PharmaSUG 2019 - Paper SS-120

An Automated, Metadata Approach to Electronic Dataset Submissions

Janette Garner, Kite Pharma, a Gilead Company

ABSTRACT

In 2014, the Food and Drug Administration (FDA) provided guidance regarding Section 745A(a), an amendment to the Federal Food, Drug, and Cosmetic Act (FD&C Act) that requires regulatory submissions (e.g., new drug applications [NDAs] or biologics license applications [BLAs]) to be submitted in electronic format. The guidance took effect at the end of 2016. This paper presents a metadata-driven solution that facilitates the generation of the dataset package for electronic dataset submission that is compliant with the FDA expectations based on published FDA guidance documents.

INTRODUCTION

When preparing an electronic submission to the FDA, the document “Providing Regulatory Submissions in Electronic Format — Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications Guidance for Industry” should be followed. This guidance implements the electronic submission requirements of Section 745A(a) of the FD&C Act for the electronic format of the content submitted to the FDA. Among the many points raised, the following are particularly relevant to those preparing study dataset package for submission:

- Files to be submitted must follow the submission data exchange standards as defined in the FDA Data Standards Catalog. This catalog describes the standards, formats, and terminologies that the FDA accepts for electronic submissions.
- Study datasets and their supportive files should be organized into a specific file directory structure when submitted in the Electronic Common Technical Document (eCTD) format. Figure 1 illustrates this directory structure that is described in the FDA technical conformance guide (pg 32). Submission of files within the appropriate folders allows the agency’s automated systems to detect and prepare datasets for review, minimizing the need for manual processing and thus expediting the review process.

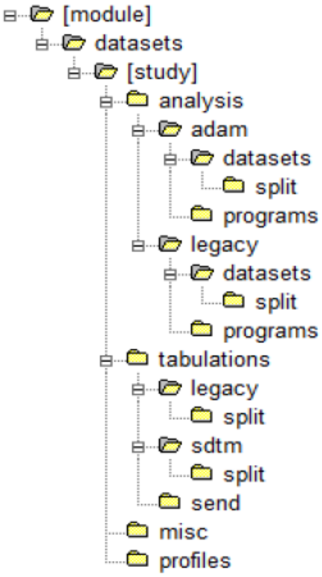


Figure 1. Directory Structure for Study Datasets (Reprint of Figure 1 from Study Data Technical Conformance Guide v 4.2.1 January 2019, pg 32)

PREPARING INFORMATION FOR AUTOMATION

Since the folder structure for electronic submission is already defined by the FDA (see Figure 1 above), we can use it (and the contents described within the FDA Data Standards Catalog) to our advantage to construct a library of utilities (Linux scripts that run Python code or SAS® macros). These utilities leverage a few files that store key metadata information including:

- file type - expectations are described within the FDA Data Standards Catalog
- source location - defined by the sponsor's standard folder structure
- destination location - defined by the eCTD structure shown in Figure 1

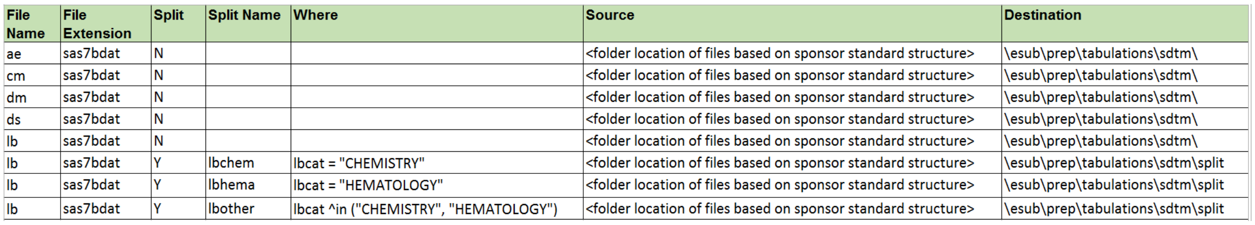
In effect, the metadata file serves as a bridge between the sponsor's folder structure and the eCTD structure used by the FDA. (Note: Figure 4 in Appendix A provides a description of each of the folders identified in Figure 1.)

