Chapter 5: Importing Data Files

You may import data files using the SEER*DMS application or by auto-loading the files from a network location. Files that you import must adhere to specifications documented in the Imports section of the Help menu. Once the data are imported, you may use the Import Manager to review summary information about each import and to access the data files.

Some types of imported data are persisted in the database; other types of data are processed but not persisted. Persisted data are stored in the record table of the SEER*DMS database and can be viewed within the SEER*DMS record editor. You may specify the Import ID in the Data Search, Worklist, or Import Manager to search for persisted records or to determine whether the records have completed the workflow. Most supplemental imports are not stored in the database after they are processed. Registry configurations determine whether supplemental data are stored in the database and the methods used to process these data (see the *Processing and Storage of Imported Data* section of this chapter for more information).

Mass change imports are unique. Mass change data are used to make direct updates to the database. An audit log entry is made in the patient set, record, AFL, facility, or person (contact) that is updated via mass change. Data in mass change files are not stored in the record table or processed via the workflow. Mass change imports are fully documented on the Imports help page within SEER*DMS.

In this chapter, you'll learn about

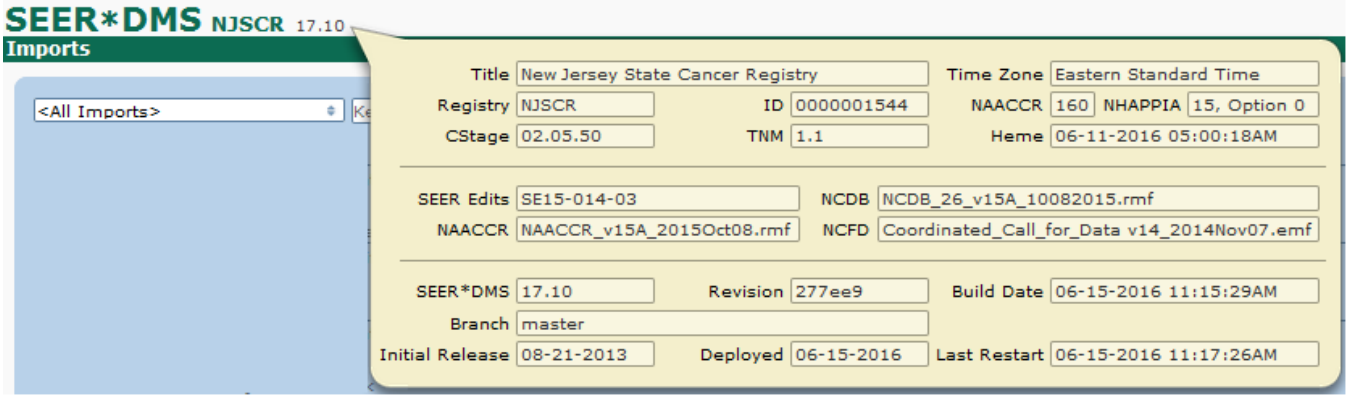
- Import Specifications and Algorithms
- Processing and Storage of Imported Data
- Import Manager
- Importing Data Files within SEER*DMS
- Auto-Loading Data Files
- Reviewing Status and Outcomes of Imports:
    - Import Information Page
    - Import Summary Report
    - Reviewing Field Frequencies
    - Consolidation Summary – Updates Tab
- Import Review Task
- Finding Records and Tasks Related to an Import
- Reports Related to Data Imports
- Technical Implementation

Import Specifications and Algorithms

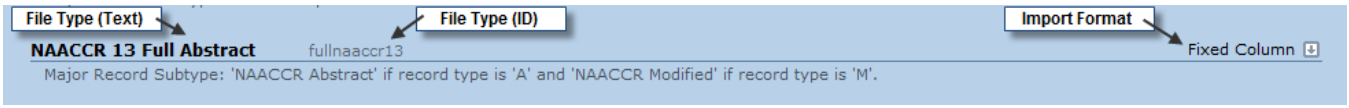
SEER*DMS supports the following Import Formats: Fixed Column, HL7, Delimited, Mass Change, and Image Files. Standard imports include the NAACCR abstract layout, NAACCR HL7, NDI Cause of Death, and the Social Security Administration linkage file. Registry-specific import types are created for other imports including state-specific death certificate data files, disease index data, motor vehicle data, pathology report data in custom file layouts, and other types of data.

SEER*DMS supports NAACCR versions 12, 13, 14, 15, and 16 of the abstract record layout. The NAACCR 16 import will be used for NAACCR 16.0, 16.1, etc. The most recent version of NAACCR currently supported by SEER*DMS is documented in the SEER*DMS system information popup. To open the popup, click the SEER*DMS version number in the top header. In the screenshot shown

below, the user clicked on the SEER*DMS version number in the header (17.10) and the NAACCR version is shown as 160. This indicates that NAACCR version 12.0 through 16.0 are supported in version 17.10 of SEER*DMS.



The file layout and configuration settings for imports are documented on the Imports help page. The File Type (in full text and as a system ID), Import Format, and a brief description are displayed for each import type.



Double-click the File Type or click the down arrow on the right-hand side to expand the documentation section. The Mass Change section contains a detailed description of the file format, validation, and entities that can be updated via Mass Change. The Linked Image file type does not have a documentation section because there is no specific file layout (any type of file can be imported as a linked image). The help sections for all other import types include the following:

- **Major Record Subtype -** The record subtype determines the routing of the data in the workflow and other processes. For example, NAACCR abstract is the subtype for all versions of NAACCR abstract data (NAACCR 12.x, 13.x, 14.x, 15.x, 16.x). Record types are defined in the Records section of *Chapter 2: Records and Patient Sets*. The types of records are listed in the *lkup_record_major_subtype* database table.
- **Line Length** (Fixed Column) – The number of characters per line in fixed column imports. If there are multiple lines per record, this is the number of characters in the first line of the record (e.g., NAACCR Extended imports have one line of coded data fields followed by separate lines for long text fields). There is no length restriction for the text lines in a multi-line import like NAACCR Extended.
- **Number of Fields and File Layout** (Delimited Files) – This section defines the number of fields, the method used to embed the delimiter character and quotes (Microsoft Excel or standard), the character used as the delimiter, and whether a single or double quote is used as the quote character. If the file includes a header line then "Ignores First Line" is set to true. This describes the two methods for handling special characters in delimited files:
    - Excel
        - All values may be enclosed in quotes, but quotes are only required if the delimiter character, newline characters, or quotes are embedded in the value.
        - To embed the delimiter character or newline within the field, surround the entire value in quotes.
        - To embed a quote within the field, surround the entire value by quotes and any quote in the value must be double-quoted

