Master List Link Assessment (MLLA) Fall 2019 Report

By Daniel Choi, L. Cully

Updated 13 January 2020

Table of Contents

Section 1 - Introduction/General Information

The following report serves as a continuation to The 2018 Master List Link Assessment report [1] from 2018 and a continuation of the Master List Link Assessment (MLLA) project. The project is a subtask to the EMDAC Metadata Cleanup Project. From the 2018 report: *"The purpose of the Master List Link Assessment project (MLLA) was to be able to automate analysis on the EMDAC Master Lists' links. The MLLA project was used to be able to check all URLs in an EOL project's Master List for any incorrect or unknown links. Furthermore, the MLLA effort was designed to provide recommendations and assistance to users who will receive these results to point them in the right direction to correct the datasets on the Master List with broken URLs. This analysis provides DMS Members with sufficient information to clean up these incorrect or unknown links in order to point to the correct data for the Earth Science Community to be able to use efficiently for any research."*

The 2019 continuation of this project was to check the 2018 software for any coding errors and improve upon the 2018 software. The software used in 2019 is the same software used in 2018 with some improvements. The software takes the links from Master Lists and identifies the links that work or are broken. The software then generates statistics for the given field projects. The

overview of the software used and other specific information can be found in the 2018 report titled, *The Master List Link Assessment*. For this Fall 2019 report, the software modules used were `linkChecker.pl`, `WrapperLinkChecker.pl`, and `results_summary.pl`.

The 2019 software can be found in subversion at:
(root)/tools/link_verification/software/linkChecker2019.pl

Note that the software identifies and divides URLs into three different types: working, broken, and unknown URLs. The software first tries to connect to a URL given from the input list. If the software successfully connects to the URL, the URL is counted as "working". If the software cannot connect to the URL, it then determines whether the link is "broken" or "unknown". A link is determined to be "unknown", if it is not a proper URL (i.e., the URL does not contain either of the phrases "http" or "ftp:"). If such phrases are not in the URL, the URL is considered to be an "unknown" link. If the URL does contain either of these phrases but connecting failed, the URL is defined and counted as a "broken" link.

Section 2 - Improvements/Corrections made to original MLLA 2018 Software

The 2018 MLLA software had a coding error where the counting of the total URLs was computed. The old software identified and then counted some URLs as documents and some as FTP links. A separate count was computed for each type. The 2018 software would first add each input URL to the total URL count. The issue with the software was that when the software checked the link type, it would add in an additional count into the total URL count. In other words, the total URL count would go up by 2 for each document or FTP link URL when it should have been incremented by only 1.

The main consequence of this is that the output would have a higher total URL count which would affect the percentages of broken and unknown URLs. Some of the figures in the 2018 report show incorrect information regarding the total URL count and incorrect percentages. The 2019 software fixes this issue so the count for the total URLs is correct. Another issue with the 2018 software was that the output file containing the statistics of the URLs could be difficult to interpret due to the word choice used. The 2019 output file was updated so there is more clarity regarding how to interpret the information.

Section 3 - Description and Results of Test Cases

The following three test cases were created to show the differences between the old 2018 MLLA processing with the new 2019 MLLA processing:

> **Test Case 1** - Run the MLLA 2018 software to confirm total broken links counts. Execute the software on the same set of field projects as done for original run in 2018. Compare the results to those found in the 2018 report. Note that the results will not be the same as when executed in 2018 since links have been corrected since that time. Next run the

2019 software with the same field projects. Compare the 2019 results with the 2018 software and results to see how the improvements affect the outcome.

Test Case 2 - Run the MLLA 2019 software on subsets of projects as shown in *Figure 4.2.3* in the MLLA 2018 Report. Create a figure similar to *Figure 4.2.3* based on the 2019 results to see the distribution of the total links and broken links.

Test Case 3 - Run the MLLA 2019 software on all field projects through 2019 including all new projects that may or may not have been included in the MLLA 2018 processing. The following results will be generated after running the software: number of total field projects, how many additional projects were added compared to the 2018 run, number of links tested, and the number of broken links.

Test Case 4 - While running the previous test cases, it seemed that the time of day or number of processes running on the system might have an impact on the execution of the wget command part of the processing. The software executes wget commands to determine if a link is broken or not. The wget command wait time is set to 10s and is hard coded. The software will do the wget test on a link a maximum of 2 tries (maximum of 20s). If a positive response is not returned during that time limit, the link is categorized as broken. This test case was added to determine if the time of day might impact the response time of the wget command. The corrected 2019 software was executed both in the morning and in the afternoon in the same day.

