4-16-2021

Metadata Remediation of Legacy Digital Collections: Efficient Large-Scale Metadata Clean-Up with a Sleek Workflow and a Handy Tool

Marina Georgieva
University of Nevada, Las Vegas, marina.georgieva@unlv.edu

Biz of Digital – Metadata Remediation of Legacy Digital Collections

Efficient Large-Scale Metadata Clean-Up with a Sleek Workflow and a Handy Tool

by Marina Georgieva (Visiting Digital Collections Librarian, UNLV Libraries Digital Collections, University of Nevada, Las Vegas; Phone: 702-895-2310) <marinik@abv.bg> *ORCID: https://orcid.org/0000-0002-2134-6719*

Column Editor: Michelle Flinchbaugh (Acquisitions and Digital Scholarship Services Librarian, Albin O. Kuhn Library & Gallery, University of Maryland Baltimore County; Phone: 410-455-6754; Fax: 410-455-1598) <flinchba@umbc.edu>

Abstract

Metadata remediation of digital collections is inevitable. At some point, each repository faces the need to clean-up digital collections legacy metadata so that it conforms to new standards. Typically, this need emerges either as a response to an updated metadata application profile, or as preparation for migration to a new digital asset management system (DAMS). Normalized metadata is critical for an improved search experience and easy discovery of digital objects.

This case study focuses on the University of Nevada, Las Vegas (UNLV) experience of cleaning up and preparing non-MARC metadata for migration to a new DAMS. The author shares her experience on cleaning up over 50,000 records in Excel for slightly over six months. Excel is a convenient, easily accessible tool with hundreds of free tutorials online. The remediation work utilizes various functions and formulas that are used to manipulate and optimize the metadata consistency.

Overview

UNLV Digital collections use a Dublin Core schema. The legacy collections employ a Dublin Core element set enriched with custom developed local fields unique for each collection. Fields and controlled vocabularies vary according to collections' peculiarities. To achieve consistency, upon a decision to migrate to a new DAMS, the metadata librarian developed a uniform metadata application profile for all digitized collections (photographic, manuscripts and oral histories). The new metadata profile omitted many custom legacy fields. It is more simplified, featuring the standard Dublin Core element set with fewer local fields intended to capture technical information or archival peculiarities. To support smooth migration to a new DAMS, all collections (legacy and new) must conform to the updated metadata profile. This decision required the clean-up of all legacy collections as part of the migration preparation.

The remediation process included review and rework of obsolete legacy fields and mapping their values to the new uniform metadata fields. The process featured extensive work with authority terms, in particular mapping terms from one vocabulary to another. Terms from Thesaurus of Graphic Materials (TGM) and Art & Architecture Thesaurus (AAT) were mapped to their Faceted Application of Subject Terminology (FAST) equivalent, and Thesaurus of Geographic Names (TGN) terms were replaced by GeoNames terms.

The metadata clean-up also involved active data manipulation in Excel using advanced functions that support large-scale remediation, such as filtering, trimming, concatenating, removing duplicates, indexing and matching of data sets, and normalizing dates.

Workflow

The whole remediation process is outlined in the workflow below. This article will focus only on certain segments.

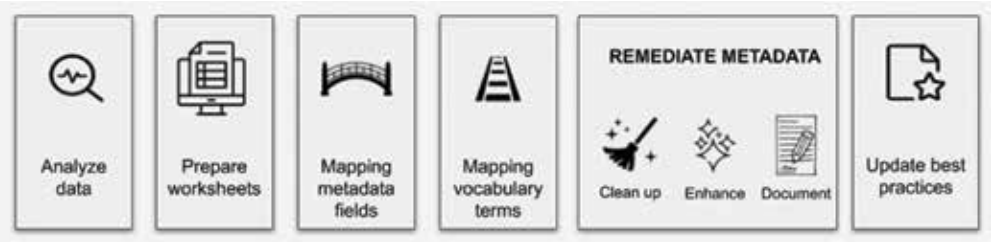


Figure 1. Segmented metadata remediation workflow

Mapping Legacy Fields to a Uniform Metadata Application Profile

The obsolete legacy collection fields were metadata rich, especially those created for grants. Examples of metadata rich collections are: Menus: The Art of Dining *http://digital.library.unlv.edu/collections/menus*, Neon Survey *http://d.library.unlv.edu/digital/collection/neo* and Nevada Test Site Oral History Project *http://digital.library.unlv.edu/ntsohp/*. Newer collections, such as Culinary Workers Union *http://d.library.unlv.edu/digital/collection/cwu,* embraced the large-scale approach with minimal metadata, but still contained collection-specific fields.

