eScholar Guide to Extracting Data

eScholar
Expand Knowledge — Improve the Future

Table of Contents

INTRODUCTION

PURPOSE

This document is intended to provide guidance on extracting data from source systems into eScholar template format. This information can be used to create individual extraction routines or a system of extraction.

DEFINITIONS, ACRONYMS, AND ABBREVIATIONS

ASCII: American Standard Code for Information Interchange. ASCII is a code for representing English characters as numbers, with each letter assigned a number from 0 to 127. For example, the ASCII code for uppercase *M* is 77. Most computers use ASCII codes to represent text, which makes it possible to transfer data from one computer to another.

EBCDIC: **E**xtended **B**inary-**C**oded **D**ecimal **I**nterchange **C**ode. Pronounced ***eb****-sih-dik,* EBCDIC is an IBM code for representing characters as numbers. Although it is widely used on large IBM computers, most other computers, including PCs and Macintoshes, use ASCII codes.

ISO Format: A date format defined by the International Standards Organization; the ISO Date Format is YYYY-MM-DD and is the required date format within eScholar

NCES: National Center for Education Statistics. NCES is the primary federal entity for collecting and analyzing data related to education. NCES publishes online handbook that offer guidance on consistency in data definitions and in maintaining data so that they can be accurately aggregated and analyzed.

THE ESCHOLAR TEMPLATES

OVERVIEW

The eScholar templates represent the format that data files must match in order to be processed by the eScholar load plans that move data into the data warehouse.

TEMPLATE CHARACTERISTICS

- The templates are represented as worksheets in an Excel workbook
- There are three types of workbooks:
    - Data Warehouse templates
    - Lookup Database templates
    - State-specific templates
- Each template maps directly to a Data Warehouse, Lookup Database or State-specific table

THE INDEX WORKSHEET

- The Index worksheet contains the list of templates included in the workbook and the order they are presented in the workbook
- The template order is alphabetical order (new with v7.1)

THE INFORMATION WORKSHEET

- The Information worksheet contains general information about the templates and a description of the contents of each template worksheet. The following are described:
    - NCES linkage
    - Minimum Compatible Version defined
    - Code column values explained
    - Overview of file formats supported
    - Explanation of Load Sequence information
    - File naming convention defined (see Template File Naming Convention section below)

INDIVIDUAL TEMPLATE STRUCTURE

- Each template includes the following header information
  - Template Number
  - Load Plan Number
  - eScholar Version
  - Minimum Compatible Version
  - Associated Data Warehouse or Lookup Database Table
- Each template includes the following columnar information
  - Field Number
  - Field Start and End Position (for creating fixed-width files)
  - Field Length
  - Scale and Data Type information:
    - Scale of any decimal columns (0 for integer columns)
    - If there is no information in this column, the field could be character or date
    - All date fields have a field name ending in **DATE**
  - Code:
    - **K** - field is a component of the logical key,
    - **U** - fields is updateable through the load plan based on matching logical key
    - **M** – field is mandatory
    - **R** – field is recommended for reporting but is not mandatory
    - How the Code values affect the building of an extract file is discussed further in the Creating a Record section below
  - NCES Info – relevant NCES Element ID, if applicable
  - Lookup Table
  - Sample Data
- Each template includes a Business Rules section
  - Any relevant rules that must be followed in creating the data file are defined
  - Specific file formats supported by the template are identified in the last business rule in each section (see Template File Formats Supported section below)
- Each template includes a Load Sequence section
  - This section defines any pre-requisite files that must be load prior to loading the template file
  - Some pre-requisites are optional and these are noted within the Load Sequence section
  - All Lookup tables are optional and so Lookup tables are specifically identified

TEMPLATE FILE FORMATS SUPPORTED

- Each template contains a business rule on the specific file formats that are supported for that template
  - This business rule is generally the last rule within the Business Rule section
  - Some templates support both delimited and fixed width options; others support only delimited

  - For templates that support a delimited file format, the following delimiters are supported:
    - Comma
    - Tab
    - Pipe
    - Additional delimiters may also work but have not been formally tested
  - For templates that support fixed width options, some support both ASCII and EBCDIC formats and others support only ASCII.
- So, a general recommendation is to create extracted data files in tab-delimited or comma-delimited format. These formats are supported by every template.

