Test Report for DS-XML Pilot

Center for Drug Evaluation and Research
(CDER)
Center for Biologics Evaluation and Research
(CBER)

April 8, 2015

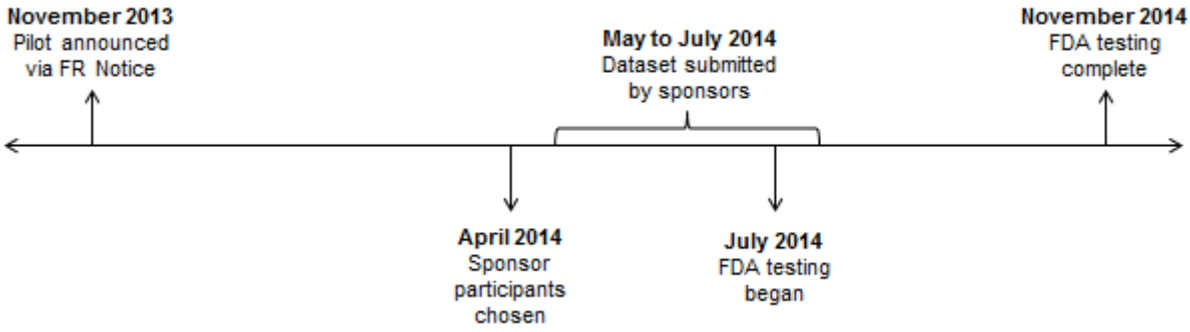
TABLE OF CONTENTS

1 Background

In the 1999 “Guidance to Industry: Providing Regulatory Submissions in Electronic Format” FDA recommended that regulatory submissions of clinical data to FDA utilize SAS Institute’s open transport called XPORT Version 5 format (XPORT)[1]. The XPORT format was developed in the late 1980s and there have been no version updates since 1999. Although XPORT is now considered by many to be an outdated transport technology it has been a basic and reliable method for transferring data across different hardware and operating systems. In recent years XPORT has proven to be a challenge due to a number of limitations, including: (1) fixed length variables, (2) variable name length (8), (3) variable label length (40), (4) character fields (200 bytes), and (5) alphanumeric character field names.

In November 2013, the Food and Drug Administration (FDA) issued a Federal Register (FR) Notice of a Pilot Project called “Transport Format for the Submission of Regulatory Study Data.” For this pilot, the FDA partnered with six sponsors to conduct an alternative analysis of an extension of the CDISC Operational Data Model (ODM)[2] XML, called Dataset-XML (DS-XML)[3] as a transport format for the submission of regulatory study data. The sponsors were selected from a list of companies that volunteered to participate in the pilot. Please refer to Figure 1 below for a timeline of the DS-XML pilot process.

Figure 1 – DS-XML Pilot Timeline



2 Test Objectives

The objective of this pilot was to test the transport functionality of DS-XML, which included ensuring that data integrity was maintained and that DS-XML format would support longer variable names, labels, and text fields.

[1] http://support.sas.com/documentation/cdl/en/movefile/59598/HTML/default/viewer.htm#creattrans.htm
[2] http://www.cdisc.org/odm
[3] http://www.cdisc.org/dataset-xml

3 Test Participants

Table 1 shows the participants in the testing process. It should be noted that while only six sponsors participated in the pilot, fourteen sponsors responded to the FR Notice request. Due to the overwhelming response, the FDA selected sponsors based on two factors: (1) the order in which their request was received, and (2) whether they had previously submitted one complete set of Phase 3 study datasets to the FDA.

Table 1 – FDA Test Participants

Role	Member Name
Participating Centers	▪ Center for Drug Evaluation and Research (CDER) ▪ Center for Biologics Evaluation and Research (CBER)
Participating Offices	▪ Office of Strategic Programs (OSP) ▪ Office of Business Informatics (OBI) ▪ Office of Translational Science/ Office of Computational Sciences (OTS/OCS) ▪ CBER Office of the Director (OD) ▪ CBER Office of Biostatistics and Epidemiology (OBE)