        - If you enclose a field in quotes then you may not have any characters between the delimiter and the quotes.
    - Standard
        - To embed the delimiter character within the field, precede the character with a backslash
        - The quote character is not a special character, all quotes between delimiters are considered part of the value
        - All values will be trimmed
- **Import Review Task** – A manual Import Review task may be limited to data that generate errors, warnings, or have duplicates. Or an Import Review task may be created for all imports of this type.
- **Processing** – This configuration setting determines whether the imported data are processed in the workflow or in import routines external to the workflow. Refer to the *Processing and Storage of Imported Data* section of this chapter for more information.
- **Stored in Record Table –** This configuration setting determines whether data of this type are stored in the record table of the SEER*DMS database (see the *Processing and Storage of Imported Data* section of this chapter for more information).
- **Algorithmic Duplicate Checking –** A matching algorithm that identifies records which may not be an exact duplicate of previously loaded records, but provide the same data as previously loaded records. The criteria used in the matching algorithm are documented on the Matching help page. This is one method for identifying records that are not exact duplicates but do not provide any more data than a previous loaded record; partial hash checking is a second and more commonly used approach to the same problem.
- **Fields Ignored for Partial Hash –** A hash representation of the original import data is created for each record. This hash is permanently maintained in the database. This hash is compared to the hash of existing records to prevent the import of duplicate records. Some fields are ignored when creating the hash. The help page indicates which fields are included in the hash and which are ignored when the hash is created.
- **Table of Import Fields –** This table shows the mapping from the import file to the SEER*DMS database. The columns of the table vary by import type, as described below.
    - Import Field – A variable name is assigned to each field in the import file. This column is shown for all imports except HL7 imports (the import field is identified in Conversion Rule for HL7 imports).
    - Field Position – Start Column, End Column, and Length are shown for Fixed Column imports. Index and Max Length are shown for Delimited files.
    - Database Field – The database field (defined by *table.fieldname*) to which the import field is mapped.
    - Ignored for Partial Hash – Indicates whether or not a field is ignored when creating the hash.
    - Conversion Rule – Processes performed during the import are documented here. This includes rules to “convert” (reformat) the data and rules to validate the fields. Warnings and errors generated during the import and reported in the Import Review task are also documented. Conversion rules for specific fields are documented in the Conversion Rule column of the table; conversion rules that do not apply to a specific field or are applied to multiple fields may be documented below the table.

Processing and Storage of Imported Data

Each import algorithm includes a setting for *Processing*; this parameter determines whether the imported data are processed in the workflow or in import routines external to the workflow. The *Processing* field has two possible values as described below; the settings used in your registry are documented for each import on the Imports help page. (Note: There is no processing parameter for Mass Change imports.)

- **Processing = Workflow**. By definition, certain data types must be processed through the workflow. Abstracts, casefinding records, death certificates, and other types of records that need to be screened for reportability must be processed via the workflow.
- **Processing = Non-workflow Import Routines**. When this setting is used, data are only sent into the workflow if manual review is required or the record is persisted. The majority of supplemental imports are processed in procedures external to the workflow (e.g., the records are matched using a pre-processing matching algorithm that is not part of workflow processes). This reduces the load on the system at the time of the import.

Stored in Record Table is an import configuration setting. Typically, supplemental import data are not stored in the record table unless a manual task is necessary to review the results of the automatic match task or resolve conflicts encountered during auto-consolidation. If SEER*DMS is able to match and consolidate the data without a manual task, the patient set is updated and the Import ID, filename, and line number are recorded in the patient set's audit log. However, SEER*DMS does support other options related to data storage. The settings used for your registry's imports are documented on the Imports help page.

- **Stored in Record Table = All data**. Imports processed via the workflow must use this setting (this includes abstracts, casefinding, death certificates, and other records that are screened for reportability). This setting is rarely assigned to supplemental imports.
- **Stored in Record Table = Not stored unless manual review required**. This is the recommended setting for supplemental imports.
- **Stored in Record Table = Matches and data that require manual review**. This option may be used for supplemental imports if all data items must be retained as a reference. The import data are matched against the database in import routines. If a match is found, the data are auto-consolidated into the patient set, a record is created and the record is linked to the matching patient set. In addition, a record is created if the results of the automatic matching or consolidation tasks require review.
- **Stored in Record Table = Non-matches and data that require manual review**. This option may be used for supplemental imports, however, it may result in a large number of unlinked records that may be listed in manual match tasks and database searches. The import data are matched against the database in import routines. A record is created for each of the non-matches. Matching data are auto-consolidated; records are not created for data that can be matched and consolidated in automatic tasks. A record is created if the results of the automatic matching or consolidation task require review.
- **Mass Change**. Direct database updates made via mass change are documented via audit log entries. The original mass change file can be accessed via the Import Info page.

As noted above, supplemental data are stored in the database if a manual matching or manual consolidation task is created. Deterministic algorithms are recommended for supplemental matching (criteria resulting in "possible" matches are rarely defined for supplemental data). Therefore, in most registries, a manual supplemental matching task is only created if the incoming data is a "perfect" match to two or more patient sets. The matching algorithms used in your registry are documented on the Matching help page. A manual consolidation task is created if the incoming data conflict with the patient set's data and an auto-consolidation "failure" is created. Auto-consolidation rules and failures are documented on the Auto-Cons help page.

