How to Use MDSConnect

What is MDSConnect:

MDSConnect is an open-access version of the Library's MARC records. It includes nearly 25 million MARC records, as distributed in the unabridged 2014 Retrospective file sets. These MDS record sets are made available at no cost to the end user, primarily for research, education and development use.

What is a MARC record:

MARC records can be best described as individual bibliographic cards in a traditional card catalog. Like a card in a card catalog, they are portable and follow international standards and Library of Congress cataloging policies. All records use the MARC formats, which are international standards for the representation and communication of bibliographic and related information in machine-readable form. More information about the MARC formats is available at: www.loc.gov/marc

MDSConnect provides data in three file types that represent the same data, but is formatted differently.

- XML files normally are identified with having .xml as a part of their filename. They are normally larger in size as they contain formatting information within the files themselves.
- MARC8 and UTF-8 are identified with having .utf8 or .marc8 as a part of their file names. They are very similar to each other and are delineated with hexadecimal data element identifiers. They also require an outside schema or data map to be able to separate and identify the data elements within the files.

How is the data organized in MDSConnect?

Files are first organized by content type. More specifically they are categorized by MARC record type (or what the subject matter). Each of the nine available datasets is contained in a parent folder.

Each parent folder contains "G-zipped" files or parts. "Parts" will appear with the name of the dataset first and the format as the suffix.

For example: BooksAll.2014.part24.utf8.gz signifies the Books All dataset in UTF8 format. "Parts" are numbered across all formats (UTF8, MARC8 and XML) and are not necessarily in numeric order.

Users can sort data by clicking the arrows at the top of each field.

Downloading & Opening files

Downloading:

To download a file (part), simply click on the file name and you will be given the option of opening or saving the file.

Opening Files:

These files were originally packed/compressed using the G-Zip application. There are a number of free applications that can allow you to unpack these files so that they can be used by your target application. They include but are not limited to:

- PeaZip http://www.peazip.org
- 7-Zip www.7-zip.org
- Gzip www.gzip.org
- WinZip www.winzip.com

Open the XML data file (in Excel 2010 or later):

1. Open Excel, choose File → Open
2. Browse for the file d160102.records.xml and click open.
3. For the dialog box that appears(Open XML), choose the option titled "As an XML table"
4. A second dialog box may appear indicating that the specified XML source does not refer to a schema. Click OK to have Excel to create the schema for you.
5. You will now be able to view your MARC data in a user friendly manner where the data is categorized by attribute into the spreadsheet columns.

A more automated approach:

There are many enterprise applications which can automate the processing of MARC data into a much larger relational database. Some relational databases contain tools that allow developers to write customized applications to perform ETL (Extract-Transform-Load) functions. So that source MARC data can essentially be unpacked and distributed to multiple tables within a relational database. The source data can be fed into the system via a piecemeal approach and can process XML, MARC 8, or UTF-8 file formats. Source files can be automatically picked up from a source directory or internet location. Once they are consumed by the processing database, they are systematically unpacked to identify and separate the data elements from each MARC record. From this point an SQL script will move the data elements to the respective tables in the database.