As we developed approaches to map obsolete fields to new fields, we strived to preserve the research effort. This often resulted in keeping valuable information and fitting it in new appropriate fields. The description field was a placeholder for non-normalized legacy values such as full sentences or notes. Names of people, organization and places were mapped to controlled vocabulary fields like contributor, collaborator, interviewer, geographic map (for locations) after proper normalization.

Preserving All Metadata

A good example of a collection that preserved all collection specific metadata is the **Neon Survey project**. Most legacy values were mapped to description. We appended the obsolete field label in front of the metadata string before transferring to description. Data that could be normalized (typically names) was placed in the creator or contributor fields.

	A	B	C	D
1	altLabel	uri	prefLabel	prefLabelStr
2		http://id.loc.gov/authorities/subjects/sh85002177	Agent Orange	Agent Orange -- http://id.loc.gov/authorities/subjects/sh85002177
3		http://id.loc.gov/authorities/subjects/sh85005173	Animal experimentation	Animal experimentation -- http://id.loc.gov/authorities/subjects/sh85005173
4	Nuclear Freeze Movement	http://id.loc.gov/authorities/subjects/sh85005721	Antinuclear movement	Antinuclear movement -- http://id.loc.gov/authorities/subjects/sh85005721
5		http://id.loc.gov/authorities/subjects/sh85007351	Arms control	Arms control -- http://id.loc.gov/authorities/subjects/sh85007351
6		http://id.loc.gov/authorities/subjects/sh85007355	Arms race	Arms race -- http://id.loc.gov/authorities/subjects/sh85007355
7		http://id.loc.gov/authorities/subjects/sh85015136	Boarding schools	Boarding schools -- http://id.loc.gov/authorities/subjects/sh85015136
8		http://id.loc.gov/authorities/subjects/sh85015475	Bombings	Bombings -- http://id.loc.gov/authorities/subjects/sh85015475
9		http://id.loc.gov/authorities/subjects/sh96011801	Book of Mormon	Book of Mormon -- http://id.loc.gov/authorities/subjects/sh96011801
10		http://id.loc.gov/authorities/subjects/sh85017004	Broadcasting	Broadcasting -- http://id.loc.gov/authorities/subjects/sh85017004
11		http://id.loc.gov/authorities/subjects/sh85019492	Cancer	Cancer -- http://id.loc.gov/authorities/subjects/sh85019492

Figure 2. View of a lookup table for LCSH terms mapped to FAST

The Blend Approach

The collection **Menus: The Art of Dining** used the blend approach. It preserved valuable collection specific data and omitted irrelevant fields. The table outlines all collection specific fields on the left and lists preserved and omitted fields on the right. Preserved metadata was mapped to *description*, *creator*, *contributor*, *subject* and *staff note* depending on whether the values could be normalized or were free text strings.

Original collection specific fields	Preserved fields	Omitted fields
• Individual contributors • Group contributors • Type of site • Restaurant/Site location • Type of menu • Type of cuisine • Language • Meals served • Price range • Illustrator • Chef • Sommelier • Manager • Printing Company • Street Address • City • State/Province/County • Country • Collection Subject (FAST) • Review • Groups (unknown restaurants)	• Type of cuisine • Language • Illustrator • Chef • Sommelier • Manager • Printing company • Street address • City • State/Province/County • Country • Collection Subject (FAST)	• Individual contributors • Group contributors • Type of site • Restaurant/Site location • Type of menu • Meals served • Price range • Review • Groups (unknown restaurants)

Omitting Legacy Metadata

A collection that omitted all legacy collection specific metadata is the **Nevada Test Site Oral History Project**. The disposed metadata was unstructured, redundant, or of lesser value.