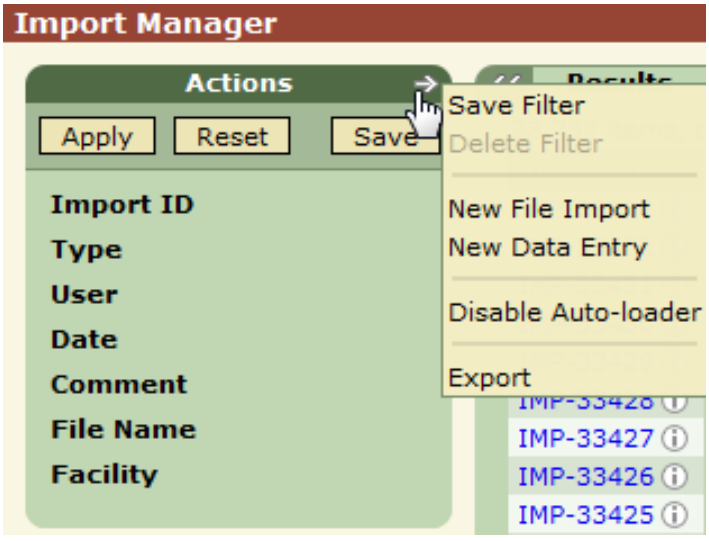
Import Manager

Requires system permission: *import_electronic* or *import_manual*

You can use the Import Manager to search for imports by date, file type, facility, filename, the user who initiated the import, and other fields. You can then open the Import Info page for a specific import. This allows you to view or download the original data files, review field frequencies on the Statistics tab, check the workflow status of imported records, review the records in the Data Search, and review warning messages that may have been generated when the data are imported.

To access the Import Manager, select **System > Imports and Data Entry**. The following data columns are shown in the Import Manager:

- **Import ID** – A unique ID assigned to each import. Click the ID to view details about the import as described in the *Import Information Page* section of this chapter.
- **Import Date** – Date and time that the import was initiated. This is the timestamp recorded when the data are analyzed and may be earlier than the date and time that the records are loaded.
- **Facility** – The facility selected by the user or set in the autoloader configuration file for this import. This facility is used by default as the record's reporting facility when the reporting facility is not available on the record.
- **User** – The user who initiated the import. "seerdms" will be listed as the user for autoloader imports and data migrated at the time of deployment.
- **Records** – The number of records imported. This value will be zero until the Import Review task is completed and the data are successfully loaded.
- **Type** – The import type is displayed in this column. The import algorithm is shown if it is a file import. The other possible values are Data Entry or Migrated Data.
- **Comment** – The import-wide comment entered on the Import Source page when starting a new import or entered in the <comments> tag in the autoloader XML.
- **Action** – Click **Lookup** in the Action column to search the database for records loaded in the import. This is simply a short-cut to the SEER*DMS Data Search.



The Actions menu in the manager allows you to:

-Save or Delete the Filter – you can save and name your filter settings; or you can delete a saved filter. SEER*DMS filters are described in Chapter 3.

-Start a **New File Import** (*import_electronic* permission required)

-Start a **New Data Entry** session (*import_manual* permission required)

-Disable or Enable Auto-loader (this menu item is a toggle) – This requires the *system_administration* permission. Files copied to the autoloader folders will be ignored by the system when the autoloader is disabled. Those files will be imported when the autoloader is re-enabled.

Use the filters on the left side of the screen to search for an import. SEER*DMS filters are described in more detail in *Chapter 3: Using SEER*DMS.* Importing Data Files within SEER*DMS Requires system permission: *import_electronic* and *contact_view*

You may use the SEER*DMS interface to upload data files from any local or network disk drive that you can access. The files will be uploaded through your Web browser and are subject to HTTP size limitations. If you are uploading a huge file, then you should upload a compressed version of the file. The SEER*DMS uploader will accept Zip or GZip files.

Importing Data Files within SEER*DMS

*To import data files using the SEER*DMS interface:*

1. Select **System > Imports** to access the Import Manager.
2. Select **New File Import** from the **Actions** menu.
3. Specify the import **Facility** by typing a facility ID (the FAC prefix is not required), typing the facility name, or by using the lookup to select a facility. You also have the option of specifying a **Contact** at that facility (not required).
4. Select the **Import Type**. The drop-down list includes all file imports defined for your registry. Once you have selected an import type, click the information icon to review the file specifications.
5. Enter notes describing the import in the **Comments** field (optional). This comment will be written to the record or patient set audit log if this import results in an automatic update. For example, this comment will be written to the audit log of data modified by mass change and it will be written to the audit log of data modified via auto-consolidation.
6. Add files to the import.
    a. To do this, you may drag-and-drop files from Windows Explorer onto the drop zone. Or you may click the **Browse** button or the drop zone to browse. You may upload a set of files in a Zip archive. SEER*DMS will unzip and add each file from the archive. You may also upload a file compressed with GZip.
    b. SEER*DMS will do a preliminary evaluation of each file. If the file is a duplicate of a file from a previous import, a warning message will be displayed. This is only a warning; there are rare situations in which you may choose to upload a duplicate file and ignore this warning (it is recommended that you consult with IMS technical support staff in these situations).
    c. To remove a file from the import, click the Remove link in the Action column.
7. To import additional files that are different file formats, but from the same facility:
    a. If you would like to import files that are different Import Types, check **Upload additional import types**. You will be able to initiate a new import; the values entered in the Import Source section of the screen will be retained.
8. Click **Import**.
9. SEER*DMS will analyze the data files. If the results of the analysis require manual review, an Import Review task will be created and must be completed in order to load the data. A manual review is required if there are errors, warnings, or duplicates in an import; and a manual review is always required for some import types (this is determined by the Import Review Task setting for the import; parameter settings are shown on the imports Help

page). Refer to the *Import Review Task* section of this chapter for instructions on completing this task. The Import Review Task provides an error summary and frequencies of all data fields. Once you review these statistics, you will have the option of loading the data or terminating the import for some or all data files.

10. If you are importing files of different Import Types, check the "Upload additional import types" box at the bottom of the screen. Add files and revise comments, if necessary.

Auto-Loading Data Files

SEER*DMS will automatically load data files from a designated location on the SEER*DMS server. That server location is called the "autoloader". The autoloader is a folder on the server with sub-folders organized by import facility. The autoloader should be used for "lights out" data transfers; that is, fully automated transfers of files. Login to SEER*DMS and use the Import Manager to manually move data files to SEER*DMS.

Typically, a registry uses an automated script to move files from the registry's network to the autoloader. A secure transfer mechanism (SFTP or SCP) is used to move the files from a registry computer that is connected to the SEER*DMS server via a VPN connection. The IMS technical team can assist registry staff to implement the transfer script.

The autoloader can be disabled via the Actions menu in the Import Manager. If the autoloader is enabled, an automatic workflow task will import files in the main auto-loader directory and the sub-directories within the autoloader directory. Each folder must contain its own copy of a configuration file (autoloader.xml). This file defines the import settings (Import Facility, Contact, Comments, and File Type for each file). The Import Facility and File Types are required; all other elements are optional.