4 Test Methodology

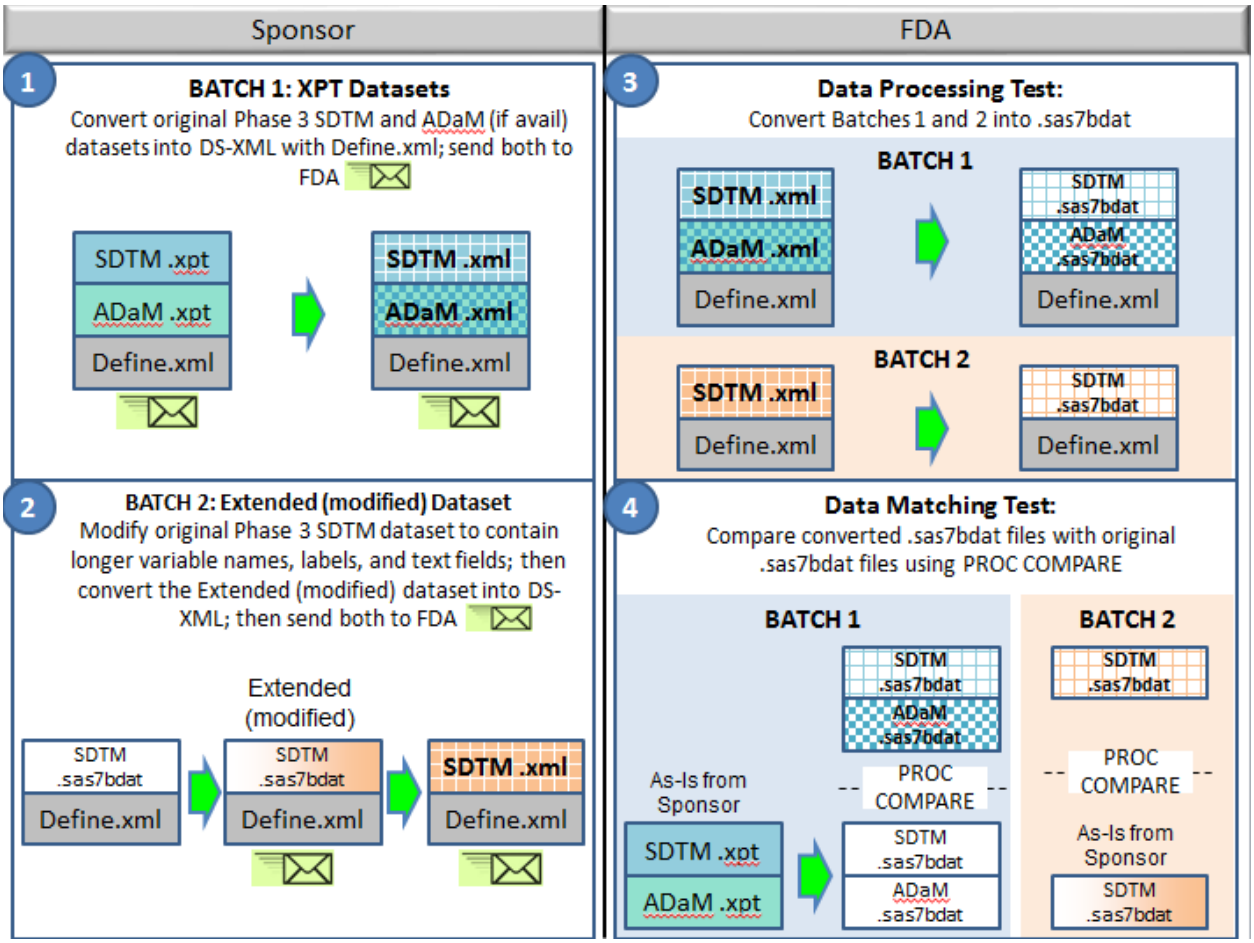
To meet the test objectives, two types of tests were performed on each dataset received from participating sponsors: (1) a data processing test and (2) a data matching test. Table 2 describes the expected outcomes for each test.

Table 2 – Pilot Tests and Expected Outcomes

Pilot Tests	Expected Outcomes
Data Processing Test	▪ The XML data can be converted to a sas7bdat file. ▪ The converted data is readable and can be viewed by FDA data analysis software (e.g., JMP).
Data Matching Test	▪ Data integrity can be preserved during transport from SAS to DS-XML and vice versa such that there is no loss of information, either metadata (data type) or variable values. ▪ The converted data from the Data Processing Test matches the original sas7bdat file.

Figure 2 provides additional context to the test methodology, describing how sponsors provided their datasets to FDA and how the FDA used the datasets for testing.

Figure 2 – Test Methodology



A description of each of the steps shown in Figure 2 is provided below. Note that all DS-XML datasets were compressed by the sponsor and submitted using .zip files via CD/DVD.

1. Sponsor converts the original Phase 3 SDTM and ADaM (if available) XPT datasets into DS-XML format with Define.xml and sends them to the FDA **(Batch 1)**.
2. Sponsor modifies the original Phase 3 SDTM SAS dataset to contain longer variable names, labels, and text fields. Sponsor then converts the new extended (modified) SAS dataset into DS-XML format and sends them with Define.xml to the FDA **(Batch 2)**.
3. FDA performs a **Data Processing Test on Batch 1 and Batch 2** by converting the XML datasets into .sas7bdat.
4. FDA performs a **Data Matching Test on Batch 1** by comparing the converted SDTM and ADaM SAS datasets (from XML) with the converted SDTM and ADaM SAS datasets (from XPT) using PROC COMPARE. FDA also performs a **Data Matching Test on Batch 2** by comparing the

converted SDTM SAS dataset (from XML) with the original SDTM SAS dataset using PROC COMPARE.

