Axcelerate 5.11.0

Incoming Data Specification

Published: 2017-Mar-07

Contents

1 Incoming Data Specification Overview

This document describes commonly used requirements for incoming data, for Axcelerate OnDemand and Axcelerate Cloud or licensed Axcelerate installations. The described examples and fields are typical for a third party production, including:

- Structured Data:

  native, imaged or mixed productions delivered in a structured format, pre-processed by another vendor
- Unstructured Data:

  raw native deliveries of loose files and emails or containers
- Recommind databases: Axcelerate Ingestion or Axcelerate Review & Analysis databases created by customers or vendors with licensed Axcelerate installations.

To ensure the fastest upload, incoming data should include these files:

- Metadata file
- Opticon load file for images
- TIFF images and/or native files
- Extracted text files (one per document)

Note: Deviations from this specification and the structures outlined below may incur additional charges and not be processed in an expeditious fashion. Depending on the actual installation, there may be additional requirements, especially with regard to metadata fields, which have to be specified separately.

1.1 Delivery and Notice for Recommind Hosting

The files should be placed on either a USB hard drive, optical media such as DVD and CD, or with prior approval a secure sFTP site. Advance notice in writing that media is being delivered; cover letters detailing with which database(s) to associate the media and outlining any special load requests is appreciated. If the media is encrypted, passwords should be sent under separate cover.

1.2 Example Reference Files

This example delivery (a third party production) consists of

- A metadata file called documents.csv, in CSV format. For better display, | is used as separator.
- An Opticon file called documents.opt for references to images.
- Four documents with their respective image files.

The documents are delivered in this folder structure:

Main folder	Contained files/folders	Contained folders	Contained files
\Delivery	documents.dat		
	documents.opt		
	\nativefile	\001\	123.doc 555.doc XYZ.doc LMN.doc
	\images	\001\	IMG_11.TIF IMG_12.TIF IMG_31.TIF ... IMG_106.TIF

The content of the reference files is shown in tables here, for better visibility.

1.2.1 documents.csv File

File content	Explanation
BEGDOC|ENDDOC|BEGATTACH| ENDATTACH|TEXTPATH|NATIVELINK	Header
ABC_0000001|ABC_0000002|ABC_0000001|ABC0000007||\nativefile\001\123.doc	First document has 2 pages. Its attachment family starts with the first document page (ABC_0000001) and ends with the last page (ABC_0000007) of the attachment.

<table>
<tr><th>File content</th><th>Explanation</th></tr>
<tr><td>ABC_0000003|ABC_0000007|ABC_0000001|ABC_0000007||\nativefile\001\555.doc</td><td>Second document has 5 pages. It belongs to the same attachment family as the first document.</td></tr>
<tr><td>ABC_0000008|ABC_0000009|ABC_0000008|ABC_0000015||\nativefile\001\XYZ.doc</td><td>Third document has 2 pages. It belongs to the same attachment family (ABC_0000008 - ABC_0000015) as the fourth document.</td></tr>
<tr><td>ABC_0000010|ABC_0000015|ABC_0000008|ABC_0000015||\nativefile\001\LMN.doc</td><td>Fourth document has 6 pages. It belongs to the same attachment family as the third document.</td></tr>
</table>

Documents.opt file

The first reference in a line refers to the respective document listed in `documents.csv`.

References	Explanation
ABC_0000001,CD_001,\IMAGES\001\IMG_11.TIF,Y,,,2	First image out of 2 for the first document. Y marks the first image for a document, 2 is the (optional) number of images for one document.
ABC_0000001,CD_001,\IMAGES\001\IMG_12.TIF,,,,	Second image out of 2 for the first document.
ABC_0000003,CD_001,\IMAGES\001\IMG_31.TIF,Y,,,5	First image out of 5 for the second document.
ABC_0000003,CD_001,\IMAGES\001\IMG_32.TIF,,,,	
ABC_0000003,CD_001,\IMAGES\001\IMG_33.TIF,,,,	

