AFP
AUSTRALIAN FEDERAL POLICE

Document Production Guideline for Cooperating Corporations

October 2021

POLICING FOR A SAFER AUSTRALIA

Report title	Document Production Guideline for Cooperating Corporations
Originating area	Corporate Crime
Date produced	October 2021
Cleared by	Co-ordinator Corporate Crime
Endorsed by	Commander Crime

Contents

About this guideline

This guideline provides information about the format in which digitized records should be produced to the Australian Federal Police (AFP) by corporations that are cooperating in investigations or responding to AFP requests for information. The guideline covers the production of original digital records, copies of digital records and digitised hard copy records.

This guideline should be read in conjunction with the AFP's Corporate Cooperation Guidance.

The production of material in conformity with this guideline should assist the AFP (and the producing entity) by reducing the time, costs and resources associated with investigations. Accordingly, compliance with these guidelines will generally be seen as an indicator of genuine and proactive cooperation.

Any feedback in relation to this guideline should be sent to:

CBR-Crime-Foreign-Bribery@afp.gov.au

Contacting the AFP to discuss the production

Questions about any aspect of how to produce material to the AFP, should be directed to the AFP case officer designated to the investigation or named in any requesting correspondence. Discussing the scope of a request, including the meaning of certain terms, can often result in significant time and cost savings.

The AFP case officer can also be contacted to discuss any particular production. However, while alternative forms of production may be agreed upon by exception, this guideline sets out the AFP's preferred approach.

Definitions

Appendix 1 contains a number of definitions which may be of assistance when reading this document.

Preferred production methods

Electronic material and records, including their metadata, should be produced in their original native file format after it has been preserved. This can be:

- the form in which the digital material is stored—for example, a Microsoft Word document from a file server;
- the native form in which the system that stores the digital material exports them—for example, an PST file exported from Microsoft Outlook; or
- a forensic container used to preserve the digital material in the form it was stored.

Material may also be produced using an evidence review or litigation support system. If an evidence review system is used to review and produce the material the following formats are preferred:

- a Nuix case file,

- a Relativity workspace, or
- a load file format as per below.

In order to ensure structured loading can be completed in the most efficient and expedient manner, it is recommended to provide a sample delivery which includes emails, email attachments, container files and non-email files in the prescribed format below. This can then be verified and any problems can be rectified prior to delivery of the main tranche of material.

How to produce material to the AFP

All data deliveries should be accompanied by a covering letter which provides as much detail about the production process as possible. This should include, although not exclusively:

- the Document Id number ranges used, custodian and/or batch names,
- the total number of records per custodian and/or batch,
- the source location and time zone for each custodian's material or batch and the time zone used for indexing each batch,
- the process, technology, methodologies used to search for the results if not the entire holdings for each custodian per device and any decisions or outcomes of any review of the material,
- the results of any validation exercises carried out, and
- confirmation that the number of files referenced in the load file/s matches the number of files delivered.

As a matter of general principle, the following rules should be followed:

1. Deliveries should be provided on an appropriately sized CD, DVD or Hard Disk Drive (HDD), which should be encrypted to ensure safe transmission.
2. Decryption information should be provided separately on confirmation of delivery.
3. HDDs should be internal SATA variants. Where external HDDs are supplied they should be eSATA or USB3 variants.
4. CDs and DVDs should be formatted to Universal Disk Format (UDF) ISO/IEC 13346.
5. HDDs should be formatted as NTFS.
6. Deliveries should be free from computer viruses and malware.
7. Password protected or encrypted files should either be provided in a decrypted state or with corresponding passwords and decryption instructions.
8. Scanned versions of hard copy documents should be provided as OCRed PDFs that contain all of the document's contents.
9. Individual ***Document Id*** numbers must be unique, should be no more than 30 characters, and should consist of an alphanumeric prefix and sequential numbers with consistent leading zeros e.g. ABC00001, ABC00002 or be in the Australian Federal Court format.
10. ***Document Id*** numbers stamped onto image files according to the jurisdictional requirements of the operation should not obscure any other part of the image.
11. The use of delimiters must be consistent across all deliveries.

12. Load file formats should be consistent across all batches and/or custodians once the sample delivery has been verified.
13. All container files such as PST, NSF, ZIP and RAR should be extracted if delivery is via load file.
14. Each email attachment should be extracted from its parent email and provided as a separate file if delivery is via load file.
15. Grandparent/parent/child relationships within family groups should be maintained and represented in the load file as per below, and
16. Family groups should not be spilt across different deliveries and/or metadata files.

Required Files

The following files must be provided with structured datasets depending on the chosen format:

If providing as a Nuix case file:

File Type	Format/s	Description
Nuix Case folder	Folder	A folder of files that represent a Nuix case. At a minimum will include a folder named "Stores" and an .fbi file
Native Evidence	Various (EO1, AD1, file folders etc.)	Source files for the records for the supplied Nuix case if binaries have not been stored. If binaries are stored the maximum binary size for storage at processing time should be set to 500 MB.

