Catalyst Insight

877.557.4273
catalystsecure.com

DATA AND LOAD FILE SPECIFICATIONS

Contents

Fast Track Requirements and Load File Specifications

This document outlines the requirements of the Fast Track Data Upload program and the general MDB load file specifications for Catalyst Insight. Catalyst's ability to ingest load files is highly dependent on their included data and format. Please note that in the event that nonconforming data is submitted Catalyst will not attempt to correct the load file. Catalyst will instead wait for a corrected load file to be submitted or until a specific request to correct the load file has been received. All manual work will be billed at the standard rate.

Basic Fast Track Requirements

Requirements include:

- Load files must be in a MS Access database format (MDB or ACCDB).
- Documents must be compressed in a ZIP or RAR format.
- ZIP/RAR files must be smaller than 3.99GB.
- The MDB and ZIP/RAR files must have the exact same file name and are case sensitive (vol001.mdb and vol001.rar, for example).
- Use a dash (-) or underscore (_) if punctuation is needed - do not use spaces or any other punctuation in the file names. Do not use additional periods other than for the file extension (i.e. use very_large_load.rar instead of very.large.load.rar).
- The file extension for the MDB and ZIP/RAR files should either be all lower case or all upper case, but not a mix of cases.

Beg/End Control (or Beg/End Bates) Numbers

Beginning and ending control numbers (or Bates values) should be less than 20 characters. If suffixes are required (e.g., -001, -002, etc.), then the begcontrol/Bates values should not exceed 16 characters to allow for 4 suffix characters (i.e., dash, underscore, period and three numeric values).

Beginning and ending control numbers (or Bates values) should not contain any special characters, including spaces, underscores or dashes. This is because these characters can interfere with the ability to perform accurate range searches on the site.

Parent/Child Relationships

The Parent and Child documents (attachments) should be linked together using the BegAtt and EndAtt fields. The begcontrol of the Parent document should be in the BegAtt field, and the endcontrol of the last Child document should be in the EndAtt field. All documents within the same attachment range (from the Parent to the last Child) need to have the same exact BegAtt and EndAtt values.

Fields in Load Files

It is important that the same fields are provided in the load file from upload to upload. For the first delivery, a mapping will be created by Catalyst. The mapping creates a relationship between the fields in the data file to the fields on the site. Every time a new load file is provided in a different format than previous deliveries, Catalyst must create a new mapping for that type of load file. If instructions are not provided to Catalyst detailing how to handle the new fields, the upload process will be halted until the specific information has been provided. Please be sure to only provide fields that exist on the site. If a field does not already exist on the site, the mapping will fail. There is no limit to the number of load file mappings a site can have.

One Load File per Volume

When providing documents for upload to Insight, we require two files per upload—an MDB or ACCDB load file, and a ZIP or RAR file containing the documents to be loaded. There should always be one load file provided per ZIP/RAR volume.

File and Folder Names and Organization

The requirements in this section are critical to complete your upload in a timely manner. If the data delivery does not meet all of these requirements then there will be significant delays in completing the upload and you may be billed for the associated technical time.

Special Characters

File and folder names should not include characters other than alphanumeric characters or dashes. Special ASCII characters should not be used. For example, do not include ampersands (&), commas (,) or apostrophes (‘) in file or folder names. If paths to files include other types of characters, we may have to rename the files and/or folders, adding additional time and cost to the upload process. All files should be named after the begcontrol [or begbates] number (e.g., begcontrol.ext) and the naming of the files is case sensitive.

- **File Name Length:** The name of a file should not exceed 25 characters as various processes may need to be run against the files.
- **Path Length:** The path for a file should not exceed 255 characters, including the name of the file.
- **Folder Organization:** A single folder should not contain more than 10,000 files as this will impact processing of the files. If a folder contains more than 10,000 files then there may be speed impacts when reviewing the files in that folder.

File Organization

All files that correspond to a record should have an associated field in the load file that includes the full path and filename for the specific file. For example, if a record will have a native file and a text file loaded, then there should be a field called “Filepath” and a field “Textpath” in the loadfile. The data in that field will point to the specific file as it appears in the uncompressed ZIP/RAR file.

Below are two common delivery configurations that illustrate how each record has one entry in the load file that references the two potential file types.

Native File Example

BegControl	Filepath	Textpath
000001	Files001\000001.DOC	Files001\000001.TXT

Image File Example

BegControl	Filepath	Indexpath
000002	Files001\000002.PDF	Files001\000002.TXT

Extraneous Files

Please be aware that any files included in deliveries will be included in storage costs, regardless of their use on the site. For example, OCR text files that are not used because native files are being indexed and viewed, or files that do not have corresponding metadata records, will be billed even though they are not accessible on the site.