References	Explanation
`ABC_0000003,CD_001,\IMAGES\001\IMG_34.TIF,,,,`	
`ABC_0000003,CD_001,\IMAGES\001\IMG_35.TIF,,,,`	
`ABC_0000008,CD_001,\IMAGES\001\IMG_81.TIF,Y,,,2`	First image out of 2 for the third document
`ABC_0000008,CD_001,\IMAGES\001\IMG_81.TIF,,,,`	
`ABC_0000010,,CD_001,\IMAGES\001\IMG_101.TIF,Y,,,2`	First image out of 6 for the fourth document
`ABC_0000010,CD_001,\IMAGES\001\IMG_102.TIF,,,,`	
`ABC_0000010,CD_001,\IMAGES\001\IMG_103.TIF,,,,`	
`ABC_0000010,CD_001,\IMAGES\001\IMG_104.TIF,,,,`	
`ABC_0000010,CD_001,\IMAGES\001\IMG_105.TIF,,,,`	
`ABC_0000010,CD_001,\IMAGES\001\IMG_106.TIF,,,,`	

2 Specification Details

2.1 Metadata File

The metadata file contains details about the incoming records, such as Bates number, author/recipient or other fields for names, paths, and attachment information. It must be consistent in structure and delimiters.

A metadata flat data file will be provided for each data source, or CSV Merge. Multiple custodians may be aggregated into a single file, provided the file contains a custodian field containing the different values.

2.1.1 Consistent structure

1. Fields will match the fields specified under "Frequently Used Fields" on page 9 in name and content.
2. Delimiters – Concordance default delimiters
   - Text delimiter – “þ”: Hex (FE), Unicode (U+00FE), Decimal (254)
   - Field separator – (not displayable or displayed as "DC4"): Hex (14), Unicode (U+0014), Decimal (20)

   If other delimiters are used, this must be explicitly specified.
3. All rows will contain the same number of delimiters and fields.
4. The multi-value field delimiter is a semicolon (U+003B) and must be consistent across all fields.
5. The first line contains a header row with field names. The table below defines the accepted field name labels for the header row.
6. Extracted text is delivered separately from the metadata file as loose text files, one per document.

2.1.2 Date format

1. Date formats must be consistent across all fields, i.e. the sent date should have the same format as the last modified date, for example.
2. Dates and times can be concatenated into a single field, if nothing else is specified. They may occur in two different fields, if this is required.
3. The default date format is configurable. If nothing else is specified, use `MM/DD/YYYY` and `HH:MM:ss (zzz)`.
4. If a time is not available, such as the estimate date for a coded document, then 12:00 am, or 00:00 should be assigned, i.e. 12/21/1999 00:00.

5. Invalid times or dates or missing times or dates in important date fields will be replaced by `01/01/1901 00:00` by default when they are loaded into an Axcelerate project.

2.1.3 Unique key

1. One field must contain a value unique across the Axcelerate project. Typically this is a Bates number or control number. This should be a unique value for the record across all deliveries.
2. This key cannot have spaces, but any alpha-numeric character and all ASCII characters are accepted, except these: < > & / \ ? * " $ | : , ;
3. A unique key combination is also needed for attachment families. All documents of an attachment family must have the same attachment start and and attachment end number. Otherwise, attachment families cannot be identified.

Example: Unique keys and attachments in documents.csv

<table>
<tr><th>File content</th><th>Explanation</th></tr>
<tr><td>BEGDOC|ENDDOC|BEGATTACH|ENDATTACH|TEXTPATH|NATIVELINK</td><td>Header</td></tr>
<tr><td>ABC_0000001|ABC_0000002|ABC_0000001|ABC_0000007||\nativefile\001\123.doc</td><td>First document has 2 pages. Its attachment family starts with the first document page (ABC_0000001) and ends with the last page (ABC_0000007) of the attachment.</td></tr>
<tr><td>ABC_0000003|ABC_0000007|ABC_0000001|ABC_0000007||\nativefile\001\555.doc</td><td>Second document has 5 pages. It belongs to the same attachment family as the first document.</td></tr>
<tr><td>ABC_0000008|ABC_0000009|ABC_0000008|ABC_0000015||\nativefile\001\XYZ.doc</td><td>Third document has 2 pages. It belongs to the same attachment family (ABC_0000008 - ABC_0000015) as the fourth document.</td></tr>
<tr><td>ABC_0000010|ABC_0000015|ABC_0000008|ABC_0000015||\nativefile\001\LMN.doc</td><td>Fourth document has 6 pages. It belongs to the same attachment family as the third document.</td></tr>
</table>

2.1.4 Frequently Used Fields

Except for the mandatory fields, additional specifications are possible.

