**MIR Migration Steps**

Overview:

MIR data resides on the Lincoln file server at \\AIONELI93AP1\MIR\_Truncated\_Library. This document outlines the steps performed to import the csv format files into the Lincoln server (AIONELI93AP1) SQL Server RDBMS, do the required data manipulation, and coordinate replication to the DISC Science PROD SQL Server RDBMS. The import transferred 342,799 MIR csv files with 1765 absorbance values each from the Lincoln server to DISC.

The MIR data was built up over 2 years. The initial import was completed in late September 2021.

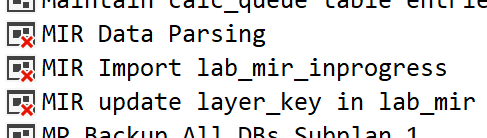
First Step

Overview: Bring the csv data into the mir\_test\_copy1 temporary database. This process captured the fileserver path and filename for the csv file along with the absorbance and wavelength fields for 342,799 files with 1765 rows resulting in a 605,040,235 row table.

* The SSIS package is located at \SSISDB\MIR\_test1\MIR Import\MIR for

SSL\_Repo lab\_mir\_inprogress.dtsx in the Integration Services Catalogs on AIONELI93AP1.

* The job MIR Import lab\_mir\_inprogress was used to run the SSIS package. Prior to the SQL Server maximum server memory configuration being changed from 2 TB to 13 GB this job took 30 hours to run.
* CEC, Tammy, and Scarlett kept in loop b/c of sever performance concerns
* Screenshot of this job and subsequent step jobs on AIONELI93AP1:



Second Step

Overview: The filename, foldername, lab\_proj\_name, and rep\_num (replication number) fields were extracted out of the fileserver path and filename field. The temporary database mir\_test\_copy1 was used and the target table is mir\_import.

* The MIR Data Parsing job has steps for
  + Parse the foldername
  + Parse the lab\_proj\_name
  + Parse the rep\_num
* CEC, Tammy, and Scarlett kept in loop b/c of sever performance concerns

Third Step

Overview: Exported the mir\_import table from temporary database mir\_test\_copy1 to the lab\_mir table in SSL\_Repo. The layer\_key field from dbo.layer in SSL\_Repo was added in SSL\_Repo

* The MIR update layer\_key in lab\_mir job is used for this process
* CEC, Tammy, and Scarlett kept in loop b/c of sever performance concerns

Fourth Step

Overview: At this point in the initial replication of the lab\_mir table in SSL\_Repo on the Lincoln server was replicated by DSO to DISC Science production. Because of access privileges DSO also needed to copy the lab\_mir table to DISC Science Dev2A.

* CEC, Tammy, and Scarlett kept in loop b/c of sever performance concerns

Fifth Step

Overview: After replication was completed to Dev2A the lab\_mir\_array table was populated from lab\_mir with the STRING\_AGG function. The STRING\_AGG function requires SQL Sever 2019 and a nvarchar field. In this case absorbance is a nvarchar(MAX) b/c of the 14,000 character string.

* This MIR Populate lab\_mir\_array table job in SciDev2A has this script
* DSO ran this job b/c of access privileges

Sixth Step

Overview: The field labsampnum was added to lab\_mir\_array in ncsslabdata in SciDev2A

UPDATE dbo.lab\_mir\_array

SET labsampnum = natural\_key

FROM dbo.lab\_mir\_array AS lma

inner join dbo.layer l

ON lma.layer\_key = l.layer\_key;

Seventh Step

Overview: Add the d\_wavelength\_array\_id. Originally is was believed that a single wavelength array was used. As it turns out there are 2. A domain table d\_mir\_wavelength\_string was specified in ncsslabdata.

1st Array

Use ncsslabdata

UPDATE lab\_mir\_array

SET d\_wavelength\_array\_id = '1'

WHERE left(wavelength, 11) = '4001.65608,';

2nd Array

Use ncsslabdata

UPDATE lab\_mir\_array

SET d\_wavelength\_array\_id = '2'

where left(wavelength, 11) = '4001.63495,'

The row counts from this step are:



Eighth Step

The domain table lab\_mir\_wavelength\_string\_d specified in ncsslabdata was requested to be replicated to sdmOnline.

Summary and Next Steps

1. Don’t do processing on Lincoln server – move that step to DISC.
2. The Excel macro to process OPUS files for publication to the Truncated Library share requires OPUS software and was developed by a Rick Nesser and a Bruker programmer. A system to parse out the new csv files prior to those being added to the Truncated Library share is needed.
3. Currently checking with DSO on Lincoln file share from SciDev2A
4. Need to know required frequency of imports – quarterly, monthly?