Original collection specific fields	Omitted fields
• Date of birth • Place of birth • Digitization specifications • Related collections • Photo credit • Tests, experiments, operations • Agencies, organizations, laboratories • Treaties, laws and legislation • Identified individuals • Locations at the Nevada Test Site • Related locations • Local subject • Topical theme • Is part of	All collection specific fields

This collection was remediated with bare minimum metadata for several reasons: (1) long textual strings that could not be normalized; (2) specific technology and nuclear scientific terms not listed in any controlled vocabulary and (3) project specific fields that did not provide value outside the project context.

Mapping Subject Terms from Obsolete Controlled Vocabularies to id.gov

Mapping controlled terms from one vocabulary to another involved several steps: manual searching for terms, verifying scope notes, selecting equivalent terms, and recording them in a table. We followed the same process for all controlled vocabularies (TGN, LCSH, AAT, TGM) as we mapped them to FAST or GeoNames.

While verifying scope notes and mapping, we compiled a list of terms. This is what we refer to as look up tables. See Figure 2.

Terms in different collections often repeat, so we automated the process by using the look up tables to search for existing terms. The Excel function **index and match** was used to replace the old terms with the new counterparts. It automatically populated the new fields with the appropriate FAST term from the look up tables. We sorted all empty results to identify the missing terms which later we manually looked up on the authority website id.gov to add to our table. The final step of mapping was a second round of de-duping redundant terms generated after concatenating terms from multiple legacy fields.

Working with Name Authorities

Encountering names of people or businesses in legacy fields added an extra step to the workflow. It featured extracting and compiling all non-normalized names along with the people's dates and places of birth. The clean-up process included normalizing the names and recording them in our systems TemaTres and ContentDm. TemaTres is a linked data ready system that displays relations among all agents, such as family and employment relationships, occurrence in digital collections, cataloger's notes, etc. The system promotes consistency of metadata, especially for similar names as it gives biographical details about the agent and disambiguates among multiple name variants. All new TemaTres entries were mapped to the appropriate fields, such as *interviewer, narrator, creator, contributor,* or *subject*.

Nevada Test Site Oral History Project had hundreds of legacy non-normalized names. After compiling a list of them, we found the authority forms in Library of Congress Name Authority File (LCNAF) (*https://id.loc.gov/authorities/names.html*) or in Virtual International Authority File (VIAF) (*http://viaf.org)*.

For locally prominent people that did not have authority forms, their names were normalized to conform to the format *last name, first name, YYYY birth – YYYY death* before we recorded it in our system. After the clean-up, all newly normalized names were compiled in separate lookup tables for automated replacing of obsolete legacy terms.

Normalizing Metadata in Excel Using Functions and Formulas

Cleaning up metadata fields in Excel is an efficient process supported by numerous free tutorials. Remediation is a multi-step task to manipulate the data and it often involves applying various functions and formulas in a specific sequence.

Typically, our metadata remediation workflow followed this pattern:

1. **trimming** all extra spaces that may surround the values (leading and trailing spaces, occasionally double-spacing between words)

2. getting rid of the **end delimiter** (comma, semi-colon, period)
3. data **evaluation** to determine the clean-up approach and the combination of formulas and functions
4. **two types of clean-up approaches** depending on the metadata fields:
   a. controlled vocabulary fields (subject, creator, contributor, interviewer, location, date, material type, etc.)
   b. free-text fields (description, title, citation)

Details on cleaning *Controlled vocabulary fields* and *Free text fields* are available in the online Appendix at *http://bit.ly/Metadata-Remediation*, section *Normalizing metadata in Excel using functions and formulas.*

For more information on metadata fields and commonly used formulas for remediation refer to the online Appendix at *http://bit.ly/Metadata-Remediation*, section *Table with frequently remediated fields and most used formulas*.

Most Challenging and Time-Consuming Metadata Fields

Subject

Subject is highly utilized for searching, so it requires consistent metadata across all digital collections. Subject remediation took much time as it merged all obsolete topical metadata terms from various vocabularies (TGM, TGN, AAT, LCSH) in a new subject field. After merging, it required extensive manual work to map legacy terms to FAST equivalent. This additional workflow step included intellectual labor of verifying scope of legacy terms and matching to the appropriate FAST counterpart. To keep the work organized, we compiled tables with terms we already verified and mapped. Later these tables helped for automated mapping of repeating subject terms.