Axcelerate Field	Mandatory (M) /Optional (O)	Multiple values possible (Y/N)	Description
Beg Doc	M	N	Beginning control number for document (the unique key used for the data)
End Doc	M	N	Ending control number for document

Axcelerate Field	Mandatory (M) /Optional (O)	Multiple values possible (Y/N)	Description
Beg Attach	M	N	Beginning control number for first page of parent document
End Attach	M	N	Ending control number for last page of last attachment
Location	O	N	File system path or Internet URL, either to the loose file, or to the container the document belongs to (e.g. a PST archive).
Custodian	O	Y	Data's custodian, owner of the files
Document Date	O	N	An aggregated date field based on the following criteria: • For loose files: modification date/time (or creation date/time if last modified date is not available) • For emails: sent date/time (or delivery date/time if sent date is not available) • For attachments: inherits the date/time from the parent email. This field is commonly known as *Sort Date*. The individual dates can also be provided separately and be mapped to the date fields below.
Modification Date	O	N	Last modified date (stored by host file system) ℹ **Note:** This is a file system date, not the application date.
Creation Date	O	N	Creation date (stored by host file system) ℹ **Note:** This is a file system date, not the application date.
Sent Date	O	N	Email date/time sent

Axcelerate Field	Mandatory (M) /Optional (O)	Multiple values possible (Y/N)	Description
Application Last Modified Date	O	N	Last modified date (stored by the application)
Application Create Date	O	N	Creation date (stored by the application)
Document Title	O	N	Aggregated field, based on this information: • *Title* metadata field (if available) or filename of non-email files • Subject for emails The individual title/filename/subject fields can also be provided separately and be mapped to the fields below.
Title	O	N	*Title* metadata field within non-email file
Filename	O	N	Filename of non-email file
Subject	O	N	Email subject
Sender	O	N	Email sender, or sender of a chat message
Recipient	O	Y	Email recipient(s), or chat message recipient(s)
Email CC	O	Y	Email CC(s)
Email BCC	O	Y	Email BCC(s)
Importance	O	N	Email importance flag
Read/Unread	O	N	Email read/unread flag
Author	O	N	*Author* metadata field within non-email file
File Name	O	N	Filename of non-email file
File Extension	O	N	File extension

Axcelerate Field	Mandatory (M) /Optional (O)	Multiple values possible (Y/N)	Description
File Size	O	N	File size in Bytes
Folder Name	O	N	Email folder (i.e. folder within a PST or NSF file)
Message ID	O	N	Internet MessageID for emails. Always use in combination with **References** field for thread detection by header analysis.
References	O	N	References to other items for internet messages. Always use in combination with the **Message ID** field for thread detection by header analysis.
MD5 Hash	O	N	The MD5 hash is based on actual file content and, for emails, on composite of metadata fields for emails. If this field is filed, duplicate detection is possible in the target system.
Store Name	O	N	Name of container file (PST name, NSF name, Opentext database name), including extension
--	O	Y	Review tags or other work product assigned to documents during a previous review. The tags to be mapped must be communicated to Recommind prior to data being loaded. See "Frequently Used Fields" on page 9.
--	O	N	Relative path to text file, e.g. `\textfile\001\123.txt` This field is mandatory if text files are part of the incoming data.

Axcelerate Field	Mandatory (M) /Optional (O)	Multiple values possible (Y/N)	Description
--	O	N	Relative path to native file, e.g. `\nativefile\001\123.doc` This field is mandatory if native files are part of the incoming data. ❗ **Caution:** The path must not contain spaces.

2.1.4.1 Tags assigned during review

There is no specific "Tags" field, but a number of default and custom fields. If tags are required, a custom specification of these fields must be added to this standard specification.

Nested fields should be sent as individual fields with the main field as a prefix.

Example:

```
Responsive
```

```
Responsive - Responsive Type
```

2.2 Native Files

Native files and references to them must meet the following requirements:

1. The incoming metadata file contains a relative path to the native file, the NATIVELINK field (see "Frequently Used Fields" on page 9).
2. Filenames matching a Bates number are acceptable, for example `PROD_006789.xls`.
3. There are no more than 1000 native files per directory.
4. The path to the native file has less than 255 characters, no spaces and only consists of ANSI characters.

2.3 Text Files

Extracted text files and references to them must meet the following requirements:

1. There is not more than one extracted text file per document, with the content of all document pages.

   ⓘ **Note:** Multiple single-page text files for one document are not supported.