This is the Document Type Definition (DTD) for the autoloader.xml:

```
<?xml version="1.0" encoding="UTF-8"?>

<!ELEMENT autoloader (organization,representative?,comments?,file-type)>
    <!ELEMENT organization (#PCDATA)>
    <!ELEMENT representative (#PCDATA)>
    <!ELEMENT comments (#PCDATA)>
    <!ELEMENT file-type (when*)>
    <!ATTLIST file-type default CDATA #REQUIRED>
    <!ATTLIST file-type rec-type CDATA ''>
        <!ELEMENT when (#PCDATA)>
        <!ATTLIST when ext CDATA #REQUIRED>
        <!ATTLIST when rec-type CDATA ''>
```

- **organization** – This element is required. Specify the Import Facility using this element. You may specify a complete facility ID or the organization name. It is recommended that you use the facility ID. If you specify the name then the text must be an exact match to a name in the SEER*DMS Facility List.
- **representative** – This element is optional. You may specify a complete contact ID or the contact's name. It is recommended that you use the Contact ID.
- **comments** – This element is optional. These are import-wide comments that are displayed in the Comment column of the Import Manager.
- **file-type** – This section is required. Use this section to specify the File Type of files to be loaded via this folder. To do this, you must associate an import File Type (ID) with one or more extensions. The File Type ID is case sensitive but the file extension is not (*.pdf and

*.PDF are equivalent to the autoloader). You must also assign a default File Type for unknown extensions. As shown below, the import help page includes the File Type ID for each import available in your registry.

- **rec-type –** This element is optional. You may specify the major record subtype for a linked image import. Use the short name specified in the *lkup_record_major_subtype* database table. If rec-type is not specified, the value defined in the import algorithm will be used (this is specified in the XML system file for the linked image import).

In the following example, all files placed into this autoloader folder will be autoloaded as Linked Image files. Image Data Entry tasks will be created and the screen layout for Path Rpt records will be shown by default. The user may select a different record type in the data entry task.

```
<?xml version="1.0"?>
<!DOCTYPE autoloader PUBLIC "-//IMS//DTD Autoloader Definition//EN"
"autoloader.dtd">
<autoloader>
   <organization>FAC-0001</organization>
   <comments>Autoloaded via the linked image folder for FAC-0001.</comments>
   <file-type default="linked-image" rec-type="path_rpt">
   </file-type>
</autoloader>
```

The example below allows different file types in the same folder and uses a default file type for undefined extensions. A *.pdf file will be uploaded as a linked image file. A *.txt file will be uploaded as an HL7 NAACCR file. All other files will be uploaded as NAACCR 13 data. Although this example shows the use of one folder for multiple types of files, that is an unusual use case. Typically, a single folder is used for all abstract records from a specific facility and a different folder is used for all HL7 records from a facility.

```
<?xml version="1.0"?>
<!DOCTYPE autoloader PUBLIC "-//IMS//DTD Autoloader Definition//EN"
"autoloader.dtd">
<autoloader>
   <organization>FAC-0001</organization>
   <comments>This file was autoloaded.</comments>
   <file-type default="fulllnaaccr13">
      <when ext="pdf">linked-image</when>
      <when ext="txt">hl7-naaccr</when>
   </file-type>
</autoloader>
```

Technical Specifications

- Permissions for files: rw-rw-r-- or 664 (permissions for folders are given below)
- WinZip files will be auto-extracted. The extracted file or files will be autoloaded.
- All subdirectories must have an autoloader.xml file to support automated pickup. If a directory is missing the autoloader.xml file, no file will be imported from that directory (but any directory with an xml file within that directory will still be processed).
- On a regular basis, the Quartz scheduler fires off an automatic workflow task to gather any files residing in the autoloader folders and import them into the system. The frequency of the directory scan is determined by the *system.schedule.autoloader* configuration parameter. This uses a standard crontab string to set the schedule.
- The autoloader will wait for a stable directory tree before it starts processing any files and directories under that tree. If the autoloader detects a change in the directory tree, it will

sleep for n-seconds before trying to scan the directory tree again (as specified in the *importer.autoloader.quiet_time* configuration parameter). This technique prevents the autoloader from starting to load files before all the import files have been copied into the proper directory. A single Import ID is assigned to all files loaded in a single task.

- If an error is generated, the file is not loaded and it is copied to the folder defined by the *importer.autoloader.failure.dir* configuration parameter (typically, the folder is named *.errors*). An email notification is sent to each user who is a member of the role defined in the *importer.autoloader.email_role* configuration parameter.
- Imports loaded via the autoloader are attributed to the user defined by the *importer.autoloader.user_name* configuration parameter. This is set to seerdms by default.

Autoloader Folders

You should create autoloader folders to meet your needs based on these considerations:

- Permissions for folders: rwxrwxr-- or 774 (group must have executable permission)
- The main autoloader directory is defined by the *importer.autoloader.dir* configuration setting.
- Any number of subdirectories can be created within the main autoloader folder.
- A folder is ignored by the autoloader if its name starts with a period (".").
- Copies of files should be placed in the autoloader folders. The registry should maintain an archive of the data files in another location.
- An import facility must be specified for each import in SEER*DMS. When using the autoloader, one facility must be specified in each autoloader.xml configuration file. Therefore, you must create a separate folder for each Import Facility. If you want to specify different facilities then you can create folders by facility.
- The level of complexity that you create is up to you. Registries are encouraged to experiment and devise a plan during the beta-testing period or on the registry test server.

Reviewing Status and Outcomes of Imports

After an import is started, information about its status and outcomes can be viewed on the Import Information Page.

Import Information Page

Requires system permission: *import_electronic* or *import_manual*

To access the Import Information page, select **System > Imports and Data Entry** and click the ID of an import shown in the Import Manager.

The Import ID is shown at the top of the left panel, followed by the Import Summary and the Current Status.

- **Import Summary:**
  - Number of files in the import.
  - Number of records in all files.
  - Number of valid records in all files. This is the number of records loaded during the import process. Valid = (Records – (Errors + Duplicates + Ignored))
- **Current Status:**

- Number of records in the database. This is the number of records that are persisted. This number will be less than the number of valid records in the Import Summary if not all records were persisted or if imported records were deleted. Click the link to review the records in the Data Search.
- Number of records per workflow task. Click the link for each task type to view the tasks in the worklist.