TEMPLATE FILE NAMING CONVENTION

- The Naming convention for template-formatted files is as follows:
  - DistrictCode_PrimaryTargetTable_YYYYMMDDHHMM.xxx
  - District Code is the eScholar-supplied value
  - Primary Target Table is identified on line 5 of each template
  - The Date and Time information is required to handle processing more than one file for the same template per day
  - xxx represents the file extension.
    - csv – comma-delimited
    - tab – tab-delimited
    - del – other delimiter
    - txt – ASCII flat file
    - fil – EBCDIC flat file
- The naming convention is described in the Information worksheet within each template workbook.
- The extracted file must conform to the naming convention in order to be processed by the automation tool.

TEMPLATE FILE DEPENDENCIES

- Files may have dependencies
  - Dependencies are noted at the bottom of each template in the Load Sequence Section.
  - Some dependent tables are optional – if this is the case, the table will be identified as optional in the Optional column within the Load Sequence Section
  - By following the dependencies, extractors can come up with a prioritized list of when to work on each template

CREATING EXTRACT FILES

RESOURCES AVAILABLE

- Data Dictionary
  - If available, a Data Dictionary provides *client-specific* information on the elements to include in the extract files
  - The Dictionary contains a definition of each element to be collected, including any required values
  - The definitions may refer to existing data collections
  - The Dictionary also includes the mapping of the elements to the eScholar template and field
- Template Field Names
  - eScholar template Field Names can also be used for an initial level of understanding of the data represented by the field, especially in the absence of a Data Dictionary
- NCES Element IDs
  - Each template contains references to NCES element IDs, where applicable. If there is a match between an NCES definition and what is intended for the eScholar warehouse field, the NCES element ID is provided.
  - The appropriate NCES handbook to use as reference for the element number is provided in the Rules section of each template, usually in the third to last rule.
- eScholar Best Practices
  - eScholar's Best Practices documents contain *general* field by field descriptions and recommendations for all elements in all templates
  - The documents are divided up into data domains and the Overview section of each document contains a list of the included templates.
  - Some templates are associated with more than one domain. For example, the Staff Attendance template is included in both the Staff domain and the Attendance domain. The information is identical in each section.
  - FAQs are provided, if applicable.

DISTRICT CODE

- This value is a required field in every record of every template-formatted file
- This value is supplied by eScholar using state-specific guidance
- Any extraction tool will need to have this value available to insert into each record, if the value is not already stored in the source system

SCHOOL YEAR

- Many templates contain a School Year Date field.
- A school year in eScholar begins on July 1st and ends on the following June 30th
- The school year is represented as the last day of the year, i.e. June 30th.
- For example, the school year that begins on July 1, 2004 and ends on June 30th, 2005 would be represented by the value 2005-06-30 and would be referred to as the '2005 school year'.
- Since School Year Date is a date field, the format must be ISO: YYYY-06-30.

DATE FORMAT

- Any date field include in any template-formatted record must be in ISO format: YYYY-MM-DD. This includes the full ten characters, including the dashes between Year and Month and Month and Day.
- There are no exceptions to this rule: any record containing a date field that is not in ISO format will be rejected by the eScholar load plan
- All date fields are identified on the templates with a Field Name ending in DATE

FIELD LENGTHS

- Field lengths are identified in each template in the Length column. This information is critical because the eScholar load plans will reject any record which contains a field that is greater than the designated length. Obviously this is of particular importance with delimited files: if the number of positions between delimiters is greater than the designated length for that field number for any field in the record, the record will be rejected.
- When creating a delimited file, there is no need to use all of the positions – it is not necessary to pad with blanks, for example
- When creating a fixed-width file, every position must be accounted for
- Do not use leading spaces with any character fields

TEXT DATA

- Text Data
  - For ease of reading reports generated using the data warehouse, we recommend that text fields be formatted with initial caps.
- Text Qualifiers

- If you are using a delimited format, in particular comma-delimited, it is important to use an appropriate text qualifier to handle situations where the delimiter is part of the field value. For example, some name fields are designated to be in the form Last Name, First Name. Enclosing the value within a qualifier such as **"** will enable the value to be successfully loaded.