Description

Typically, description came with pre-filled information, but also it served as a storage place for valuable data from obsolete fields. The most challenging and time-consuming part of the process was to decide what legacy metadata to preserve, whether it brings value to researchers, and how to present it in a structured way in a free text field.

Tips and Tricks for Efficient and Smooth Remediation

Remediation requires attention on many levels. These tricks helped me stay efficient and deliver high quality output.

Sorting

Appropriate sorting is critical as each subset of data becomes easier to manipulate. The filtered data is more manageable, allows patterns to emerge, outlines discrepancies and facilitates data manipulation. Efficient work with large sets of data (some outnumbering 40,000 lines) is achieved by sorting on several criteria in multiple fields at the same time. See Figure 3.

Color-coding

Large data sets take weeks of work and it is easy to get lost and perform redundant actions. To avoid repetitiveness and ramp up efficiency, a color-coding system simplifies the progress tracking. Just a glimpse on the color-coded data displays what is completed, what is in progress and what is pending, as well as if anything needs revision. Defining a color legend keeps the color-coding consistent among all spreadsheets. Upon remediation and before sharing the clean data with the migration team, all color-coding is removed.

Version control

Version control keeps all cleaned fields safe and gives an option to revert one step in case something goes wrong. Although Excel has built-in version control, we use another approach: to save versions of our files upon remediating each field. For example, after finishing the subject field, we save a version of the file. Then, we make a copy of it and on the new copy, we continue working on the description field. In case the data gets mismatched or the formulas get messy, we can rework the description field from scratch. This keeps the previously finished fields safe as the older file versions are not affected.

Each version controlled file comes with a tab that contains a log. The log outlines all modifications and provides completion dates. See Figure 4.

Worksheets

Worksheets are helpful for remediating fields rich in controlled terms. They provide a clean workplace for massive subsets of data often featuring tens of thousands of terms. The obsolete data is extracted from the main file and copied in a separate worksheet where remediation takes place. Upon completion, the clean data is moved back to the main file where it replaces the legacy data. Working in separate worksheets allows more streamlined manipulation of data and easier progress tracking. It also guarantees if anything goes wrong, the legacy data will not be affected.

Compound objects

Compound objects are digital objects with two or more pages, which we refer to as "children." Their remediation can be challenging. In the spreadsheet each child (page) is represented as a new line. If children have item-level metadata, sorting and filtering of data must be handled carefully. During the process of sorting/filtering, if children are left behind or sent to the wrong parent, this may result in shifting metadata to wrong lines. In other words, children will get wrong metadata or will remain empty.

Best practice to avoid metadata shifting is to apply A-Z sorting. Typically, we sort by digital IDs as our compound objects have consistent file numbering convention: all children inherit the parent digital ID and get unique numerical extensions. When we sort by digital IDs, the children are always properly arranged.

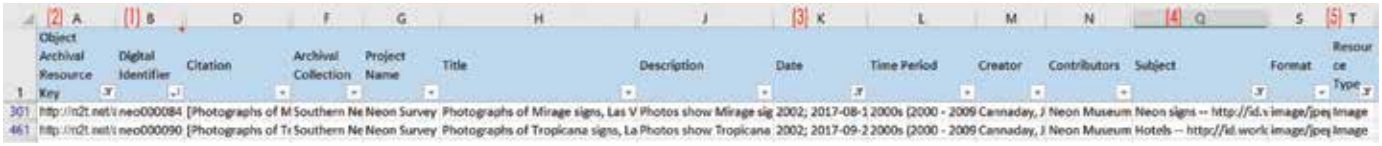
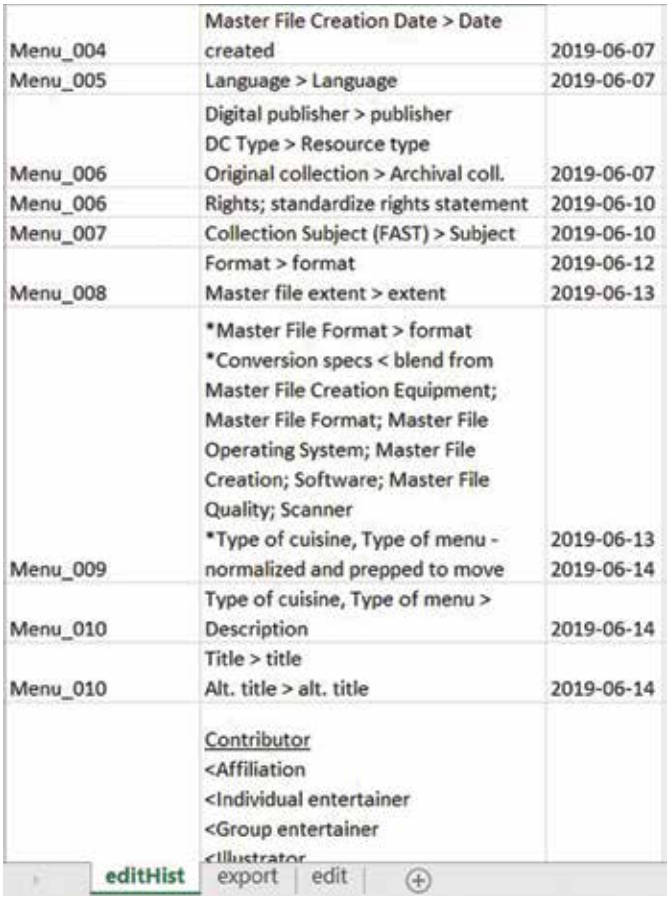