The center panel displays the following information about the import.

- **Facility** – The facility selected by the user who initiated the import or set in the autoloader configuration file. This facility is used as a default value for reporting facility when reporting facility cannot be set using data fields on an imported record.
- **Contact** – This is an optional field that may be used to identify a representative at the import facility who may be contacted about data provided by that facility.
- **Import Type** – If it is a file import then the import algorithm is displayed; the other possible import types are data entry and migrated data.
- **Import Date** – Date and time that the import was initiated. This is the timestamp recorded when the data are analyzed and may be earlier than the date and time that the records are loaded.
- **User** – The user who initiated the import. "seerdms" will be listed as the user for autoloader imports and data migrated at the time of deployment.
- **Comments** – The import-wide comment made by the user who started the import or entered in the <comments> tag in the autoloader XML.

The files tab is displayed below the Info panel. The following information is listed for each file on the Files tab that is displayed on the Import Info page and in Import Review tasks.

- **Filename** – Click the filename to open or download the file. You may also click the "Download All" link. Note: if you are logged into your test server, the filename will only be a link if the file was imported on the test server.
- **Records** – The total number of records in the data file. The word "stopped" will be shown in this column if the import was terminated because the number of errors exceeded 200.
- **Errors** – The number of errors in the file. An import error is generated when a field fails import validation or when the record is not formatted correctly. A limited number of fields are validated during the import process; the field-level validation for each import is documented on the Import Help page. The number of errors will be a hyper link if the value is greater than zero. Click this value to review the error messages. SEER*DMS stops processing the import file when the error count exceeds 200.
- **Warnings** – The number of warnings for the file. Import warnings are generated for a limited number of data fields. These include invalid values for some date fields, and warnings when a text field is truncated in order to meet database constraints. Refer to the Import Help page to see which fields are validated; and whether that validation results in an import error or warning. The number of warnings will be a hyper link if the value is greater than zero. Click this value to review the warning messages.
- **Duplicates** – If two records within the same import are exact duplicates, one will be loaded and the other will be counted here. This count also includes records which are not loaded because they are exact duplicates of records loaded in an earlier import (based on hash values); and records which are not exact matches but provide the same data as previously loaded records. Non-exact duplicates may have differences in non-essential fields but provide the same data in key fields. Refer to the Imports Help page to determine if non-exact duplicate checking is performed for a specific import type; if used, it will be implemented as "Algorithmic Duplicate Checking" or as "Fields Ignored for Partial Hash". To

determine the specific types of duplicates, click the number of duplicates and review the messages. The types of duplicates are:

  - File-dup – records that were not loaded because they were exact duplicates of records in the same import (the duplicate may be in the same file or in a different file within the same import)
  - Hash-dup – records that were not loaded because they are full or partial hash duplicates of records loaded in an earlier import
  - Alg-dup – records that were not loaded because they are a duplicate as determined by Algorithmic Duplicate Checking defined for this import.

- **Ignored** – To reduce the number of extraneous supplemental records in the database, some data types are evaluated to determine if they provide useful data. A record that is deemed as not useful is "ignored" (it is not processed further and is not loaded into the record table). Rules to ignore records are defined in importer algorithms of certain import types. For example, some supplemental record types are ignored if social security number is null (e.g., this rule applies to CMS records in some registries). Importer rules are documented on the Imports Help page.
- **Valid** – The number of records which can be loaded. Valid = (Records – (Errors + Duplicates + Ignored)). If you complete the import and this is a data type that is persisted, this is the number of new rows that will be created in the record table and the number of records that will enter the workflow.
- **Summary** – Click view to open a CSV report that shows the import summary of each data line. This report indicates whether or not the line of data was processed. If the data line was persisted in the database then this listing provides the Record ID that was assigned. If the data line was fully processed in import routines then this listing indicates the outcome of those processed. Please review the Import Summary Report section of this chapter for a more detailed description.

Import Summary Report

The Import Summary Report is a CSV report generated as each file is processed. The results are static and only show the results of import processing; to review workflow outcomes use RPT-008A or review the workflow item history.

To open the Import Summary report, click **View** in the Summary column of the Import Info page. The report includes the columns listed below.

- **Line Number** – The line number in the data file.
- **Import Status** – This will either say "imported" or it will provide a reason that the data line was not imported. The possible values are:
  - IMPORTED – Data line was imported.
  - ALG_DUP – There is an Algorithmic Duplicate matching algorithm for this record type; and the data line is a duplicate of a record based on that matching algorithm.
  - ALG_IGNORED – Data line was ignored because of a rule that is implemented in import procedures.
  - DATA_IGNORED – Data line was ignored because of a rule defined in the import layout (XML file).
  - FILE_DUP – The data line is an exact match of another line in the same import. It may be a FILE_DUP of a line in the same file or another file in the import.
  - HASH_DUP – The hash value of this data line matches the hash value of a record in the database.

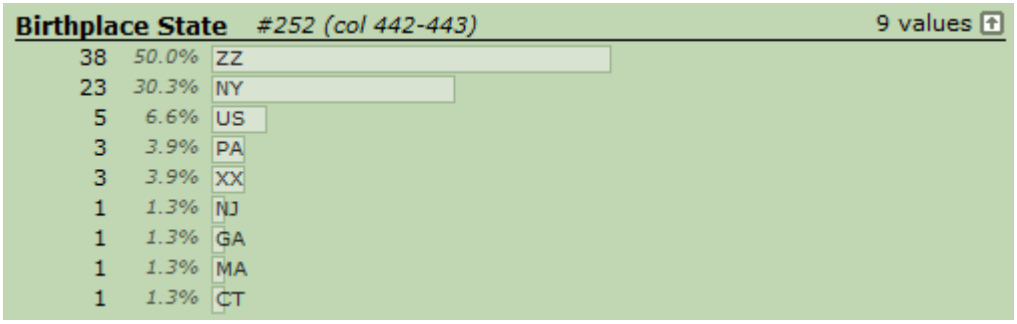
- **Persisted Record** – The record ID assigned to the data. A value will be shown if the data line was loaded into the record table (whether or not the record was deleted later).
- **Non-workflow Status** – The final status of the record if it was processed in import routines. The possible values are:
    - NO_MATCH
    - MULTIPLE_MATCH
    - AUTO_CONS_FAILURE
    - AUTO_CONS_SUCCESS
- **Matched Patient Set** – This column will only contain a value if an attempt was made to auto-consolidate the record in import routines. This column will be blank for any record that is processed in the workflow rather than import routines.
- **Comment** – Text field containing other information. For example, the number of changes made during auto-consolidation will be listed here.