2. The character encoding in the text files must be consistent - ideally UTF-8.
3. The incoming metadata file contains a relative path to the extracted text or OCR, in the TEXTPATH field. See: "Frequently Used Fields" on page 9.
4. There are no more than 1000 text files per directory.
5. The path to the text file has less than 255 characters, no spaces and only ANSI characters.
6. Filenames matching a Bates number are acceptable, for example `PROD_006789.txt`

2.4 Images

Images and the Opticon file (*.opt) must meet these requirements:

Black and white images:

single page TIFF
1bit color-depth
Photometric interpretation: MinIsWhite (Black and White)
Compression: CCITT – Group 4
300 dpi (default and recommended)
Byte Order: little-endian
Fill Order: TIFF fill order 1

ⓘ **Note:** Any private tags are ignored during loading/merging.

Color images:

singe page TIFF
24bit color-depth
Photometric interpretation: RGB
Compression: LZW
300 dpi (default and recommended)

ⓘ **Note:** Any private tags are ignored during loading/merging.

The image link is delivered separately from the metadata file in a file following the Opticon load file format specification.

The Opticon load file format is a text-delimited file containing all information necessary to link the image with the database. There is one line entry per image file.

The image file entries must be in correct order, i.e. in the same order as documents occur in the metadata file. Pages must be in the same order as they occur in the documents.

The field delimiter is a comma (U+002C).

Example:

The following is a 5-image Opticon load file example. It details 4 documents with their images.

documents.opt

The first reference in a line refers to the respective document listed in `documents.csv`.

References	Explanation
`ABC_0000001,CD_001,\IMAGES\001\IMG_11.TIF,Y,,,2`	First image out of 2 for the first document. Y marks the first image for a document, 2 is the (optional) number of images for one document.
`ABC_0000001,CD_001,\IMAGES\001\IMG_12.TIF,,,,`	Second image out of 2 for the first document.
`ABC_0000003,CD_001,\IMAGES\001\IMG_31.TIF,Y,,,5`	First image out of 5 for the second document.
`ABC_0000003,CD_001,\IMAGES\001\IMG_32.TIF,,,,`	
`ABC_0000003,CD_001,\IMAGES\001\IMG_33.TIF,,,,`	
`ABC_0000003,CD_001,\IMAGES\001\IMG_34.TIF,,,,`	
`ABC_0000003,CD_001,\IMAGES\001\IMG_35.TIF,,,,`	
`ABC_0000008,CD_001,\IMAGES\001\IMG_81.TIF,Y,,,2`	First image out of 2 for the third document

References	Explanation
ABC_0000008,CD_001,\IMAGES\001\IMG_81.TIF,,,,	
ABC_0000010,,CD_001,\IMAGES\001\IMG_101.TIF,Y,,,2	First image out of 6 for the fourth document
ABC_0000010,CD_001,\IMAGES\001\IMG_102.TIF,,,,	
ABC_0000010,CD_001,\IMAGES\001\IMG_103.TIF,,,,	
ABC_0000010,CD_001,\IMAGES\001\IMG_104.TIF,,,,	
ABC_0000010,CD_001,\IMAGES\001\IMG_105.TIF,,,,	
ABC_0000010,CD_001,\IMAGES\001\IMG_106.TIF,,,,	

2.4.1 Fields in the Opticon Load File for Images

Field	Mandatory (M) /Optional (O)	Description
ALIAS	M	Should match your BEGDOC field (see "Frequently Used Fields" on page 9) for the first page of a record, subsequent lines are the interior pages of the document, up to the next Unique key

Field	Mandatory (M) /Optional (O)	Description
VOLUME	O	This entry is the name of the volume where the image resides. This is typically the volume name of a CD or server.
PATH	M	This is the full path and file name (and extension) of the image. File name and path should only consist of ANSI characters and must have no spaces. They have less than 255 characters.
DOC_BREAK	M	Enter a 'Y' to denote whether this image marks the beginning of a document.
FOLDER_ BREAK	O	A 'Y' denotes that this image marks the beginning of a folder. (not used)
BOX_BREAK	O	A 'Y' denotes that this image marks the beginning of a box. (not used)
PAGES	O	This entry is the number of document pages. (not used)

2.5 Filenames and Paths

Filenames and paths can have any character allowed for filenames in Windows, but must not contain spaces. They should only consist of ANSI characters. The paths must not have more than 255 characters.

2.6 Encoding

Opticon files are ANSI encoded. For all other files, UTF-* encoding is expected.