5 Test Results Summary

Table 3 provides a summary of the results for the XPT and XML datasets (Batch 1) and table 4 provides a summary of the results for the extended (modified) SAS datasets[4] (Batch 2). Please refer to Section 4 for a description of the expected outcomes for the data processing test and data matching test. A test is given a 'Pass' result if the expected outcome for that test was actually observed. In contrast, if the actual outcome was observed to be different than the expected outcome, then the test is given a 'Fail' result. A 'Pass' result in the table below constitutes a 'Pass' on both the data processing and data matching tests.

Table 3 – Batch 1 Results

Sponsor	Test Result
Sponsor 1	Pass
Sponsor 2	Pass
Sponsor 3	Pass
Sponsor 4	Pass
Sponsor 5	Pass
Sponsor 6	Pass
Total Test Cases	**6**

Note: The test period is July 2014 to November 2014.

Table 4 – Batch 2 Results

Sponsor	Test Result
Sponsor 1	Pass
Sponsor 2	Pass
Total Test Cases	**2**

[4] The extended (modified) SAS datasets are the modified .sas7bdat files with longer variables, labels, and text fields. This dataset was considered optional as the higher priority was given to testing the transport functionality of DS-XML.

Note: The test period is July 2014 to November 2014.

Four sponsors submitted the extended (modified) SAS dataset for Batch 2. However, only two sponsors provided the completed dataset with Define.xml file. Therefore, table 4 shows two test results.

Prior to testing, the expectation was that the DS-XML format would support longer variable names, labels, and text fields. The FDA tested this functionality with the extended (modified) SAS datasets from two sponsors. For both sponsor submissions, the DS-XML format was able to facilitate the longer variable names, labels, and text fields.

5.1 Other Observations

Table 5 below shows the original and post-conversion file sizes for each sponsor, including the percent change observed between XML file sizes and XPT file sizes. For all six sponsors, the file size increased after conversion to DS-XML.

Table 5 – File Size Comparison

Company	Original File Size (XPT)	XML File Size – after conversion	Percent Change
Sponsor 1	4.69GB	17.07GB	263.97%
Sponsor 2	53.6MB	187MB	248.88%
Sponsor 3	3.78MB	11.7MB	209.52%
Sponsor 4	2.58GB	2.77GB	7.36%
Sponsor 5	2.61GB	7.32GB	180.46%
Sponsor 6	1.55GB	1.75GB	12.90%

6 Issues

Table 6 provides a summary of the issues encountered during the testing period.

Table 6 – Summary of Test Issues

#	Test Batch	Issue
1	Batch 1	Testing was not successful initially due to a memory issue caused by the large dataset size. This issue was resolved after the SAS tool was updated which addressed the high memory consumption issue.
2	Batch 1	Testing was halted at the data conversion on two sponsors' datasets due to the following error: "Some code points did not transcode." This was caused due to non ASCII data in the datasets and the Define.xml. The two sponsors sent the updated data for re-testing and the issue was resolved.
3	Batch 2	The files were successfully converted except for one difference on a variable label. This issue was due to an error in the

		define.xml file and only occurred in one sponsor's dataset. Testing was completed with the awareness of the variable label difference.

7 Conclusion

In completing the data processing and data matching tests (see section 4, Test Methodology), the following outcomes were observed based on the test results:

- Based on the six study datasets submitted in Batch 1, it was observed that DS-XML can transport data and maintain data integrity.
- Based on the two extended (modified) SAS datasets that were submitted in Batch 2, it was confirmed that the DS-XML transport format can facilitate a longer variable name (>8 characters), a longer label name (>40 characters) and longer text field (>200 characters).
- Based on Issue 2 in Table 6, DS-XML requires stricter encoding in data.
- Based on issue 3 in Table 6, DS-XML requires consistency between datasets and Define.xml.
- Based on the file size observations, DS-XML produced much larger file sizes than XPORT, which may impact the Electronic Submissions Gateway (ESG) and may lead to file storage issues.

In summary, the purpose of the pilot was to conduct an initial analysis of CDISC's DS-XML as an alternative solution to the challenges of SAS XPORT V5 transport. Additional testing will be needed to evaluate cost versus effectiveness as an alternate transport format. As indicated previously in the FR Notice[5], FDA envisions conducting several pilots to evaluate new transport formats before a decision is made to support a new format. FDA would like to thank the sponsors who participated in the pilot as well as those who responded to the request. FDA would also like to thank SAS and CDISC for the support they provided during the pilot.

[5] https://www.federalregister.gov/articles/2013/11/27/2013-28391/transport-format-for-the-submission-of-regulatory-study-data-notice-of-pilot-project