Metadata files can be created for each directory branch of the eCTD structure for electronic submissions. That is, one can create the following files:

- Metadata file corresponding to the SDTM dataset related documents within the Tabulations branch.
- Metadata file corresponding to the ADaM dataset related documents within the Analysis branch.
- Metadata file corresponding to the Analysis Programs related documents within the Analysis branch.
- Metadata file corresponding to the Misc related documents within the Misc branch.
- Metadata file corresponding to the Patient Profiles related documents within the Profiles branch.

Note: The Misc branch above refers to datasets that do not qualify as analysis, tabulations, or profiles. This is commonly used for lookup tables. Importantly, non-CDISC datasets that support the analysis or tabulations would still go under these branches, but would be placed within the appropriate "legacy" folders, as shown in Figure 4 in Appendix A.

Figure 2 shows an example of a metadata file containing information on the SDTM datasets.



File Name	File Extension	Split	Split Name	Where	Source	Destination
ae	sas7bdat	N			<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\
cm	sas7bdat	N			<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\
dm	sas7bdat	N			<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\
ds	sas7bdat	N			<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\
lb	sas7bdat	N			<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\
lb	sas7bdat	Y	lbchem	lbcat = "CHEMISTRY"	<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\split
lb	sas7bdat	Y	lbhema	lbcat = "HEMATOLOGY"	<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\split
lb	sas7bdat	Y	lbother	lbcat ^in ("CHEMISTRY", "HEMATOLOGY")	<folder location of files based on sponsor standard structure>	\esub\prep\tabulations\sdtm\split

Figure 2. Metadata File Containing Information on the SDTM Datasets

The "File Name", "File Extension", "Source", and "Destination" columns are fairly self-explanatory, but the other 3 columns warrant a bit more discussion. "Split" is set to "N" by default, and "Split Name" and "Where" are left blank. If the source file is greater than 5 gigabytes (GB) in size, Section 3.3.2 of the Study Data Technical Conformance Guide states that the file should be split into smaller datasets no larger than 5 GB. In such instances, "Split" is set to "Y" and the names of each of the smaller datasets are identified by separate rows under "Split Name". "Where" identifies the logic used to subset the original dataset to obtain the split dataset.

It is possible to reduce the number of metadata files by consolidating some of those listed above. However, this may lead to additional work if one of the components requires updating. For example, suppose that the analysis programs and ADaM datasets were in the same file. If we made any update to the analysis programs, this would require that both the analysis programs be re-converted from .SAS to .TXT and ADaM datasets re-converted from .SAS7BDAT to .XPT in order to preserve traceability (e.g., refreshing the .TXT analysis programs and .XPT ADaM datasets). This is necessary to ensure that the timestamp of the metadata is later than the timestamp of the ADaM datasets (which uses that metadata file). Failing to do this creates an inconsistency that could result in an audit finding. Furthermore, separate metadata files allow more than one programmer to work on a package concurrently. This may reduce the timeline for submission.

AUTOMATING ELECTRONIC DATASET SUBMISSION USING METADATA INFORMATION

It is worth noting that the process below was conceived assuming a Linux environment, but can be adapted to other environments provided that the utilities have the ability to modify folder contents. Moreover, the process described below is based upon Module 5 which only covers the submission of clinical data. Other modules (e.g., Module 4, which consists of only non-clinical data) can also consist of SAS® datasets and programs. This process can be adapted to satisfy the needs of those modules or even to use programs written under and datasets prepared by other statistical software (e.g., R).

The dataset submission package is generated in 4 steps.

1. The utility ***mkmd*** generates the relevant metadata files assuming a standard folder structure used by the sponsor, the eCTD folder structure and the data exchange standards defined by the agency.
2. The programmer converts datasets using the utility ***prepxpt*** and/or programs such as ***sas2txt*** to create the submission files, such as the define.xml, SDTM annotated CRF, etc.
3. Another utility, ***mkmod***, pulls completed files from the source locations and copies these files into the destination location under the eCTD folder structure.
4. Final quality assurance checks can be performed on the assembled package. The utility ***chkesub*** can be created to programmatically perform as many checks as possible while minimizing the checks that must be done manually. This quality assurance step should be made before providing files to the regulatory operations group for publishing and, ultimately, submission to regulatory agencies.