3 Changes to this Document

Date	Topic title	Text before change	Text after change	Remarks
2015-03-18	"Frequently Used Fields" on page 9		Field list was reworked.	
2015-03-27	"Frequently Used Fields" on page 9		Added *Application Create Date* and *Application Last Modified Date* to field list.	
2015-10-13	"Frequently Used Fields" on page 9	*Email From* *Email To*	*Sender* *Recipient*	
2015-11-05	"Text Files" on page 13	-	There is not more than one extracted text file per document, with the content of all document pages. **Note:** Multiple single-page text files for one document are not supported.	
2017-01-12	"Frequently Used Fields" on page 9	Internet MessageID for emails. -	Internet MessageID for emails. Always use in combination with **References** field for thread detection by header analysis.	
2017-01-12	"Frequently Used Fields" on page 9	-	Added *References* field to the list.	

4 Contact Us

About Recommind

Recommind provides the most accurate and automated enterprise search, automatic classification, and eDiscovery software available, giving organizations and their users the information they need when they need it.

Visit us at http://www.recommind.com.

Support

For support issues on Recommind products, visit the Recommind Ticketing System at https://rts.recommind.com.

Documentation

Find Recommind product documentation, Knowledge Base articles, and more information at the Recommind Customer Portal at https://supportkb.recommind.com. For login access to the site, contact your product support:

- For : SearchSupport@recommind.com
- For : Axcelerate@recommind.com

The Recommind Documentation team is interested in your feedback.

For comments or questions about Recommind product documentation, contact us at documentation@recommind.com.

5 Terms of Use

Disclaimer

This document, as well as the products and services described in it, is furnished under license and may only be used or copied in accordance with the terms of the license. The information in this document is furnished for informational use only, is subject to change without notice, and should not be construed as a commitment by Recommind, Inc., including its affiliates and subsidiaries (collectively, "Recommind"). Recommind assumes no responsibility or liability for any errors or inaccuracies that may appear in this document or any software or services that may be provided in association with this document.

Except as permitted by such license, no part of this document may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the express written consent of Recommind. Information in this document is provided in connection with Recommind's products and services. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document.

EXCEPT AS PROVIDED IN RECOMMIND'S SOFTWARE LICENSE AGREEMENT OR SERVICES AGREEMENT FOR SUCH PRODUCTS OR SERVICES, RECOMMIND ASSUMES NO LIABILITY WHATSOEVER, AND RECOMMIND DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF RECOMMIND PRODUCTS OR SERVICES INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. RECOMMIND MAKES NO WARRANTIES REGARDING THE COMPLETENESS OR ACCURACY OF ANY INFORMATION, NOR THAT THE PRODUCTS OR SERVICES WILL BE ERROR FREE, UNINTERRUPTED, OR SECURE. IN NO EVENT WILL RECOMMIND, THEIR DIRECTORS, EMPLOYEES, SHAREHOLDERS AND LICENSORS, BE LIABLE FOR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR EXEMPLARY DAMAGES INCLUDING, BUT NOT LIMITED TO, LOSS OF ANTICIPATED PROFITS OR BENEFITS.

Recommind may make changes to specifications, and product and service descriptions at any time, without prior notice. Recommind's products may contain design defects or errors known as errata that may cause the product or service to deviate from published specifications. Current characterized errata are available on request. Whilst every effort has been made to ensure that the information and content within this document is accurate, up-to-date and reliable, Recommind cannot be held responsible for inaccuracies or errors. Recommind software, services and documentation have been developed and prepared with the appropriate degree of skill, expertise and care. While every effort has been made to ensure that this documentation contains the most up-to-date and accurate information available, Recommind accepts no responsibility for any damage that may be claimed by any user whatsoever for the specifications, errors or omissions in the use of the products, services and documentation.

Trademarks and Patents

Recommind's underlying technology is patented under *U.S. Patent Nos. 6,687,696, 7,328,216, 7,657,522, 7,747,631, 7,933,859, 8,024,333, 8,103,678, 8,429,159 and 8,489,538*

Recommind, Inc. is the leader in predictive information management and analysis software, delivering business applications that transform the way enterprises, government entities and law firms conduct eDiscovery, enterprise search, and information governance. Recommind, Axcelerate, Axcelerate Cloud, Axcelerate OnDemand, and CORE’s name and logo are registered trademarks of Recommind, Inc.

Copyright