PROCEDURES FOR TRANSMITTING DOCUMENTS TO THE U.S. SENATE PERMANENT SUBCOMMITTEE ON INVESTIGATIONS

Due to security procedures at the U.S. Senate, the following are guidelines for transmitting documents to the Permanent Subcommittee on Investigations.

VIA PERSONAL DELIVERY AND/OR COURIER SERVICE (Preferred):

Deliveries during normal business hours (9:00 am to 5:00 pm) should be brought in an unsealed envelope directly to the Russell Senate Office Building Room 199 and dropped off with Subcommittee Clerk Kate Kielceski.

VIA FEDERAL EXPRESS OR OTHER COMMERCIAL CARRIERS:

Before shipping any production through FedEx or other commercial carriers, please first discuss with Subcommittee Staff. Subcommittee staff can provide an address for a secure delivery of the production, which may be located offsite.

***DO NOT SEND PACKAGES VIA U.S. POSTAL SERVICE*:**

Packages sent via the U.S. Postal Service are irradiated. Irradiation causes disintegration of the documents being shipped, often rendering them unusable. Discs have been known to arrive melted due to the irradiation process.

Any questions regarding the transmittal of documents to the Subcommittee can be directed to Subcommittee Clerk Kate Kielceski at 202-224-9868 or Kate_Kielceski@hsgac.senate.gov.

Updated March 2018

Data Delivery Standards
Permanent Subcommittee on Investigations
United States Senate

The following document describes the technical requirements for electronic productions produced to the Senate Permanent Subcommittee on Investigations ("PSI"). **Any proposed formats other than what is listed below (including databases) should not be produced without prior discussion with PSI staff.** PSI uses Concordance 10 and Concordance Image 5.

General Instructions:

1. Provide a cover letter with each production which includes the Bates range and a general description of the documents. The cover letter should also summarize the number of records, images, emails and attachments in the production.
2. Produce documents in the same form that they were created or maintained. Documents created or stored electronically should not be produced in hard copy.
3. Deliver data on CD, DVD, or hard drive. Hard drives with external power supplies are preferred. The smallest number of media is requested.
4. Label all media submitted. Include on the label at least the following information: producing party, production date, Bates range, and disk number, if applicable.
5. Provide all passwords for documents, files, or compressed archives provided in the production.
6. To the extent practicable, de-duplication of email and native file productions is preferred.
7. Overview of preferred formats for production:
    a. Paper Documents – Scanned paper converted/processed to TIFF files, Bates numbered, and includes OCR text.
    b. Email Collections – Electronic mail converted/processed to TIFF files for the email and attachment(s), Bates numbered, includes a link to the email or native file, and includes full text.
    c. Native Files – Electronic documents converted/processed to TIFF files, Bates numbered, includes a link to the native file, and includes full text.

A. Paper Documents:

1) **Image files.** Images must be Group IV TIFF files (single or multi-page files). All images should be Bates numbered. The number of files per folder should be limited to 1,000 files.

2) **Delimited Text file.** At a minimum, this file must contain an IMAGEID field (image key used to reference images in Concordance Image). The image key must be unique, fixed length, and CANNOT be the Bates number of the document. Bates numbers (endorsed on the documents and included in the delimited text file) MUST be delivered in a consistent manner for sorting purposes. For example, if the first production delivered is Bates stamped ABC-0000001-ABC-0005267, subsequent productions with the same prefix should have the same format (spaces, dashes, etc.) and the same number of digits, not ABC 0005268, ABC0005268 or ABC-00005268. The delimited text file must also include a header record. The delimiters for the file should be as follows:

    Comma – ASCII character 20
    Quote – ASCII character 254
    Newline – ASCII character 174

3) **OCR Text.** The OCR text provided to the PSI can be delivered two ways. (1) The OCR text can be delivered as multi-page ASCII files. The name of the file must match the IMAGEID field. (2) The OCR text can be included in the Delimited Text file (OCRTEXT field). Option 1 is preferred.

If possible (regardless of delivery method), please place page markers at the beginning or end of each OCR text page as shown:

*** LA000001 ***

The data surrounded by *** is the Concordance Image ImageID.

