April, 2007

U.S. Department of Justice, Antitrust Division
Standard Specifications for Summation Database Production

This document describes the specifications and procedures for producing an image-based production to the Antitrust Division as a loaded Summation database**.** The following page describes the sample files provided with these specifications. Included are two Summation databases, one loaded as a reference, the other an empty shell for your use. Please contact the Division if you have any questions about these technical details. The items below highlight areas of particular importance.

1) Database files (such as an Access .MDB) should not be produced in this manner. Database productions should be discussed with the appropriate government legal and technical staff to determine the optimal production format.

2) Pay special attention to the PARENTID and ATTCHIDS fields, which are used by Summation to keep track of email families, and have often been delivered incorrectly by vendors.

3) Note that PowerPoint and Excel files require special handling. Both PowerPoint and Excel files should be produced natively, with links referenced in the DOCLINK field. In addition, you must provide TIFF images of the entire PowerPoint. PowerPoint files should be produced in full slide image format along with speaker notes, with any speaker notes following the appropriate full slide image. For Excel spreadsheets, provide tiff images of the first 5 pages, since images of spreadsheets often comprise thousands of pages, and are thus not useful for review purposes.

4) Extracted Text should be provided with all records, except for documents that originated as hard copy. For the hard-copy records, please provide OCR text. For redacted documents, provide full text for the redacted version. The extracted text files should include page breaks that correspond to the “pagination” of the image files.

5) Before beginning production, you must produce a sample, including emails, attachments, and non-email files. We will take a day or so to evaluate that sample and confirm the technical details. If we identify a problem, you will need to resubmit the sample until we confirm there are no problems.

6) When you provide the full production, we must receive a cover letter that provides total document and page counts, so that we can verify the records in the database. The counts should be reported by custodian. With any submission, documents from an individual custodian should be confined to a single load file.

7) You must label the media provided, whether CD, DVD, or hard drive. At a minimum, the label must include a unique number (Submission #), the company providing the response, and any references necessary to link to the information in the cover letter.

Folders & Files within the ZIP file

Two Summation Databases

DOJSAMPL

The Loaded Sample Database. This folder is the Case Directory for the Summation database that is fully loaded with DOJ sample data, including Metadata fields, Images, OCR, and Native files.

DOJSHELL

The Empty Shell database. This folder is the Case Directory for the Summation database that the vendor will use to load the data. This Shell will be used for all databases produced to DOJ. This database carries the same layouts and field names as the DOJSAMPL database.

Files (Bookmarked within PDF)

Image Details & Load File Specifications

Document detailing Image, Extracted/OCR Text file, DII, and Metadata production formats.

Metadata Fields & Family Record Specifications

Document detailing requested metadata fields and production of Family Records.

Summation Database Specifications

Explains Summation Database Specifications: e.g. limit of OCRbase, limit on number of records, where native files are to be stored, etc.

Summation Submission Requirements

Explains what information needs to be provided per submission.

Sample Cover Letter Spreadsheet

Sample data that lists the specifics associated with each submission (# Images, # Records, Hard Drive #, Submission #, etc.)

Sample Deduplication - Custodian Append File

** ONLY USED WHEN DEDUPING ACROSS CUSTODIANS**

Often provided during a rolling production. As more Custodians are discovered on documents that have been de-duplicated out, this file is updated with the new custodian information.

*Note: **AppendDate** = a multi-entry date field that designates when this data was appended, so DOJ can keep track of every time that record was edited with new Custodian information.*

U.S. Department of Justice, Antitrust Division

IMAGE & LOAD FILE SPECIFICATIONS

Image Details	
Image Files	Group IV Single-Page Tiffs
	Filenames cannot have embedded spaces
	Images for a document must be in one folder
	Number of image files should be limited to 5,000 per folder
	Files should be named the <PageID>.TIF
	Ex. DOJ-005.TIF

Summation Image Load file (.dii) Specifications	
Bold indicates a constant. *Italics* indicate a variable.	
@Fulltext DOC	Indicates that there is a Text file attached to the records
@T *IMAGETAG*	Required: Unique identifier for document
@D @I *Tiff Path*	Required: Directory location designation
Image Files	Required: listing or iteration of files
IMAGETAG	Identical to the Begdoc#
TiffPath	Path to Image Files
Image Files	Individual or Iterated listing of Tiff filenames
	Ex. as Iteration: DOJ-00{3-6}.TIF
Note: 8 character file name limitation for DII file	
Note: The Fulltext line is written once at the top of the DII file.	