The results of the test cases are shown below.

Test Case 1 Results:

Note that each URL is categorized in two different ways. First, each URL is checked to see if the URL is an FTP link, document, or neither. Second, each URL is checked to see if it is a working, broken, or an unknown link. The sum of the FTP links, documents, and URLs that are not FTP links and not documents should add up to the total URL count. The sum of working, broken, and unknown links should also add up to the total URL count if the software works as intended.

A. **Old 2018 Results: (Executed original 2018 Software with no corrections. *Note that due to an error found in the original software, the results in this test case are erroneous. The software error caused the count of the total number of URLs to be almost double the actual amount.*)**

   Total number of URLs: **48877** ***(incorrect count)***

   Total number of FTP links: 252 (0.52%)
   Total number of documents: 8756 (17.91%)

Total number of working links: 23816 (48.73%)
Total number of broken links: 2258 (4.62%)
Total number of unknown links: 14 (0.03%)

B. New 2018 Results: (Original 2018 Software with no corrections but with additional/new projects. *Existing software error results in total number of URLs to be almost doubled.*)

Total number of URLs: **49788** ***(incorrect count)***

Total number of FTP links: 256 (0.51%)
Total number of documents: 8932 (17.94%)

Total number of working links: 24914 (50.04%)
Total number of broken links: 1734 (3.48%)
Total number of unknown links: 15 (0.03%)

C. 2019 Results: (Corrected Software with additional/new projects. Compare with "B." above.)

Total number of URLs: **26586**

Total number of FTP links: 256 (0.96%)
Total number of documents: 8921 (33.56%)

Total number of working links: 24893 (93.63%)
Total number of broken links: 1678 (6.31%)
Total number of unknown links: 15 (0.06%)

Comparing the old 2018 results in "A" above with the recent 2018 processing run ("B" above) shows that more links were added and that the number of broken links decreased. Over time, a number of broken links were repaired. Note that the sum of working links, broken links, and unknown links using the 2018 processing software do not add up to the total URL count. The 2019 software fixes this error. **The 2019 results show an accurate summary of the results.** Here the sum of the links adds up to the total URL count. The main difference between the 2018 and 2019 software is the total URL count. The 2018 software double counted *almost* all incoming link types. **Note that the original 2018 processing was executed in late September and early October of 2018. Test Cases 1B and 1C were completed in late October of 2019 on the same day.**

	Total URLs	Number of FTP Links	Number of Documents	Number of Working Links	Number of Broken Links	Number of Unknown Links
A. **Old 2018 Results** *(Known software errors)*	48877	252 (0.52%)	8756 (17.91%)	23816 (48.73%)	2258 (4.62%)	14 (0.03%)
B. **New 2018 Results** *(Known software errors. Recent projects included)*	49788	256 (0.51%)	8932 (17.94%)	24914 (50.04%)	1734 (3.48%)	15 (0.03%)
C. **2019 Results** *(Corrected software. Recent projects included)*	**26586**	**256 (0.96%)**	**8921 (33.56%)**	**24893 (93.63%)**	**1678 (6.31%)**	**15 (0.06%)**

Test Case 2 Results:

Figures 1 and 2 below show the total URL count and the Broken URL count for 5 year intervals for the 2018 and 2019 software respectively. The first interval for the figures was from 1970-1999 as there was not as many links in that time period. **The patterns in the figures remain the same despite the changes made in the 2019 software. The main difference between Figure 1 and Figure 2 is that Figure 1 has higher total URL counts for each time interval.**