4) **Concordance Image Cross-Reference file.** The Concordance Image cross-reference file is a comma delimited file consisting of six fields per line. There must be a line in the cross-reference file for every image in the database. The format for the file is as follows:

ImageID,VolumeLabel,ImageFilePath,DocumentBreak,FolderBreak,BoxBreak,PageCount

- *ImageID*: The unique designation that Concordance and Concordance Image use to identify an image.
- *VolumeLabel*: Optional.
- *ImageFilePath:* The full path to the image file.
- *DocumentBreak:* If this field contains the letter "Y," then this is the first page of a document. If this field is blank, then this page is not the first page of a document.
- *FolderBreak:* Leave empty.
- *BoxBreak:* Leave empty.
- *PageCount:* Optional.

B. <u>Email Collections:</u>

Preferred Format: Delimited Text with Images and Native Attachments

1) **Image files.** The producing party will provide a TIFF image for each page of the email and attachment(s). Images must be Group IV TIFF files (single or multi-page files). All images should be Bates numbered. The number of TIFF files per folder should be limited to 1,000 files. Refer to the Paper Documents section for Bates and image key numbering rules.

2) **Native files.** The producing party will provide a copy of the email and native attachment files. The number of native files per folder should be limited to 1,000 files.

3) **Delimited Text file.** The text and metadata of the email and the attachment(s) is extracted and entered in the appropriate fields and provided as an ASCII delimited text file. The email will be the "*parent*" and the attachment(s) will be the "*child.*" An email may have more than one *child.* The *child* attachment's Bates number will be listed in the *parent* email's coded fields under *CHILD_BATES.* If there is more than one attachment, list the first Bates number of each attachment and separate them by semi-colons (;). The *parent* email's Bates number will be listed in the *child(s)* attachment(s) under *PARENT_BATES.* The *child/children* will immediately follow the parent record. The following is a field definition table of the data requested.

Field Definitions - Email

Field	Comment
BEGBATES	First Bates number of email
ENDBATES	Last Bates number of email
BEGATTACH	First Bates number of attachment range
ENDATTACH	Last Bates number of attachment range
PARENT_BATES	First Bates number of parent email
CHILD_BATES	First Bates number of "child" attachment(s); can be more than one Bates number listed; depends on number of attachments
CUSTODIAN	Mailbox where the email resided
FROM	Sender
TO	Recipient(s)
CC	Carbon copy recipient(s)
BCC	Blind carbon copy recipient(s)
SUBJECT	Subject of the email
DATE_SENT	Date the email was sent
TIME_SENT	Time the email was sent; must be a separate field and cannot be combined with the DATE_SENT field
HYPERLINK	Hyperlink to the email
FILE_EXTEN	The file extension of the email; will vary depending on the email format
AUTHOR	Empty for email
DATE_CREATED	Empty for email
TIME_CREATED	Empty for email
DATE_MOD	Empty for email
TIME_MOD	Empty for email
DATE_ACCESSD	Empty for email
TIME_ACCESSD	Empty for email
PRINTED_DATE	Empty for email
FILE_SIZE	Size of email in KB
INTFILEPATH	Location of email
MESSAGE ID	Unique Identifier from the email system used to duplicate emails
CONVERSION ID	Identifier from the email system used to group and manage related emails
CONVERSATION INDEX	Identifier from the email system used to group and manage related emails
HASHVALUE	Value generated for deduplication
TEXT	Text of the email

Field Definitions - Attachment

Field	Comment
BEGBATES	First Bates number of attachment
ENDBATES	Last Bates number of attachment
BEGATTACH	First Bates number of the attachment range
ENDATTACH	Last Bates number of the attachment range
PARENT_BATES	First Bates number of parent email
CHILD_BATES	First Bates number of "child" attachment(s); can be more than one Bates number listed; depends on number of attachments
CUSTODIAN	Mailbox where the email resided