The utilities referenced in bold italics in the steps above, as well as other helpful utilities, are sponsor-defined utilities that are explained in more detail in Appendix B. Figure 3 illustrates the steps listed above.

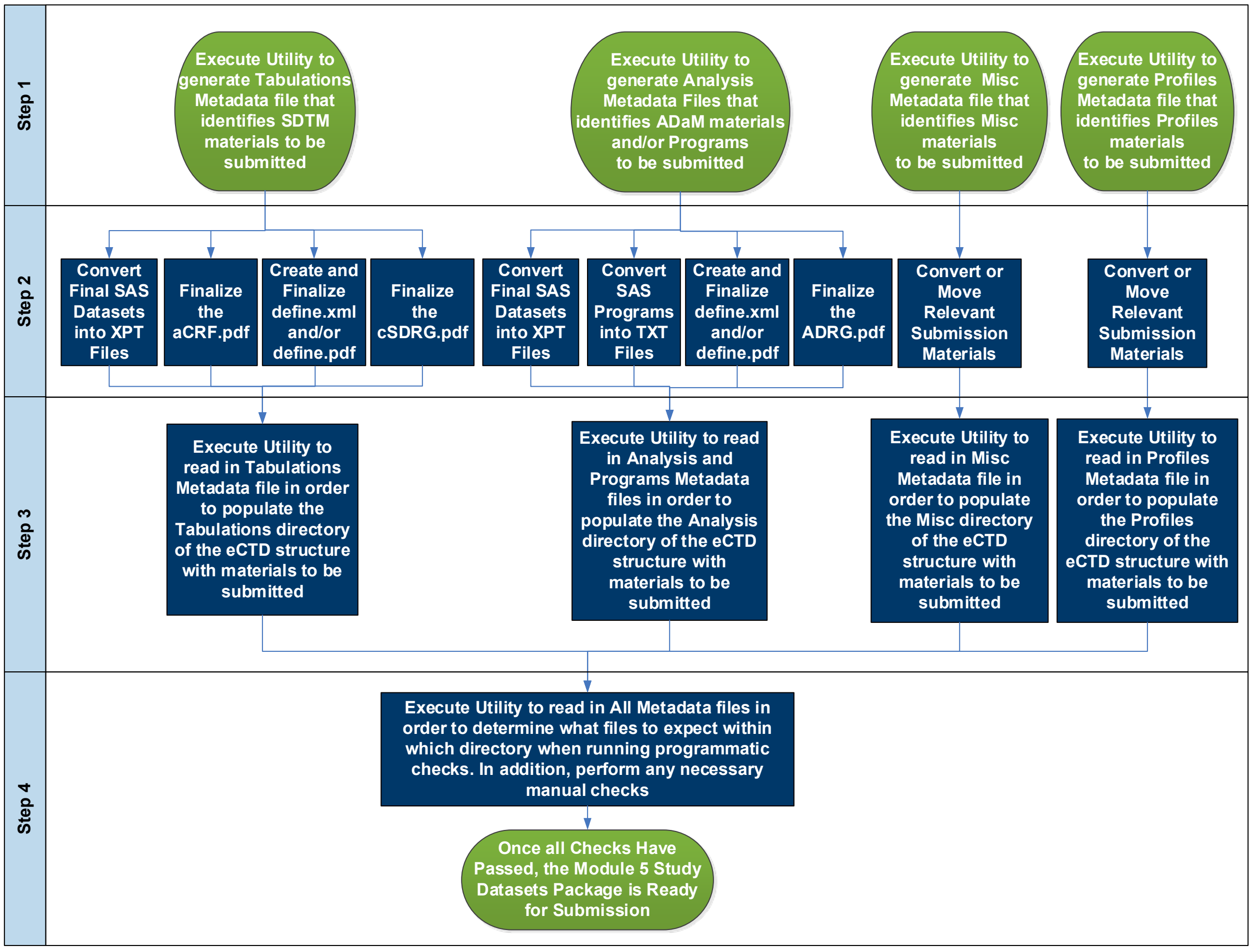

Step 1
Execute Utility to generate Tabulations Metadata file that identifies SDTM materials to be submitted
Execute Utility to generate Analysis Metadata Files that identifies ADaM materials and/or Programs to be submitted
Execute Utility to generate Misc Metadata file that identifies Misc materials to be submitted
Execute Utility to generate Profiles Metadata file that identifies Profiles materials to be submitted
Step 2
Convert Final SAS Datasets into XPT Files
Finalize the aCRF.pdf
Create and Finalize define.xml and/or define.pdf
Finalize the cSDRG.pdf
Convert Final SAS Datasets into XPT Files
Convert SAS Programs into TXT Files
Create and Finalize define.xml and/or define.pdf
Finalize the ADRG.pdf
Convert or Move Relevant Submission Materials
Convert or Move Relevant Submission Materials
Step 3
Execute Utility to read in Tabulations Metadata file in order to populate the Tabulations directory of the eCTD structure with materials to be submitted
Execute Utility to read in Analysis and Programs Metadata files in order to populate the Analysis directory of the eCTD structure with materials to be submitted
Execute Utility to read in Misc Metadata file in order to populate the Misc directory of the eCTD structure with materials to be submitted
Execute Utility to read in Profiles Metadata file in order to populate the Profiles directory of the eCTD structure with materials to be submitted
Step 4
Execute Utility to read in All Metadata files in order to determine what files to expect within which directory when running programmatic checks. In addition, perform any necessary manual checks
Once all Checks Have Passed, the Module 5 Study Datasets Package is Ready for Submission

Figure 3. Process Diagram for Generating a Dataset Submission Package

AN IN-DEPTH EXPLANATION OF THE PROCESS FOR GENERATING A DATASET SUBMISSION PACKAGE

STEP 1: EXECUTING THE UTILITY TO PREPARE METADATA FILES

The utility can determine which metadata file to create based on where the utility is executed. The programmer has the flexibility to choose which metadata file to create and not to create. Any issues like splitting datasets can also be handled in this metadata file but will require the programmer to manually modify the pre-populated metadata file before executing other utilities that are built around using these metadata files.

STEP 2: PREPARING SUBMISSION FILES