If providing in load file format:

File Type	Format/s	Description
Metadata	.DAT	A text file containing individual document records and associated metadata. This should be UTF-8 encoded.
Native	Various (.DOC .XLS .PDF etc.)	Source files for individual records. The metadata file should include file paths that indicate the location of the native file relative to the metadata file for each record. These files should be contained in a folder named "Natives" which may contain subfolders if needed. The file name for each native file should be the 'document id' number of the document record. An image file of a particular record should only be provided in place of its native where partial LPP is being claimed over the record. Only OCRed PDFs should be provided as image files.

Metadata Load File

The metadata load file is the primary file used for structured data ingestion. The following rules should be adhered to:

1. Themetadata file should be UTF-8 encoded and should not contain a byte-order mark.
2. The metadata file should not use commas, double quotes or other common characters as separators or qualifiers (delimiters) as they are likely to appear in field values. Instead, the following delimiters should be used:

Delimiter Type	Character Code	Character

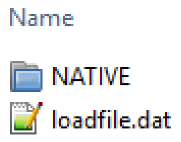
Separator	0020	(can also appear as an empty or solid box)
Qualifier	0254	þ
New line placeholder	0174	®

3. The first line of the metadata file should be a header indicating the names for each column.
4. Column names should be unique and only contain alphanumeric characters and underscores.
5. Each line in the file should represent a single document or record and individual delimited columns should contain the associated metadata, either electronically captured or objectively coded.
6. All values should be qualified by prepending and appending them with a qualifier character.
7. All newline (line break) characters contained in field values should be replaced with a newline placeholder character so that the entire record for a document appears on a single line.

Accompanying Native Files

Native source files that accompany the metadata load file usually consist of individual files such as Word documents, Excel spreadsheets, PDF documents and text files. The following rules should be applied:

1. Native files should be indicated by storing the relative file path and file name of the source file in a column named ***ItemPath*** within the metadata file against its associated record.
2. All container files such as PST, NSF, ZIP and RAR should be extracted prior to delivery.
3. The file path should be relative to the location of the metadata file and should not contain mapped drive letters or additional folders. An example directory structure is shown below:



Name

NATIVE

loadfile.dat

Metadata File Values

A number of factors must be taken into account when generating the metadata file for delivery. Particular consideration must be given to the delivery of email files as they require additional metadata fields and will likely include attachments which must be linked as part of a family group.

The metadata (DAT) file should include the following columns (fields) dependent on the type of files delivered. If additional fields are deemed relevant and included, these should be detailed in a covering note.

The Value Names italicised in the table below must appear the same in the supplied load file.

Value Name	Requirement	Description	Alias
DOCID	All records	Unique ***Document Id*** denoting the beginning of each record	Document Id
PARENT_DOCID	All attachments	**DOCID** of document's *immediate* parent	Host Ref
ITEMPATH	All records	Relative file path to the matching native file being supplied	
ITEM DATE	All records	Derived in the following format: - sent date, - received date, - last modified date, - created date	Document Date Main Date
NAME	All records	Title of the document taken from the face of the document, subject field or file name.	Title Description
DOCUMENT TYPE	All records	Classification of document into representative categories that describe the kind of document, e.g. Email, Chat, Photo.	
SOURCE TIMEZONE	All records	Timezone the material was sourced from. Must also be used for processing the files.	
AUTHOR	All non-email records	Author of source file	
CREATED DATETIME	All non-email records	Date source file was created	
MODIFIED DATETIME	All non-email records	Date source file was last modified	
LAST ACCESS DATETIME	All non-email records	Date source file was last accessed	
TITLE	All non-email records	Title of source file	
FROM	All communication records	From communication address	Sender
TO	All communication records	To communication address	Recipient
CC	All communication records	CC communication address	
BCC	All email records	BCC email addresses	
ATTENDEES	All calendar or meeting records	Attendees name or email addresses	
BETWEEN	All agreement or contract records	Names of parties to an agreement or contract	

ON BEHALF OF	All communication records	On behalf of addresses	
SENT DATETIME	All email records	Date and time email was sent	
RECEIVED DATETIME	All email records	Date and time email was received	
SUBJECT	All email records	Subject of email	
FILEEXTENSION	All records	Extension of native file	
MIMETYPE	All records	MIME type of document	
MD5HASH	All records	The calculated MD5 hash value of the document	
ORIGINAL FILE NAME	All records	Original file name	
ORIGINAL FOLDER PATH	All records	Folder where document was located. May be from file system path or folder in PST	
ATTACHMENTS	Optional	List of **DOCID** values for all attachments to the document. If there are multiple attachments, separate the values with a semicolon	
CUSTODIAN	Optional	Name of custodian	
FILESIZE	Optional	Size of native file in bytes	
LASTPRINTDATETIME	Optional	Date that document was last printed	
LASTSAVEDATETIME	Optional	Date that document was last saved	
PAGECOUNT	Optional	Number of pages in native	