Metadata Load File Delimiters	
Field separator	**Vertical Pipe (ASCII 124)**
Field encapsulate	**Carat (ASCII 094)**
Return value in data	**Tilde (ASCII 126)**
Multi-value field	**Semi Colon (ASCII 059)**
Dates format	**MM/DD/YYYY**
Note:	Hard Returns at End of Record ONLY

Text File Specifications	
A single Text file per document	
The name of the Text file should equal the first page's Bates of the document, with a TXT extension	
There must be a carriage return and line feed in the first 80 characters of text	
Text files should include page breaks that correspond to the “pagination” of the image files.	
When loading using a DII:	
The Text files should be in the same folder as the corresponding Images, and the name should match the document's first Image file, with a TXT extension.	

U.S. DEPARTMENT OF JUSTICE, ANTITRUST DIVISION

METADATA & FAMILY RECORD SPECIFICATIONS

Default File Layout (.txt)	**Field Name has 8 character limit.**								
	Field Name	**Field Description**	**Field Type**	**Hard Copy**	**E-mail**	**Spreadsheets**	**Presentations**	**Other Elec. Docs.**
	Company	Company submitting data	Note Text	X	X	X	X	X
	Box#	Production Box Number/Submission Number	Note Text	X	X	X	X	X
	Custdian	Custodian(s)/Source(s) -- formatted Last, First	Multi-Entry	X	X	X	X	X
	Begdoc#	Start Bates (including Prefix) - No spaces	Note Text	X	X	X	X	X
	Enddoc#	End Bates (including Prefix) - No spaces	Note Text	X	X	X*	X	X
	DocID	Populate with exact same value as Start Bates	Note Text	X	X	X	X	X
	PgCount	Page count	Integer	X	X	X*	X	X
	ParentID	Parent Bates, including Prefix (ONLY IN CHILD RECORDS)	Note Text	X	X	X	X	X
	Attchids	Child document list - Start Bates of each Child (ONLY IN PARENT RECS)	Multi-Entry	X	X	X	X	X
	FamlyRng	Family Start and End Bates (including Prefix) (i.e. ABC-001 - ABC-003	Note Text	X	X	X	X	X
	Prprties	Record Type -> (File, E-mail, Attachment, Hard Copy); Privilege Notations -> (Redacted, Document Withheld Based On Privilege)	Multi-Entry	X	X	X	X	X
	From	Author -- formatted Last, First	Multi-Entry		X	X	X	X
	To	Recipient -- formatted Last, First	Multi-Entry		X	X	X	X
	Cc	Cc field -- formatted Last, First	Multi-Entry		X	X	X	X
	Bcc	Bcc field -- formatted Last, First	Multi-Entry		X	X	X	X
	Subject	Subject / Document Title	Note Text		X	X	X	X
	DocDate	Document Date / Date Sent - MM/DD/YYYY	Date Keyed		X			
	Timesent	Time email was sent	Time		X			
	Datecrtd	Date Created	Date			X	X	X
	Datesvd	Date Modified	Date			X	X	X
	Datercvd	Date Accessed / Received	Date		X	X	X	X
	Filesize	File size	Note Text			X	X	X
	Attitle	File name - Name of file as appeared in original location	Note Text			X	X	X
	Applicat	Application used to create native file (e.g., Excel, Word)	Note Text			X	X	X
	FilePath	Data's source filepath information	Note Text		X	X	X	X
	Doclink	Current filepath location to the native file	Note Text			X	X	X
	FolderID	Email folder path (sample: Inbox\active) or Hard Copy Folder Information	Note Text	X	X			
	Paragrph	Paragraph # to which the document is responsive	Note Text	X	X	X	X	X
	Hash	Hash Value (used for deduplication or other processing	Note Text		X	X	X	X
	Srchtrms	List of Terms used to identify record as responsive (if search terms used)	Multi-Entry		X	X	X	X