Many of these files have existing options to assist programmers in creating them, such as the define.xml. The utility ***prepxpt*** can be created to (i) copy the SAS7BDAT dataset files from a source location, (ii) convert those files to XPT, and (iii) output them into the appropriate directory within the eCTD structure. In this scenario, the utility would read in the relevant metadata file in order to do the following:

(1) Determine the source location of the files to be submitted
(2) Identify the datasets and programs for the submission
(3) Determine the destination location of where the files should be outputted to

In addition, utilities can also use the file type captured in the metadata file in order to determine if submission files need to be converted and moved or only to be moved. For example, a .SAS file would need to be converted to .TXT and then moved, but a .TXT file can be copied directly. This step to create the submission files is optimally performed in a development area.

STEP 3: COPYING FILES TO APPROPRIATE LOCATIONS

Assuming files were converted and created in a development area, another utility, ***mkmod***, can be executed to read in all available metadata files and auto-populate the eCTD structure in a production area with completed submission files.

STEP 4: RUNNING QUALITY ASSURANCE CHECKS

There are usually a list of quality assurance checks that are performed on the dataset submission package before finalization. Such checks can ensure that the submission does not violate any of the items defined in the FDA rejection criteria. Many of these checks can be automated via another utility. The utility ***chkesub*** would read in the available metadata files to ensure consistency and accuracy between original source and the final package, such that no files are missing and extraneous files get removed. Finally, when all of the quality assurance checks have passed, the dataset submission package is ready.

The utility ***mkmd*** can be built to output the relevant metadata file depending on where it is executed. The resulting metadata files can also come pre-populated with the source and destination locations as well as the file names. This is achievable if there is a standard location where relevant files are stored (e.g., datasets, programs). The destination location can be specified assuming that there is a copy of the eCTD folder structure in the working environment.

SUMMARY

The metadata files described within this paper amount to a road map of the study dataset submission package. These metadata files are leveraged throughout the process by utilities that move, convert and run checks on files and folders for creating and validating. This metadata approach facilitates quality assurance checks, allows for governance of a process, and ensures traceability of files in the dataset submission package.