LOOKUP TABLES

- eScholar uses Lookup Tables to translate coded values into a uniform set of descriptions (for example, the Ethnic Code Lookup Table translates Ethnic Code values such as 1 or A into uniform descriptions, such as African-American or Asian-American.
- Lookup Tables that are associated with warehouse fields are denoted within the warehouse templates in the Lookup Table column
- Lookup Tables also have their own set of templates
- The Extract Tool may need to create Lookup Table files, in addition to warehouse files
- Where the warehouse template indicates that the field will be filled from a lookup table, it is our recommendation that you do let the lookup table value be used, rather than populating both the code and the description.

CREATING A RECORD

- When creating a record, each field that is coded with a K in the templates must be populated, with a few exceptions. These exceptions are noted in the Best Practices documents and/or in the rules at the bottom of each template.
- When creating a record, each field coded with an M is mandatory and must be populated. The eScholar load plans will reject records where M fields are not populated.
- When creating a record, each field identified with an R is considered valuable for district-level reporting. The eScholar load plans will *not* reject records where R fields are not populated.
- When creating a record based on a template, it is not necessary to populate every field. Since the eScholar complete data warehouse is based on one model there will almost certainly be fields that do not apply to a given client.
- However, each field from the template must be *accounted for* in the record. So, if you are building a delimited file, you must have a delimiter for each field in the template, even if there is no value in between delimiters. Similarly, if you are building a fixed-width file, you must account for the positions of each field, even if there is no value at those positions.

HOW THE ESCHOLAR LOAD PLANS WORK

DETERMINING THE CORRECT PLAN

- The file is matched up with the appropriate load plan based on the standard file name
    - See the Template File Naming Convention section above for a description of the file naming convention. The same description is included in the Information worksheet of the template workbooks.

BASIC FORMAT CHECKING

- Are the field lengths correct? – if any fields would be truncated then the entire record is rejected
- Are the data types correct? – for example, if there is any text data in a numeric field, then the record is rejected
- Are the date formats correct? – if any date field is not in ISO format, the record is rejected

DATA INTEGRATION CHECKING

- If any key fields or other fields need to be present in a parent table, the load plan will perform the necessary integration checks.
    - For tightly-coupled tables, the integration check is strict. Any integration check that fails will cause the record to be rejected
    - For loosely-coupled tables, the integration check is not as strict In these cases, the record is not rejected but a warning message is generated instead.
- The integration checks are described in detail in the Best Practices documents.

DATA TRANSFORMATIONS

- Next, the load plan will perform any necessary transformations
    - Lookup tables
    - Derived fields
    - If a corresponding Lookup Table value is not found, the record is not rejected. Rather, the code value is copied into the description field.

INSERT VS. UPDATE PROCESSING

- At this point, the record has passed all the required checks and is ready for database processing. There are two kinds of processing: insert and update
- The file record is compared to the warehouse table to see whether the record is already in the warehouse. This is performed based on logical key – the fields in the template coded with a **K**.
- If no match if found, insert processing is performed
    - The record is prepared for final insertion and then inserted into the warehouse. This may be done in batches for enhanced performance
- If a match is found, update processing is performed
    - Each field in the file record is compared to the table record to determine whether any fields have changed. Only fields identified with a U in the templates are checked. If any updateable field has changed, the table record is updated appropriately.
    - If no updateable fields have changed, no further processing is done
- Delete processing
    - The basic load plans to do not include delete processing; only the insert and update processing described above
    - For example, there is no ability to exclude a record from an input file and trigger delete processing in this way.
    - Selected delete processing is available via separate plans

ERROR LOGS

- The eScholar load plans produce error information when records are rejected based on any of the checks described above.
- These log files are provided as feedback to the client in an appropriate manner.