*Indicates field may be empty if only native files produced

Parent IDs, Attachment IDs, and Family Range Details				
Customer Notes:				
Y	Confirm Family Range definition for attached files			
Y	Confirm Field names and types.			
Y	Each member of the Family is its own record			
Family Range Definition:				
	All records will have a family range when the file or email has a parent or children			
	Family Range will start with the first page of the top most parent and go until the last child's last page			
Example:				
	Description:	Top most Email	Attachment to Doc1	Attachment to Doc1
		Doc No. 1	Doc No. 2	Doc No. 3
	Begin Bates	ABC-001	ABC-011	ABC-016
	End Bates	ABC-010	ABC-015	ABC-020
	ParentID	{empty}	ABC-001	ABC-001
	Attchids	ABC-011; ABC-015	{empty}	{empty}
	Family Range	ABC-001 - ABC-020	ABC-001 - ABC-020	ABC-001 - ABC-020

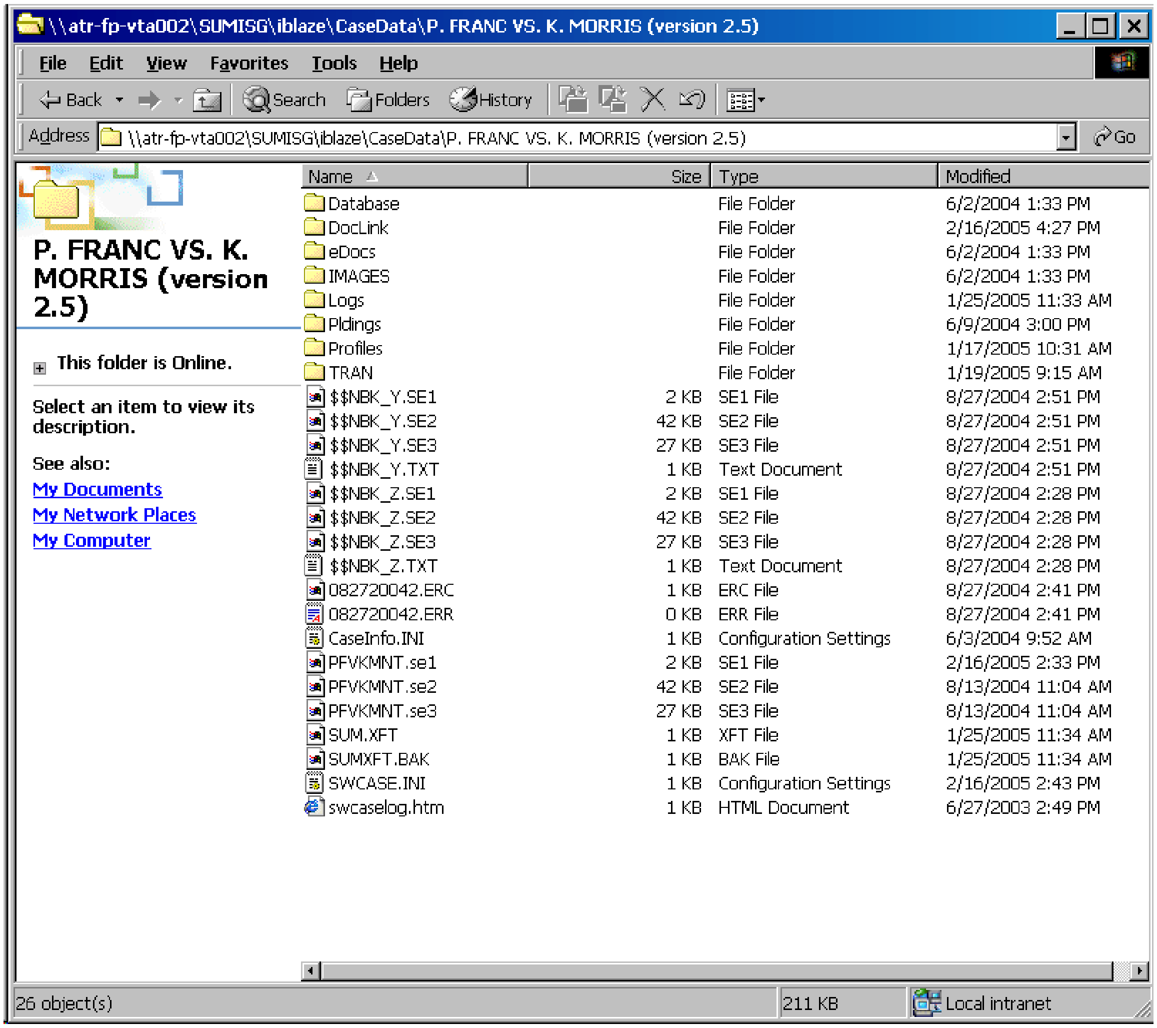
02/22/2007

U.S. DEPARTMENT OF JUSTICE, ANTITRUST DIVISION
Summation Database Specifications

- DOJ to provide empty Summation database shell. The Division currently uses Summation version 2.6.3.

- DOJ will accept loaded Summation databases with the following conditions:
  - Each database has no more that 5-6 GB OCRBase.
  - If the database will have more than 200,000 records, it must be addressed with Division staff prior to production.
  - Custodians do not cross databases, except under limited circumstances.
  - Metadata fields must be populated
  - Data should be structured in the standard Summation format, with images stored inside the Images subdirectory of the case folder.
  - When records include a Doclink field to a native file, the native file should also reside inside the case folder in a folder called DocLink.

See Example of Case Directory Structure, below.



\\atr-fp-vta002\SUMISG\iblaze\CaseData\P. FRANC VS. K. MORRIS (version 2.5)

Address: \\atr-fp-vta002\SUMISG\iblaze\CaseData\P. FRANC VS. K. MORRIS (version 2.5)

P. FRANC VS. K. MORRIS (version 2.5)