As noted above, this process was developed assuming a Linux environment and clinical data (submitted as part of module 5) saved as SAS programs and datasets. These assumptions can be relaxed to permit usage in another operating system or on non-clinical data, or even using other statistical software (e.g., R).

REFERENCES

Study Data Technical Conformance Guide v 4.2.1 January 2019, https://www.fda.gov/downloads/drugs/guidances/ucm624623.pdf.

Providing Regulatory Submissions in Electronic Format — Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications Guidance for Industry January 2019, https://www.fda.gov/downloads/drugs/guidances/ucm333969.pdf.

Providing Regulatory Submissions in Electronic Format — Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act Guidance for Industry December 2014, https://www.fda.gov/downloads/drugs/guidances/ucm384686.pdf.

FDA Data Standards Catalog v5.2 December 2018, https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm#Catalog.

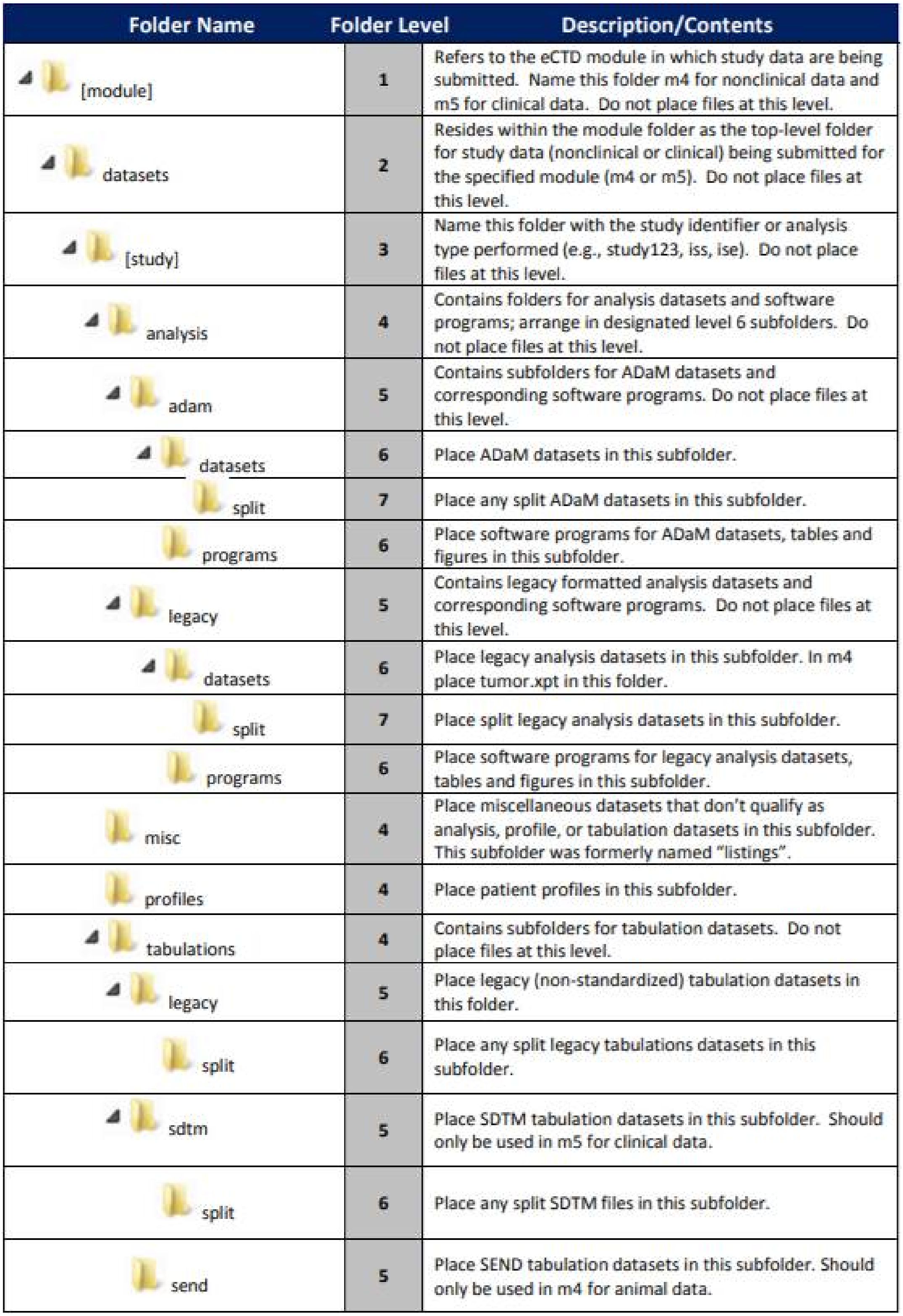
FDA Technical Rejection Criteria May 2018, https://www.fda.gov/downloads/drugs/guidances/ucm523539.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Janette Garner
Company: Kite Pharma, a Gilead Company
E-mail: janie_mei4985@hotmail.com

