of records is attempted. In addition, significant additional metadata will need to accompany any transfer.