Zero-Filled Dates

Catalyst handles incomplete dates that are zero-filled (those with 00 for the month and/or day and/or 0000 for the year) as follows:

- Any date field with a value of 00/00/0000 or 00000000 will be made NULL/empty.
- Any date with 00 for month or day will be made NULL/empty.
- Any date with 0000 for the year will be made NULL/empty.

The deleted date information is stored in the XML file in the “indexissuedetail” field. By keeping the invalid date data, the document will still come back in results when a search is run against it using the invalid dates.

Catalyst can accept true date and time data into the date fields, but for optimized search functionality we ask that any date/time fields be split into separate date fields, and separate time fields.

Document Format Guidelines

Native Files

Catalyst will attempt to index the native file and no text file is required. If this indexing is successful then search hit highlighting will be available on the preview of the native.

Compressed files must be exploded and processed prior to uploading. System and container files cannot be indexed or viewed through the site.

Metadata should be extracted from the native files and submitted with a corresponding load file.

OCR or Text Files

When delivering files that are searchable in the native format (.DOC, .XLS, etc.), extracted text files are not required because Catalyst will index the native file. If the files are not searchable (.TIF, .JPG, etc.) then OCR text files are needed in order to make the documents searchable within Insight. Requirements for delivering text files are as follows:

- Load files should not contain OCR or extracted text (i.e., text should not be submitted in a field within a Concordance DAT file). OCR or extracted text must be submitted as separate text files.
- Text files must be in multi-page format (one text file per document, not one text file per page). Catalyst cannot accept single page text files.
- Page breaks in the text files are preferred but not necessary.
- Text files must have UTF-8 encoding to ensure proper indexing.

TIFF Files

Catalyst Insight accepts single-page TIFF files, but in order for TIFF files to be searchable, text files must also be delivered. Single-page TIFF files must be loaded manually and cannot be loaded via the Fast Track system. These files must be accompanied with an additional load file, either an IPRO .LFP file or an Opticon. OPT file to indicate the document breaks.

Multi-page text files must be delivered with the single-page TIFF files (single page text cannot be loaded). The text files should be named to match the first page of each document, such as ABC001.TIF.TXT. If the text files do not contain the full TIFF file name (including the .TIF extension) plus .TXT then the files will be indexed but not visible on the site. Regardless of file naming convention, the text files must be delivered within the same folder as the image files. The associated text should not be included within the load file.

PDF Files

All PDF files must be optimized for fast web viewing or “linearized.” There are three types of PDF files, with unique instructions for each:

- PDFs with embedded text: This is the Catalyst preferred format for images. These PDFs have embedded text in the PDF file. They are created either from scanned images run through an OCR process or created from an electronic source file.
- PDFs with associated OCR text files: In this format, the OCR text is delivered in a separate file with the same name as the PDF file (ABC001.PDF and ABC001.TXT, for example).

- PDFs with embedded images only: PDFs without embedded text or associated OCR text files will not be full-text searchable. A user will have to rely on the searching of metadata in order to find these documents in the repository.

Multi-Language Documents

Multi-language documents must also be in UTF-8 format.

Delivery of Coding for Documents

When delivering document coding to be loaded to the site for fields where the data will be mapped to radio buttons, checkboxes, drop down lists or multi-select fields, there are specific formatting requirements. These requirements are as follows and only apply to editable fields:

- The values in the data should EXACTLY match the search facets for the fields on the site. This includes matching case and punctuation.
- If a value doesn't match one the existing search facets then it may or may not display on the site. The new value will not be added the field and may cause problems in searching on that field.
- Multi-valued data must use the semicolon to separate values within a field.

Fast Track—Media

Fast Track Data Media Delivery

When the volume of data is simply too large to transfer via FTP there is an option to have this data uploaded from media. If this option is selected, there are manual steps involved that lengthen the time of upload. In addition to the delivery time, there is the manual staging of the data, and a manual step where the Fast Track Data Upload is initiated.

The data on the media should not be compressed.

- The organization of the documents on the media should be reflected in the path fields in the load file.
- The delivery should include one master MDB/ACCDB load file for every 250,000 files.
- Send an email to support@catalystsecure.com and copy your Catalyst Project Consultant. In the subject line include: "Your Company Name / Site Name: Fast Track Data Upload Delivery."
- Attach to the email a completed Data Transmittal Log.
- A confirmation email will be sent indicating that the data has been received.
- Once the staging has completed and the manual step to initiate the Fast Track Upload has been performed, the upload will continue as a regular Fast Track Upload. As with the regular FTP Fast Track, a failing load file will require contact with Catalyst as to how to proceed. A replacement MDB/ACCDB load file can be provided to your Catalyst Project Consultant.
- Please use the address below when shipping the media to Catalyst. If the data will be received after normal business hours or on a weekend, then please instruct the courier to call the number listed below when making the delivery.

Catalyst Repository Systems
ATTN: [insert Catalyst Project Consultant name]
1860 Blake Street – 7th Floor
Denver, CO 80202
(303) 824-0911