APPENDIX A – STUDY DATASET AND FILE FOLDER STRUCTURE AND DESCRIPTION



Folder Name	Folder Level	Description/Contents
[module]	1	Refers to the eCTD module in which study data are being submitted. Name this folder m4 for nonclinical data and m5 for clinical data. Do not place files at this level.
datasets	2	Resides within the module folder as the top-level folder for study data (nonclinical or clinical) being submitted for the specified module (m4 or m5). Do not place files at this level.
[study]	3	Name this folder with the study identifier or analysis type performed (e.g., study123, iss, ise). Do not place files at this level.
analysis	4	Contains folders for analysis datasets and software programs; arrange in designated level 6 subfolders. Do not place files at this level.
adam	5	Contains subfolders for ADaM datasets and corresponding software programs. Do not place files at this level.
datasets	6	Place ADaM datasets in this subfolder.
split	7	Place any split ADaM datasets in this subfolder.
programs	6	Place software programs for ADaM datasets, tables and figures in this subfolder.
legacy	5	Contains legacy formatted analysis datasets and corresponding software programs. Do not place files at this level.
datasets	6	Place legacy analysis datasets in this subfolder. In m4 place tumor.xpt in this folder.
split	7	Place split legacy analysis datasets in this subfolder.
programs	6	Place software programs for legacy analysis datasets, tables and figures in this subfolder.
misc	4	Place miscellaneous datasets that don't qualify as analysis, profile, or tabulation datasets in this subfolder. This subfolder was formerly named "listings".
profiles	4	Place patient profiles in this subfolder.
tabulations	4	Contains subfolders for tabulation datasets. Do not place files at this level.
legacy	5	Place legacy (non-standardized) tabulation datasets in this folder.
split	6	Place any split legacy tabulations datasets in this subfolder.
sdtm	5	Place SDTM tabulation datasets in this subfolder. Should only be used in m5 for clinical data.
split	6	Place any split SDTM files in this subfolder.
send	5	Place SEND tabulation datasets in this subfolder. Should only be used in m4 for animal data.

Figure 4. Study Dataset and File Folder Structure and Description (Reprint of Table 2 from Study Data Technical Conformance Guide v 4.2.1 January 2019, pg 33)

APPENDIX B – LIST OF PROPOSED UTILITIES TO SUPPORT THE METADATA APPROACH TO ELECTRONIC DATASET SUBMISSIONS

Table 1 below records the various utilities proposed withiyn this paper as well as provides a brief description of each utility. All of the utilities listed below, with the exception of ***mkmd***, would use the metadata file generated by ***mkmd*** to perform the specific task described. Please note that these are all sponsor-defined utilities and the name of the utility is not important. Rather, what matters is what the utility does.

Table 1. Utilities to Support the Metadata Approach to Electronic Dataset Submissions

Utility Name	Description of Utility
mkmd	Generates populated metadata files.
prepxpt	Prepares the XPT files from the SAS7BDAT files. Also generates a report alerting the programmer to any findings such as files ≥5GB depending on where it is executed.
sas2txt	Converts the SAS programs into legible TXT files and move these files into the appropriate submission material development area.
chkesub	Runs a detailed check on module5 and generates a findings report.
mkmod	Moves final versions of relevant files from the development area into the appropriate folders under the production area that is based on the eCTD structure.
prepesub	An umbrella script that reads in the metadata specs and determines the appropriate utility to execute in order to automatically populate the development area for submission materials upon execution.