Figure 3. Excerpt from Neon Survey collection. These 2 lines are sorted out of 1,390 rows with data by applying several criteria in five fields. (1) Column B is sorted A-Z (alphabetical arrangement), (2) Column A is sorted to exclude all cells that are blank (no data), (3) Column K is sorted to show only lines that contain "2002," (4) Column Q is filtered to show only lines that contain the word "hotel," (5) Column T is filtered to exclude all lines that contain the word "text."



Menu_004	Master File Creation Date > Date created	2019-06-07
Menu_005	Language > Language	2019-06-07
Menu_006	Digital publisher > publisher DC Type > Resource type Original collection > Archival coll.	2019-06-07
Menu_006	Rights; standardize rights statement	2019-06-10
Menu_007	Collection Subject (FAST) > Subject	2019-06-10
Menu_008	Format > format Master file extent > extent	2019-06-12 2019-06-13
Menu_009	*Master File Format > format *Conversion specs < blend from Master File Creation Equipment; Master File Format; Master File Operating System; Master File Creation; Software; Master File Quality; Scanner *Type of cuisine, Type of menu - normalized and prepped to move	2019-06-13 2019-06-14
Menu_010	Type of cuisine, Type of menu > Description	2019-06-14
Menu_010	Title > title Alt. title > alt. title	2019-06-14
	Contributor <Affiliation <Individual entertainer <Group entertainer <Illustrator	

editHist | export | edit

Figure 4. Version control log that keeps track of new changes in each file with date of completion, file name and brief description of the modifications. Menus: The Art of Dining collection