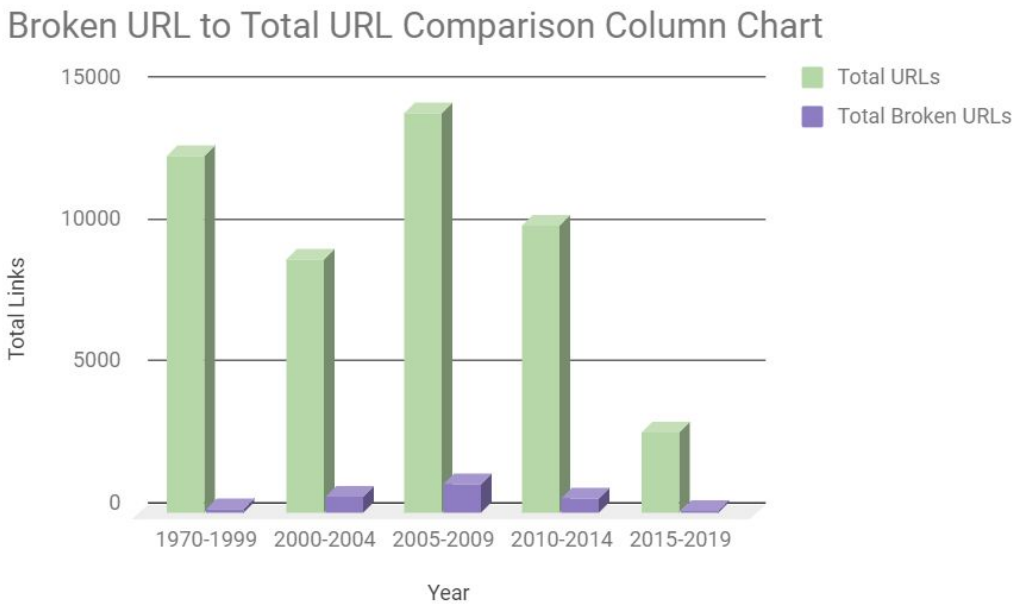

Figure 1: The above image compares the total count of URLs to the total count of non-unique broken URLs within that 5-year period for the 2018 software. (See section 4.1 below for the list of projects included.)

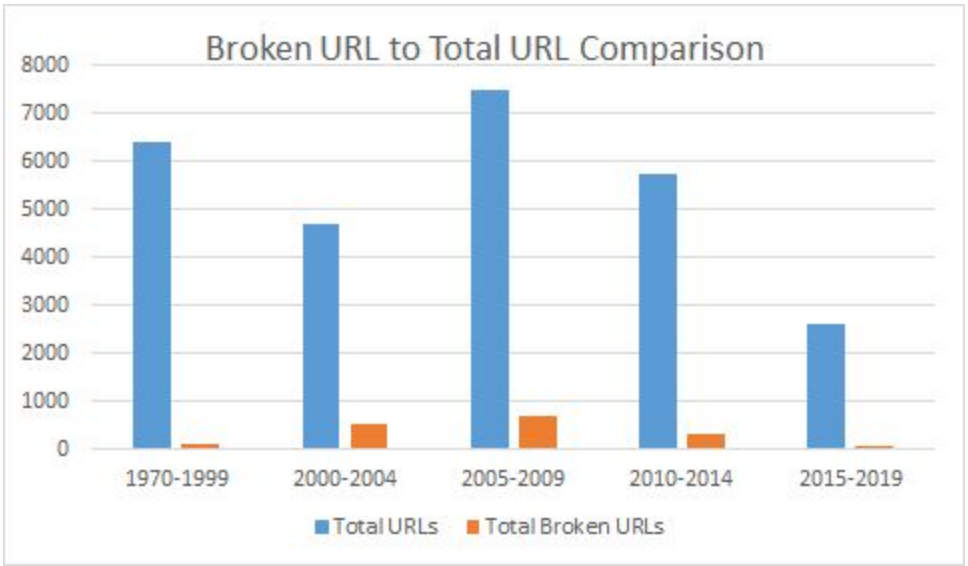

Figure 2: The above image compares the same information as Figure 1 but with the results for the 2019 software. (See section 4.1 below for the list of projects included.)

Test Case 3 - 2019 Run with Corrected Software and Recent Projects included

Number of projects processed: 113 *(See section 4.1 below for the list of projects included.)*
Number of projects added since 2018 run: 12
Total number of urls: 26902
Total number of working links: 25217 **(93.74%)**
Total number of broken links: 1669 (6.20%)
Total number of unknown links: 16 (0.06%)

The information above shows the results of the 2019 processing. The 2019 run used all of the projects used in the 2018 run as well as 12 recent projects.

Test Case 4 Results:

The results in Test Case 1 had a few differences in 1B and 1C. Total URL count aside, the URL counts for the other categories were different. Since the method for counting these different URL types (besides total URL) are the same for the 2018 and 2019 processes, it is strange that these numbers were different. Test cases 1B and 1C were run on the same day but at different times. Test Case 4 was created to see how time of day might affect the results of these processes. For this Test Case 4, the 2019 process was run in the morning and in the afternoon on December 3, 2019 to see if the time affects the results.