This folder is Online.

Select an item to view its description.

See also:
My Documents
My Network Places
My Computer

Name	Size	Type	Modified
Database		File Folder	6/2/2004 1:33 PM
DocLink		File Folder	2/16/2005 4:27 PM
eDocs		File Folder	6/2/2004 1:33 PM
IMAGES		File Folder	6/2/2004 1:33 PM
Logs		File Folder	1/25/2005 11:33 AM
Pldings		File Folder	6/9/2004 3:00 PM
Profiles		File Folder	1/17/2005 10:31 AM
TRAN		File Folder	1/19/2005 9:15 AM
$$NBK_Y.SE1	2 KB	SE1 File	8/27/2004 2:51 PM
$$NBK_Y.SE2	42 KB	SE2 File	8/27/2004 2:51 PM
$$NBK_Y.SE3	27 KB	SE3 File	8/27/2004 2:51 PM
$$NBK_Y.TXT	1 KB	Text Document	8/27/2004 2:51 PM
$$NBK_Z.SE1	2 KB	SE1 File	8/27/2004 2:28 PM
$$NBK_Z.SE2	42 KB	SE2 File	8/27/2004 2:28 PM
$$NBK_Z.SE3	27 KB	SE3 File	8/27/2004 2:28 PM
$$NBK_Z.TXT	1 KB	Text Document	8/27/2004 2:28 PM
082720042.ERC	1 KB	ERC File	8/27/2004 2:41 PM
082720042.ERR	0 KB	ERR File	8/27/2004 2:41 PM
CaseInfo.INI	1 KB	Configuration Settings	6/3/2004 9:52 AM
PFVKMNT.se1	2 KB	SE1 File	2/16/2005 2:33 PM
PFVKMNT.se2	42 KB	SE2 File	8/13/2004 11:04 AM
PFVKMNT.se3	27 KB	SE3 File	8/13/2004 11:04 AM
SUM.XFT	1 KB	XFT File	1/25/2005 11:34 AM
SUMXFT.BAK	1 KB	BAK File	1/25/2005 11:34 AM
SWCASE.INI	1 KB	Configuration Settings	2/16/2005 2:43 PM
swcaselog.htm	1 KB	HTML Document	6/27/2003 2:49 PM

26 object(s) | 211 KB | Local intranet

02/22/2007

U.S. DEPARTMENT OF JUSTICE, ANTITRUST DIVISION
Summation Submission Requirements

Via e-mail or on CD-ROM, the DOJ has provided a DOJShell Summation database directory. Please categorize your submissions by placing the *DOJShell* folder under sequential Database folder names. The folder naming scheme should be 2 to 3 letters (indicating your company) followed by 3 numbers. For example:

> For the first 3 databases from ABC Co., the root of the piece of media (External, CD, or DVD) should display the following folders: ABC001, ABC002, and ABC003. Each of these folders should contain the **loaded** DOJShell Summation Case Directory.