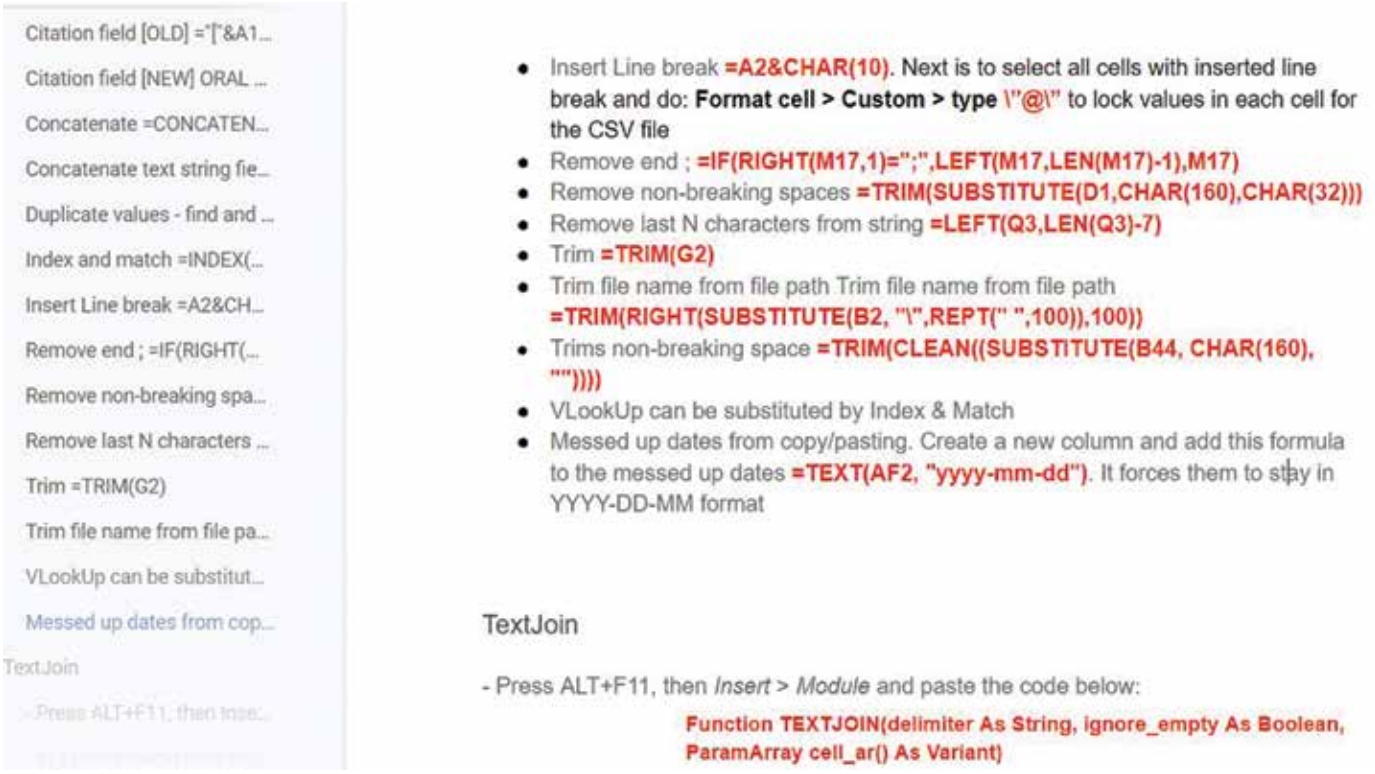
Library of frequently used formulas

Keeping a document with frequently used formulas saves time and boosts remediation efficiency. All formulas are supplemented by brief descriptions when to use them and how they work. This best practice promotes consistent metadata clean-up across all collections. Additionally, it helps with analysis and decision-making on choosing the sequence of actions for each data set. See Figure 5.

Conclusion

Although it may seem overwhelming to work with large sets of data, it is vital to remember that data can be further divided into multiple data subsets for easier and more manageable manipulation.

Developing a segmented workflow is critical for smooth, efficient, and successful operations. Segmentation ensures predictable data manipulation, structured remediation, and effortless progress tracking that results in successful project completion. Segmentation is complemented by version control for a more robust workflow and allows unforeseen modifications even after the project is completed. Remediation projects completed in Excel yield quick turnover and high-quality output.



- Insert Line break **=A2&CHAR(10)**. Next is to select all cells with inserted line break and do: **Format cell > Custom > type** \"@\" to lock values in each cell for the CSV file
- Remove end ; **=IF(RIGHT(M17,1)=";",LEFT(M17,LEN(M17)-1),M17)**
- Remove non-breaking spaces **=TRIM(SUBSTITUTE(D1,CHAR(160),CHAR(32)))**
- Remove last N characters from string **=LEFT(Q3,LEN(Q3)-7)**
- Trim **=TRIM(G2)**
- Trim file name from file path Trim file name from file path **=TRIM(RIGHT(SUBSTITUTE(B2, "\",REPT(" ",100)),100))**
- Trims non-breaking space **=TRIM(CLEAN((SUBSTITUTE(B44, CHAR(160), ""))))**
- VLookUp can be substituted by Index & Match
- Messed up dates from copy/pasting. Create a new column and add this formula to the messed up dates **=TEXT(AF2, "yyyy-mm-dd")**. It forces them to stay in YYYY-DD-MM format

TextJoin

- Press ALT+F11, then *Insert* > *Module* and paste the code below:

Function TEXTJOIN(delimiter As String, ignore_empty As Boolean, ParamArray cell_ar() As Variant)

Figure 5. Excerpt from the library of formulas with brief descriptions how they work