June 5, 2015

Ms. Margaret Grafeld
Deputy Assistant Secretary
for Global Information Services (A/GIS)
Suite 8000, SA-2
515 22nd St. NW
Department of State
Washington, DC 20037

Dear Ms. Grafeld:

This letter concerns the recent test transfer of data from the Department's 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. NARA has completed evaluation of the test data.

The SMART test transfer arrived at NARA on one DVD in a compressed format. Accompanying the test transfer were a cover letter and the XML Schema Definition, which defines the fields in the XML file. 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.

NARA staff performed technical and archival evaluations of the data, examining the records for issues which may affect access, authenticity, or comprehension. These evaluation revealed several major issues, as well as several minor issues, and other questions. The technical issues are prefaced with a "T" and the archival questions are prefaced with an "A".

T-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.

T-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 at a resolution of these images below the NARA minimum of 300dpi.

T-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.

T-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.

T-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.

T-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.

T-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").

T-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").

T-9. Minor Issue: At least one PDF record contained images which were not viewable (i.e. " 11-ISLAMABAD-506.eml.pdf.pdf").

A-1. Why do all XML files have same name? All 24,000 messages were named "manifest.xml". This will 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.

A-2. Why are there both PDF and XML versions of the records? Which version is considered the record? or does the record consist of both? 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?

A-3. 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.

A-4. 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.

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

These technical and archival issues and questions must be resolved before the actual transfer of records is attempted. In addition, significant additional metadata will need to accompany any transfer.

We appreciate the Department's cooperation and look forward to receiving your explanations and answers to the issues and question noted above. We will consider action on the MOU complete when the Department has addressed these issues to NARA's satisfaction.

Sincerely,

PAUL M. WESTER. Jr.
Chief Records Officer

(Please be sure to send cc's of the letter that goes out to Greg Lepore and Lynn Goodsell in NWME and to David Langbart in NWCT.)