Reviewing Field Frequencies

Requires system permission: *import_resolution*

It is critical that imported data are valid. Some import files need to be pre-processed, while others can be imported as is. If there were problems with the data preparation, either at the sending facility or within the registry, an invalid import file could incorrectly modify a large volume of system data. This type of problem is very difficult to remedy once the data are loaded into SEER*DMS; each record immediately begins to move through the workflow and affect changes in patient set data. It is the responsibility of registry staff to define and implement comprehensive quality control procedures to prevent import errors. This should include but not be limited to the review of input field frequencies to identify invalid or unacceptable values. Field frequencies are shown on the Frequencies tab in Import Review tasks and Import Info screens.

The fields listed on the Frequencies tab are database fields mapped directly to import fields and database fields assigned via import conversion rules. The fields are listed in alphabetical order. You may use the linked letters at the top of the list to view fields which begin with that letter or use your browser's search tools to search for a fieldname.



The figure above shows the type of information shown for each data field. The following describes the information that is displayed and the text actually used in the example above.

- **Field name** - shown in bold above the line: **Birthplace State**
- **NAACCR Item Number and/or File position –** If this is a NAACCR data item, the item number will be displayed (#252 in the example above). If the field is mapped directly to an import field, the import field's position will be displayed as column position for fixed column

imports or field index for delimited imports. (File position is not displayed for fields that are assigned based on conversion rules; and file position is not shown for HL7 fields.)

- **Number of different values** – only shown if there are more than five different values for the field. The number of different values ("*9 values*" in the example) is displayed on the far right of the page, just above the line. Up and down arrows enable you to toggle the display to show all values or hide all but the top five values. The figure above shows all 9 values.
- **Statistics and bar chart -** For each value encountered for the field: the number of records, percent of records, and value. The value is shown within a histogram bar. In the example, 38 records have a value of "ZZ" for Birthplace State. 38 records equals 50% of the number of valid records in the file. Note: frequencies are not shown if the number of differing values exceeds display limitations. For example, it would be impractical to show a frequency for street address which would be unique for every record; instead, "too many values" will be displayed.

Consolidation Summary – Updates Tab

The Updates Tab is displayed if the imported data are processed in import routines rather than the workflow. Typically, Supplemental imports are the only imports processed in import routines (this is described in the *Processing and Storage of Imported Data* section of this chapter). This tab shows a summary of the import's match results and of updates made to patient sets via automatic and manual consolidation of the import's data. The values shown on the Summary Tab are:

Patient Sets Updated. This is the total number of Patient Sets modified via the consolidation of this import's data. This includes updates made in automated and manual consolidation tasks.

No Matches. This is the number of records in the import file with a match score of 0 for all patient sets. These records were not used in consolidation and did not create a manual task to review possible or multiple matches.

Manual Match Tasks. This is the number of records in the import file that triggered manual match tasks (Supplemental Match tasks). This count and the number of manual consolidate tasks give an indication of the amount of manual labor associated with an import. A manual match task is required if, based on registry algorithms, a person must review the results of the automated match. Registry management should reconsider the matching algorithms if supplemental data result in large numbers of manual match tasks. Deterministic algorithms are recommended for supplemental matching (criteria resulting in "possible" matches are rarely defined for supplemental data). Therefore, in most registries, a manual supplemental matching task is only created if the incoming data is a "perfect" match to two or more patient sets. A manual task would be created if the supplemental matching algorithm includes criteria that results in scores greater than 0 but less than 1000; and the only matches to an incoming record are patient sets meeting those criteria.

Manual Consolidate Tasks. The number of Consolidate FUP tasks triggered by data in this import. A manual Consolidate FUP task is created if the incoming data conflict with the patient set's data and an auto-consolidation "failure" is created. Auto-consolidation rules and failures are documented on the Auto-Cons help page.

Counts of Updated Fields in Patient Set Data. This is a list of Patient Set fields modified via the consolidation of this import's data; and the number of changes per field. For some fields, the count may exceed the total number of Patient Sets Updated, for example, the count for Survival Time may include changes to the field on 2 or more CTCs within the same Patient Set.

Import Review Task

Requires system permission: *import_resolution*

Once an import is initiated, SEER*DMS will analyze the data files and create a manual Import Review task prior to loading the data into the database. The Import Review Task provides an error summary and a listing showing the frequency of each value for all data fields.

To complete the Import Review tasks:

1. Click the **Import Review** link in the **My Tasks** section of the home page. Click the Task ID to open the task. At this stage, the files have been checked for errors and analyzed to create frequencies of values in all data fields. The records have not been loaded into the system. (Note: Import analyses are considered stale after one week. You will not be able to complete the import if the analyses are stale. Click re-analyze.)
2. Carefully review the number of Records, Errors, Warnings, Duplicates, Ignored, and Valid in the Import Files section of the page. Each of these fields is described in the *Import Information – Files Tab* section of this chapter.
3. If an incorrect file type was specified, select the correct **Import Type**. An asterisk is displayed next to the original type selected when the import was initiated (the field frequencies shown in the Frequencies tab are based on the layout for the original type).
4. If the import type is correct, review each file with errors or warnings:
    a. Click the number displayed in the errors or warnings column.
    b. Review the messages. Determine whether you want to reject the entire file or import records that did not cause errors (records with warnings cannot be excluded).
    c. Click **Close** to close the Import File Problems window.
    d. Set the appropriate **Action** for this file in the left panel.
        i. Select *Reject* to reject all records in the file (records in other files will not be affected).
        ii. If all problems are caused by duplicate records, you may use *Accept valid only* to accept the non-duplicate records. If there are other types of problems, this action is not recommended.
5. To close the Import Review task:
    a. If an Action is specified for each file, click **Import** to load the acceptable data and exit the task. If the data cannot be loaded or a new file type was selected for a file, a second Import Review task will be created and assigned to you. Otherwise, SEER*DMS will begin to load and process the data. An email message will be sent to you when the import step is complete. The data will then be processed in the workflow or in import matching and auto-consolidation routines. The Import Info page and system reports can be used to review the status of imported data.
    b. If you would like to close the Import Review task without importing any of the files, click **Reject All**.
    c. To exit without removing the task from the workflow, click **Cancel**. The task will be assigned to your user account.