In addition, the following rules should be applied to the data:

1. Multi-level families cannot be represented without including both the **DOCID** and **PARENT** fields. The following table provides an example of field values for a multi-level family:

DOCID	PARENT DOCID
ABC001	
ABC003	ABC001
ABC006	ABC001
ABC010	ABC006

2. If there are multiple email addresses in an email field, the values should be separated by a semi colon (';').
3. If an email address contains a semi colon character, then double quotes should be used to indicate that the semi colon character is content and not a separator.
4. Double quotes should be used when needed in email addresses instead of single quote characters to avoid confusion with names that have apostrophes.
5. The format of emails should be consistent across all email fields.
6. The format of all date-time fields should be consistent across all date fields in the format
7. *YYYY-MM-DD HH:MM:SS*

8. The hour portion in date-time fields should use 24-hour time.
9. If a field cannot be delivered as a date-time value and must instead be delivered as separate date and time fields, the formatting should be *YYYY-MM-DD* for dates and *HH:MM:SS* (24-hour) for time.
10. In such cases, the field names should be like the ***DATETIME** fields (e.g. **SENT DATETIME**) except they should be suffixed with ***DATE** and ***TIME** (e.g. **SENT DATE** and **SENT TIME**).
11. Source time zone is to be supplied in standard UTC offset time Z or +/-hh:mm. The data being supplied should have also been indexed using its source time zone to ensure any records without an inbuilt time zone had its date fields applied correctly.

Sample Delivery File

Sample delivery file has been included in Appendix 2.

The following principles apply:

1. In the example metadata file UTF8 character code ÿ (0255) has been used as a separator in place of 0020 as it is easier to read.
2. The natives folder may contain subfolders for simplified management of large deliveries. A single folder within the natives folder should not contain more than 10,000 files in the one folder.

Databases and Proprietary Systems

There are a number of instances where further discussion should be sought between parties prior to data being delivered. This is particularly the case for data originating from a structured source such as an MS Access or SQL database, an accounting package, or from a proprietary system.

Appendix 1 - Definitions

Document Number

Document number is a standard method of organising legal documents which involves annotating each document or page with a unique identifying number, e.g. DOCID.

Byte Order Mark (BOM)

The BOM is a Unicode character used to denote the byte order of a text file and in some instances how the text is encoded.

Delimiters

A delimiter is a character used to denote the boundary between independent sections of a plain text document.

Family Group

A family group is a collection of one or more child files that are linked to a parent e.g. an email and an attachment.

Metadata

Metadata is commonly referred to as data which provides information about one or more aspects of data or 'data about data'.

Native

Native refers to files that are electronically stored in the format in which they were created e.g. Microsoft Word documents stored as .DOC or .DOCX files and Microsoft Excel spreadsheets stored as .XLS or .XLSX files.

NSF File

A Notes Storage Facility file is a container file format used to store email originating from Lotus Notes software.

NTFS

The New Technology File System is a high performance file system developed by Microsoft and used across its Windows platform.

PST File

A Personal Storage Table file is a container file format used to store email, calendar and other items originating from Microsoft software such as Outlook and Windows Messaging.

SATA, eSATA, USB

Examples of computer interfaces for connecting devices such as hard disk drives.

Universal Disk Format (UDF)

The ISO/IEC 13346 open specification is a vendor-neutral file system for storing electronic data on a wide range of different media.

ZIP and RAR Files

ZIP and RAR files are examples of container file formats used for data compression and are often used when archiving data. A compressed file will typically contain one or more source files. Other examples of compressed container files are ARC, GZIP and TAR.

Appendix 2 – Sample Delivery Files

Metadata File

```
þDOCIDþÿþPARENT DOCIDþÿþITEM DATEþÿþNAMEþÿþDOCUMENT
TYPEþÿþISEMAILþÿþFROMþÿþTOþÿþCCþÿþBCCþÿþSENT DATETIMEþÿþSUBJECTþÿþAUTHORþÿþCREATED
DATETIMEþÿþMODIFIED DATETIMEþÿþTITLEþÿþITEMPATHþ

þABC001þÿþþÿþ2000-01-31 14:20:30þÿþMeeting next
weekþÿþEmailþÿþTRUEþÿþjohn.smith@company1.comþÿþjane.smith@company2.comþÿþmanagers@comp
any2.comþÿþþÿþ2000-01-31 14:20:30þÿþMeeting next weekþÿþþÿþþÿþþÿþþÿþnatives\ABC001.msgþ
þABC002þ

þABC002þÿþABC001þÿþþÿþImportant DocumentþÿþDocumentþÿþþÿþþÿþþÿþþÿþþÿþþÿþþÿþJohn
Smithþÿþ2000-01-21 00:10:04þÿþ2000-01-22 21:13:52þÿþImportant DocumentþÿþNatives\ABC02.docxþ
```