This is a living document created by Chris Coletta on Wed Feb 22, 2023

# Introduction

The purpose of most of those scripts is to take USDA-supplied discovery artifacts (mostly large IDMS data dictionary text files), scrape them for their content, and reformat that content for ingestion into Erwin. For your convenience, I have provided them in their original Jupyter Notebook format, as well as converted them into simple Python scripts for convenient inspection. Unless they are run inside a Jupyter notebook environment, some of the ancillary exploratory outputs -- like histograms of line lengths, pivot tables of number of lines in a file, etc -- will not be saved to the converted Python script .py files since they are meant to be viewed interactively in Jupyter rather than saved to an intermediate file.

# (EDS-Internally) Deliverable Documents

Coletta current state assessment documents and data model intermediate representations work product [folder](https://drive.google.com/drive/folders/12GpczBPSTqKTPM95tKgZGCOrInWg4M-R?usp=sharing):

1. 2023-02-21\_IDMS\_Copybooks\_FARMS\_FLAT\_STRUCTURE.txt
   1. This is a COBOL copybook that contains the current state assessment of the IDMS database tables that are relevant to FSA production activities.
      1. The mechanism for excluding tables that are outside of FSA scope or are not production was accomplished by identifying which IDMS “schema”/”subschema” (in the IDMS context, schema/subschema is organizational mechanism for grouping tables into functional domains.)
      2. FSA DBAs Gary White and Lester Maslow were the ones who identified the relevant schemas/subschemas as: FARMS, FARMS0, FARMS1, FARMS2, RMS, SCMAC01, SCMACCT, SCMFA01, SCMHN01, SCMUA01
   2. As of this writing no entity relationships are represented in this file, nor could it be because COBOL copybooks are a format that does not provide a mechanism for representing those relationships
      1. Samee and Chris have discussed changing format of this deliverable to one that could represent relationships, like SQL DDL.
2. more…

# (EDS-Internally) Deliverable Scripts

Coletta [Python Code folder](https://drive.google.com/drive/folders/1Fa4Bu0_UvKQIZRzqpheY358N6FI2ZhsQ?usp=sharing).

Here is a list of the most important ones and a brief description of what they're meant to do:

1. 2023-02-03\_schema\_source\_exploratory.py
   1. This is a script that is under ongoing revision. It is meant to scrape the IDMS schema source given to us by the IDMS DBAs ([the ones in here](https://drive.google.com/drive/folders/14rEm-4HxKyTgajKqNQJBgJsviZtwmzzf?usp=sharing)) and format it into a spreadsheet. We were told that schema source is taken as input to create various data dictionary reports, and is a low-level representation of the physical IDMS data model.
   2. Output is a spreadsheet with table names, field names, data types, etc
2. 2023-02-22\_CreateCopyBooks-FLAT-v3.py
   1. Takes the output from script 1 above and formats it into a COBOL copybook/data layout file, which is directly imported into Erwin.
   2. The idea is that the data modeling process can continue in Erwin with hand edits and improvements. The model can then finally be exported for consumption by AWS RedShift.
3. 2023-01-25\_IDMS\_scrape\_IDMS\_Data\_Dictionary.ipynb
   1. This is a big one!
   2. Inputs:
      1. Input 1: a lightly edited version of the 200k line Data Dictionary file we got last year ([this one](https://drive.google.com/file/d/1VWL-8ITiRcp7lowMShaE6qbVcDQYy5eR/view?usp=sharing)), that contains a report formatted "DATA DICTIONARY REPORTER REL 17.0". This file contains critical information such as a list of records (tables), their entities (fields), schema (a unit of organization within the data model), and mainframe program that uses each record.
         1. Confusingly, we were also provided supplementary Data Dictionary reports ([these ones](https://drive.google.com/drive/folders/1c83wQSkyFfI84A0g32X9vbiygrpDklkl?usp=sharing)), that are formatted completely differently, with formatting labeled "SCHEMA RECORD DESCRIPTION LISTING", and thus this script won't work on them :-(
      2. Input 2: Bob Lovelace's spreadsheet ([this one](https://docs.google.com/spreadsheets/d/1hF9GqCwdIshIuYc68AG_BRvOCZZZWP8d/edit?usp=sharing&ouid=113878876011349130267&rtpof=true&sd=true)) where PLAS subsystem 2-letter codes are explained
      3. Input 3: [file\_lengths.txt](https://drive.google.com/file/d/18Bdpgf08HtJ7TjSDk6jpfl4xvUXSsXaT/view?usp=sharing) contains line counts for all 16 thousand mainframe programs.
   3. Outputs:
      1. Does an analysis of the 16 thousand mainframe code files and uses Bob Lovelace's spreadsheet to try to characterize them.
      2. More outputs…
4. More scripts…