The cover letter for each submission of loaded Summation databases should include information about the loaded Summation database(s) included on each External Hard Drive or other piece of media submitted, preferably in spreadsheet format.

Include the following for each submission:

A. For each piece of media:

1. Assign a unique identifier for each piece of media that is also readily identifiable *on* the piece of media (i.e. Submission #; Serial number is also acceptable), and

2. Identify the Databases on the piece of media.

B. For each Database:

1. The Custodians included;

2. The total number of records;

3. The number of records for each Custodian (e.g., ABC001 contains 183,000 records: Jones - 150,000 records, Smith - 13,000 records, Doe - 20,000 records);

4. The Bates number ranges (and any gaps therein) for each Custodian;

5. The total number of native files in each Database;

6. The number of native files for each Custodian (e.g., ABC001 contains 15,980 native files: Jones - 1,500; Smith - 5,250; Doe - 9,230);

7. The total number of images included in each database; and

8. The total number of images included for each Custodian (e.g., ABC001 contains 15,980 images: Jones - 1,500; Smith - 5,250; Doe - 9,230).

Custodian	Hard Drive Nbr	Volume Name	Begin Bates Nbr	End Bates Nbr	Intentionally Left Blank	Nbr of Records	Nbr of Images	Nbr of Native Files	Summation DB Name	Date Produced
Doe, John	20810	DOJ001	DOJ-00000001	DOJ-00005825		258	5,825	13	DOJSHELL	mm/dd/yyyy
Doe, Jane	20810	DOJ001	DOJ-00005826	DOJ-00009536		365	3,710	52	DOJSHELL	mm/dd/yyyy
Daniels, Jack	20811	DOJ002	DOJ-00009537	DOJ-00015263		1,150	5,726	156	DOJSHELL	mm/dd/yyyy
Weiser, Bud	20811	DOJ002	DOJ-00015264	DOJ-00018273		600	3,009	20	DOJSHELL	mm/dd/yyyy
Brown, Charlie	20814	DOJ003	DOJ-00018274	DOJ-00025225		1,315	6,951	68	DOJSHELL	mm/dd/yyyy
Flinstone, Fred	20814	DOJ003	DOJ-00025226	DOJ-00035625	DOJ-00025698 - DOJ-00025982	2,023	10,115	85	DOJSHELL	mm/dd/yyyy

```
Submission#|Begdoc#|Custdian|AppendDate
^DOJ001^|^DOJ0000018^|^Richards, Joseph;Anderson, Ryan^|^03/07/2005^
^DOJ001^|^DOJ0000020^|^Johnson, Bill^|^03/07/2005^
^DOJ001^|^DOJ0000022^|^Heralds, Steve^|^03/07/2005^
^DOJ001^|^DOJ0000024^|^Ponston, Bobby^|^03/07/2005^
^DOJ001^|^DOJ0000026^|^Henderson, Morris;Williams, Don^|^03/07/2005^
^DOJ001^|^DOJ0000028^|^Smith, Ricky^|^03/07/2005^
^DOJ001^|^DOJ0000030^|^Martins, Jeff^|^03/07/2005^
^DOJ001^|^DOJ0000032^|^Jeffries, Mark^|^03/07/2005^
^DOJ001^|^DOJ0002833^|^Hayes, Jared;Anderson, Ryan^|^03/07/2005^
^DOJ001^|^DOJ0002879^|^Williams, Ronald;Stevenson, Richard^|^03/07/2005^
^DOJ001^|^DOJ0002890^|^Morton, Linda^|^03/07/2005^
```