Finding Records and Tasks Related to an Import

Each imported record travels through the workflow, triggering the automated and manual "tasks" that must be performed to process the data. There are a variety of tools in SEER*DMS for determining the current location of a record in the workflow and the end result of the processing. Use these tools to determine whether there are open tasks related to the records and if the records have been consolidated into the patient set data.

To view a summary of open tasks related to the import:

1. Select **System > Imports and Data Entry**. The most recent imports will be listed on the first page of the manager by default. All data entry sessions, autoloader imports, imports initiated in the system interface, and migration imports will be listed in the manager. You may use the filters to search for a specific import. The filters correspond to the fields described in the *Import Manager* section of this chapter.
2. Click the **Import's ID**.
3. The **Current Status** will be displayed in the left panel.
4. Click one of the links to view or access the tasks via the worklist. Clicking the "Records in Database" link is simply a shortcut to the data search.

To search the worklist for open tasks related to the import:

1. Select **View > Worklist**, or click the **Worklist** link on the Home page.
2. Use the **Import** filter to find tasks related to records loaded in a specific import. Some filters may be pre-set when you enter the worklist. Clear all filters other than the Import filter.
3. Click **Apply**.
4. Record-based tasks for records loaded in the import will be listed. When a record moves through the workflow and is linked to a patient set, the focus of the task switches from the record to the patient set. These patient set tasks will not be listed in the workflow for this import; use RPT-061A to obtain a complete list of tasks initiated by records in an import.

To review a sample of records loaded in a specific import:

1. You may initiate your search from the Import Manager or the Data Search:
    a. To search from the Import Manager:
        i. Select **System > Imports and Data Entry.**
        ii. Click the **Lookup** link listed in the **Action** column for the import of interest.
        iii. A record Data Search will be executed using the import's ID as a search field.
    b. To initiate the search from the Data Search:
        i. Select **View > Data Search**.
        ii. Set **Search Type** to Record Search.
        iii. Specify the Import ID in the Import filter.
        iv. Click **Apply**

Reports Related to Data Imports

Select **View > Reports** to review the current list of system reports provided in SEER*DMS. The reports listed below are the system reports related to data imports that were included in SEER*DMS when this manual was printed. If you require information regarding imported data that is not included in a system report, you may generate an external report. See *Chapter 24: Creating Reports and Extracting Data* for more information.

Report ID	Title	Description
RPT-008A	Status of Imported Records	The linkage status of each record in an import. It includes import information (line number, file name, etc); record information (ID, type, facility, event date); linkage status (deleted, linked at patient level, linked to CTC, unlinked); and if it was used to update a record or patient set.
RPT-009A	Imported Records	Use this report to generate a listing of records in a specific import or a listing of records imported on a specific date.
RPT-009B	Records Submitted by Facility	A listing intended to be sent to the reporting facility showing the records provided by that facility during a specific time period.
RPT-061A	Workflow Location of Imported Records	For each record imported, this report lists the current location of the record in the workflow.
RPT-070A	Import Summary	This report gives summary information for each import including file names, number of errors, warnings, duplicates; and the total number imported
RPT-085A	Mass Change Report (Records)	Changes made to records, contacts, or patient sets via mass change Imports. These reports query the audit log tables of records, contacts, and patient sets to identify changes made via direct updates defined in mass change imports.
RPT-085B	Mass Change Report (Contacts)	
RPT-085C	Mass Change Report (Patient Sets)	

Technical Implementation

The SEER*DMS import module provides a means to process large quantities of data in a known format. The module is designed to accept a wide variety of file types. This is an obvious necessity given the nature of data streams entering a cancer registry. The types of data received at cancer registries are ever changing; and the file layouts of standard types are updated on a regular basis. To meet this requirement, the SEER*DMS import module is based on an XML construct. Each XML file defines the import specifications and algorithms described previously in this chapter, the file layout, database mapping, formatting rules, and conversion rules.

There is a separate XML file for each import type. Each XML file serves two primary functions: it tells the SEER*DMS application how to load and process the data file; and it provides the help text displayed in the SEER*DMS online help system. If you go to Help > Imports, you will see a listing of every import type in your registry. There is one XML file for each of these imports. The help system provides a readable view of the XML. The actual XML code can be viewed in the System

Administration module (the *system_administration* permission is required to access system files, as described in *Chapter 27: System Management*).

There are five categories of imports, as listed below. Four of these categories are implemented via XML and there is a separate document type definition (DTD) for each of these categories. Mass change is unique in that it is the only import type that is not XML-based. The requirements for mass change imports are consistent for all registries; therefore, mass change imports are implemented via Java classes.

- Fixed Column
- Delimited
- HL7 - Pathology Laboratory Reports
- Linked Image
- Mass Change

There are six main sections in the Fixed Column and Delimited XML files:

- **derive-from** – determines whether sections of the import definition are "derived" or inherited from a base import. For example, the implementation of a registry's NAACCR v16 import will often include registry-specific logic. The import logic shared by all registries for NAACCR v16 is maintained in a base import. Each registry's import inherits the logic of the base import and extends it with registry-specific rules.
- **major-subtypes** – rules to set the major subtype of records created by the import. Major record subtype is used throughout SEER*DMS, for example, it determines the path of a record through the workflow. An auto-coding polisher may change the subtype of a record after it is imported.
- **partial-hash** – defines the fields to exclude (or include) in the partial hash. This section would show the "exclude" fields if that list is shorter than the "include" list. The hash is saved in the database. This hash is compared to the hash of existing records to prevent the import of duplicate records. The import help page includes a column indicating which fields are excluded from the partial hash.
- **file-characteristics** – this section provides the location of the fields in the data file; the mappings between the fields in the data file to database fields; defines formatting rules (eg, zero-left-pad); and associates conversion-rule with fields, as necessary.
- **conversion-rules** – scripts to implement validation and import conversion rules. An example of an import conversion rule is the conversion of a string to a date object; parsing addresses, etc. This is not used for significant changes or recoding of data items; significant changes to data values are done in auto-coding polishers so that an audit log entry is created. The conversion rules are implemented in the Groovy scripting language. The index determines the order in which the rules are executed. The scripts can use Java utility methods. The utility methods provide the ability to share logic among rules, use more complex logic, and allow the scripts to use persisted data items.
- **extra-information** – this section can be used by IMS staff to provide additional help text, not covered by the text provided in the other sections of the XML. To view the help text for an import, go to Help > Imports.