FROM	Empty for attachment
TO	Empty for attachment
CC	Empty for attachment
BCC	Empty for attachment
SUBJECT	Empty for attachment
DATE_SENT	Empty for attachment
TIME_SENT	Empty for attachment
HYPERLINK	Hyperlink to the native attachment
FILE_EXTEN	The file extension will vary depending on the document type
AUTHOR	Attachment/native file metadata
DATE_CREATED	Attachment metadata
TIME_CREATED	Time the attachment was created; must be a separate field and cannot be combined with the DATE_CREATED field
DATE_MOD	Attachment metadata
TIME_MOD	Time the attachment was modified; must be a separate field and cannot be combined with the DATE_MOD field
DATE_ACCESSD	Attachment metadata
TIME_ACCESSD	Time the attachment was accessed; must be a separate field and cannot be combined with the DATE_ACCESSD field
PRINTED_DATE	Attachment metadata
FILE_SIZE	Size of file in KB
INTFILEPATH	Path where attachment file was stored
HASHVALUE	Value generated for deduplication
TEXT	Text of the attachment

The delimited text file must include a header record. Please refer to the Paper Documents section for ASCII character assignments.

4) **Full Text.** When the full text is not provided in the ASCII delimited text file or if text exceeds 12MB in the TEXT field, the full text provided to the PSI can be delivered as multi-page ASCII files. The name of the file must match the image key field. Any document in which text cannot be extracted should be OCR'd, particularly in the case of PDFs without embedded text.

5) **Concordance Image Cross-Reference file.** The Concordance Image cross-reference file is a comma delimited file consisting of six fields per line. There must be a line in the cross-reference file for every image in the database. Refer to the Paper Documents section for file format.

C. Native Files:

Preferred Format: Delimited Text with Images and Links to Native Files:

1. **Image files.** The producing party will provide a TIFF image of the native files. Images must be Group IV TIFF files (single or multi-page files). All images should be Bates numbered. The number of TIFF files per folder should be limited to 1,000 files. Refer to the Paper Documents section for Bates and image key numbering rules.

2. **Native files.** The producing party will provide a copy of the native files. The number of native files per folder should be limited to 1,000 files.

3. **Delimited Text file.** An ASCII delimited file containing the metadata associated with the file, text extracted from the native file, and a directory path to the native file. The fields to be included in the production are as follows:

FIELD	COMMENT
BEGBATES	First Bates number of native file
ENDBATES	Last Bates number of native file
CUSTODIAN	Individual from whom the documents originated
HYPERLINK	Hyperlink to native file
AUTHOR	
DATE_CREATED	
TIME_CREATED	Time the document was created; must be a separate field and cannot be combined with the DATE_CREATED field
DATE_MOD	
TIME_MOD	Time the document was modified; must be a separate field and cannot be combined with the DATE_MOD field
DATE_ACCESSD	
TIME_ACCESSD	Time the attachment was accessed; must be a separate field and cannot be combined with the DATE_ACCESSD field
PRINTED_DATE	
FILE_SIZE	Size of file in KB
PATH	Path where native file was stored
HASHVALUE	Value generated for deduplication
TEXT	Text extracted from native file.

The delimited text file must include a header record. Please refer to the Paper Documents section for ASCII character assignments.

4. **Full Text.** When the full text is not provided in the ASCII delimited text file or if text exceeds 12MB in the TEXT field, the full text provided to the PSI can be delivered as multi-page ASCII files. The name of the file must match the image key field. Any document in which text cannot be extracted should be OCR'd, particularly in the case of PDFs without embedded text.

5. **Concordance Image Cross-Reference file.** The Concordance Image cross-reference file is a comma delimited file consisting of six fields per line. There must be a line in the cross-reference file for every image in the database. Refer to the Paper Documents section for file format.

If Unable to Comply with Format Described Above:

Any proposed formats other than what is listed above should not be produced without prior discussion with PSI staff.

All documents derived from word processing programs, email applications, instant message logs, spreadsheets, and wherever else practicable, may be produced in text searchable PDF format. Spreadsheets shall also be provided in their native form. Audio and video files shall be produced in their native format, although picture files associated with email or word processing programs shall be produced in PDF format along with the document it is contained in or to which it is attached.

Native files may be delivered in Custodian named folders.

If PDFs are delivered, all PDF files must meet the following requirements:

1. All PDFs must be unitized i.e. each PDF represents a discrete document; a single PDF cannot contain multiple documents
2. All PDFs must contain embedded text to include all discernable words within the document, not selected text.
3. The PDF file will be named as the Bates range, with all document text contained within.

# # #