The table below shows the results of this Test Case. The number of URLs remained the same throughout the day. The number of working links increased as time passed. Unlike 1B and 1C, the URL counts here did not change as much overall. Based on these results, the time of day and a variety of other possible items (e.g., anything that might impact the wget command response time, employees may correct broken links, etc.) may have an impact on results.

	Total URLs	**Number of FTP Links**	**Number of Documents**	**Number of Working Links**	**Number of Broken Links**	**Number of Unknown Links**
A. **Morning Results**	27041	256 (0.95%)	9045 (33.45%)	**25178 (93.11%)**	**1847 (6.83%)**	16 (0.06%)
B. **Afternoon Results**	27041	256 (0.95%)	9045 (33.45%)	**25195 (93.17%)**	**1830 (6.77%)**	16 (0.06%)

Section 4 - Summary of Current State of Master List Links for All Projects

The following projects were used for the 2018 processing:

- ACE-ENA
- ADELE SPRITE
- AMMA
- ARCSS
- ARISTO2015
- ARISTO2016
- ARISTO2017
- ATLAS
- BAMEX
- BEST
- BSIERP
- CASES-97
- CASES-99
- CONTRAST
- CSET
- CUPIDO
- DBO
- DC3
- DEEPWAVE
- DYNAMO
- ECLIPSE
- EPIC
- FRAPPE
- FRONT
- FRONT-DE2
- FRONT-ROSE
- FRONT-STEP
- GRAINEX
- HAIC-HIWC
- HAIC-HIWC 2015
- HCRTEST
- HIPPO-1
- HIPPO-2
- HIPPO-3
- HIPPO-4
- HIPPO-5
- HIWC-FL
- ICEBRIDGE
- ICEBRIDGE-2015
- ICE-L
- ICE-T
- IDEAL
- IDEAS-1
- IDEAS-2
- IDEAS-3
- IDEAS-4 C130
- IDEAS-4 GV
- IFRACS
- IHOP
- INDOEX
- ITEX
- ITOP
- LATTE
- LPB
- MATERHORN-X
- MILAGRO
- MITTS
- MPEX
- NAME
- NOREASTER
- OHHI
- ORCAS
- OWLES
- PACDEX
- PACMARS
- PACS
- PASE
- PECAN
- PERDIGAO
- PLOWS
- POST
- PREDICT
- RAINEX
- RELAMPAGO

- RICO
- SAANGRIA-TEST
- SABIRPOD
- SALLJEX
- SAS
- SBI
- SHEBA
- SNOWIE
- SOCRATES
- SOCRATES-TEST
- SPRITES-II
- START08
- T28
- TCI
- TIMREX
- TORERO
- T-PARC
- TREX
- VERTEX
- VOCALS
- VORTEX2
- VORTEX-SE 2016
- VORTEX-SE 2017
- VORTEX-SE 2018
- WAMME
- WE-CAN
- WINTER

Section 4.1 - Summary of Master Lists Included in Processing

All of the projects above were used in the Fall 2019 MLLA processing run. In addition, the following projects were also included in the 2019 processing. The following projects **were not included** in the 2018 processing:

- CCOPE-2015
- CHEESEHEAD
- ECLIPSE2019
- HIGHWAY
- HIPPO
- ICICLE
- MASCRAD
- MESO18-19
- OTREC
- SAVANT
- TORUS 2019
- WE-CAN-TEST

Section 5 - Conclusions and Recommendations

Fixing the errors in the 2018 software changed the output of the software significantly as compared to the 2019 processing. The old software gave significantly larger (almost double) total URL counts which resulted in lower percentages of counts in all link categories. This can be seen in Test Case 1A and 1B above. Test Case 2 shows that the distribution of Total URL and Broken URL counts are very similar except the magnitude of Total URLs are different. The 2019 software properly generates URL counts that add up to the correct results. The output in Test Case 3 shows the current state of the Master List as of late October 2019. The projects used for each run are shown in section 4.1 above.

The overall recommendation is to execute this process again in March 2020 (3 months) to compare results from the corrected software and thereafter on a 6 month schedule. It is

also recommended that additional test cases should be run on the processing software to ensure all errors have been found.

References

[1] A. Robinson, C. Connell, and S. Patnam, “The Master List Link Assessment.” [Online]. Available: https://docs.google.com/document/d/1mPZZcI6dd3d-ahFErvwn9vlUq6C30V8mQ6_c4j9P8L0/edit.