An example of a registry-specific NAACCR v15.0 configuration file is shown below (some sections were abbreviated to save space). As you see in the derived-from section, this import is derived from the generic NAACCR 15 import (fullnaaccr15). Note that only one field is listed in the file-characteristics section. The definitions for all other fields are inherited from the base import. The logic for major-subtypes, conversion-rules are also inherited. Registry-specific rules for partial-hash and the parsing of the State Requestors data item are defined in this registry's import.

```
<?xml version="1.0"?>
<!DOCTYPE fixed-column-file-type SYSTEM "FixedColumnType.dtd">

<fixed-column-file-type id="ut-naaccr-15">
    <derive-from alg-id="fullnaaccr15" derive-types="yes" derive-hash="no"
             derive-mappings="yes" derive-rules="yes" />

    <!-- Subtype rules are entirely derived from the NAACCR 15 import type -->
    <major-subtypes/>

    <partial-hash exclude = "Record_Type, Registry_Type,…"/>

    <file-characteristics>
        <line-length>22824</line-length>
        <import-field name="State_Requestor_Items" start-column="2340"
            end-column="3339" dest-system-field-name="stateRequestorItems"
             trim="none" rule-id="parse-state-requestor-items"  />
    </file-characteristics>

    <conversion-rules>
        <conversion id="other-state" index="6" >
            <doc>Copy RECORD_REGISTRY.FROM_OTHER_STATE_REGISTRY from import facility.</doc>
            <script>
                <![CDATA[
                    if (record.sourceSubmission != null &&
                          record.sourceSubmission.sendingFacility != null)
                        record.registryData.fromOtherStateRegistry =
                           record.sourceSubmission.sendingFacility.state
                ]]>
            </script>
        </conversion>

        <!-- This rule override the one from NAACCR to parse UT-specific items -->
        <conversion id="parse-state-requestor-items" index="6" >
            <doc>
                <![CDATA[
                    Parse the following fields from stateRequestorItems:<br/>
                        2340 - Used for major record subtype.<br/>

                    If vendor name starts with 'SRABS':<br/>
                        2343 - 2352 - RECORD_EXTENDED.FOLLOW_UP_CONTACT_PHONE_NUM<br/>
                        2353 - 2354 - RECORD_EXTENDED.FOLLOW_UP_CONTACT_PHONE_OWNER
                ]]>
            </doc>
            <script>
                <![CDATA[
                    if(!SeerStringUtils.isBlank(record.stateRequestorItems)) {
                        if (record.vendorName != null &&
                            record.vendorName.toUpperCase().startsWith('SRABS')) {
                            record.followUpContactPhoneNumber =
                            SeerStringUtils.trim(record.stateRequestorItems.substring(3,13))
                            record.followUpContactPhoneOwner =
                           SeerStringUtils.trim(record.stateRequestorItems.substring(13,15))
                        }
                    }
                    record.stateRequestorItems =
                      SeerStringUtils.trim(record.stateRequestorItems, false, true)

                    if (record.stateRequestorItems != null &&
                             record.stateRequestorItems.length() == 0)
                        record.stateRequestorItems = null
                ]]>
            </script>
```

```
        </conversion>
    </conversion-rules>
</fixed-column-file-type>
```

The linked image XML file is the simplest. It provides a mechanism to associate an image file (PDF, JPG, etc) with a record type. The main section in the linked image import's XML is major-subtypes. For information on processing linked images, see *Chapter 6: Data Entry*.

There are five main sections in the HL7 XML files:

- **derive-from** – determines whether sections of the import definition are "derived" or inherited from a base import. For HL7 imports, this allows registries to derive from standard imports like the NAACCR HL7.
- **message-structure** – defines the overall structure of the HL7 message in terms of the segments. This syntax lists the valid segments, indicates whether a segment can repeat, and defines the required order of the segments.
- **partial-hash** – defines the fields to exclude (or include) in the partial hash.
- **mapping**– this section provides the location of the fields in terms of the segment ID and field number; the mappings to database fields; and the conversion rules.
- **extra-information** – this section can be used by IMS staff to provide additional help text, not covered by the text provided in the other sections of the XML. To view the help text for an import, go to Help > Imports.

An excerpt from the HL7-NAACCR XML file is shown below. This excerpt shows the message-structure and partial-hash sections; and one example of a mapping rule. To view the complete XML file, please go to System > Administration. Select System Files in the Module drop-down. Select HL7_NAACCR_ImportType.cfg.xml in the Files drop-down. Refer to the NAACCR website (www.naaccr.org) for more information on the message structure for HL7 pathology reports.

```
<?xml version="1.0"?>
<!DOCTYPE hl7-file-type SYSTEM "HL7_Type.dtd">

<!--
   Implementation based on "Pathology Laboratory Electronic Reporting Version 4.0"
      http://www.naaccr.org/StandardsandRegistryOperations/VolumeV.aspx
-->

<hl7-file-type id="hl7-naaccr" name="HL7 NAACCR" review="errors-warnings-only">

  <message-structure regex=
       "MSH(,SFT)?,PID(,NK1)*(,PV1)?(,ORC)?(,OBR(,NTE)?(,OBX(,NTE)?)+(,SPM(,OBX)*)*)+(,DSC)?"/>

   <!-- all MSH fields should be included here; refer to #7113 before making changes to this list --->
   <partial-hash exclude="MSH-1,MSH-2,MSH-3,MSH-4,MSH-5,MSH-6,MSH-7,MSH-8,MSH-9,MSH-10,MSH-
11,MSH-12,MSH-13,MSH-14,MSH-15,MSH-16,MSH-17,MSH-18,MSH-19,MSH-20,MSH-21"/>

   <mapping id="sending-fac">
      <desc>MSH-4 is copied into [:sendingLab:].</desc>
      <body>
         <![CDATA[
            record.sendingLab = message.getSegment('MSH').getFieldAsString(4)
         ]]>
      </body>
   </mapping>

...
```

```
...
...
...
</hl7-file-type>
```