Transport for the Next Generation

Version 1.0
Created 30 Apr 2017

A White Paper by The PhUSE Alternative Transport Format Working Group - Part of the PhUSE Emerging Trends and Technologies Computational Science Symposium Collaboration.

This white paper does not necessarily reflect the opinion of the institutions of those who have contributed.

Table of Contents

Introduction

PhUSE initiated The PhUSE Alternate Transport Format Project following discussions between the FDA and PhUSE representatives to come up with some suggestions for the replacement of the venerable (but seriously outdated) SAS® Version 5 (V5) Transport format (hereafter referred to as SAS V5 transport format). This format is the standard for the submission of clinical, nonclinical and analysis datasets to the FDA and other regulatory agencies.

The SAS V5 Transport format dates from 1989 and was first available as part of SAS version 5. Since that time, there have been many changes to the industry with respect to the process for submissions and the approaches to data curation and manipulation - but none to the format itself.

This SAS V5 transport format is commonly referred to as either “XPORT” (due to the LIBNAME keyword “XPORT” used during file creation) or “XPT” (due to the convention of using a file extension of “xpt”). This is a non-proprietary format with published specifications. It is useful in that a SAS dataset can be converted by SAS to a SAS V5 transport file for use outside of the SAS environment, e.g. there are non-SAS applications which can import SAS V5 transport files. Similarly, non-SAS applications can create files which follow this non-proprietary format and these can then be converted by SAS to SAS datasets.

NOTE: There is a proprietary format related to the SAS CPORT/CIMPORT procedures. This is useful for migrating SAS datasets across platforms and SAS versions within a SAS environment. This format does not have published specifications and CPORT files are not accepted by the FDA. Therefore, the CPORT format is completely outside the scope of this white paper.

This white paper will aim to cover the following aspects:

- building the case for replacing the SAS V5 transport format for submission data
- enumerating 'pain points' of the SAS V5 transport format
- desirable attributes for a transport standard that would serve both current and future needs

In order to ease the process of delivery, the team decided that the project proceed in two phases;

- Phase I - build the case for replacing the standard transport format and enumerate the characteristics of a replacement transport format.
- Phase II - using a standard dataset, generate alternate formats and then use the criteria identified as part of Phase I to score each representation. The SAS V5 transport format will also be scored using the same criteria in order to try and be as objective as possible.

Background

This section will cover the current state of the industry. We will explore the dominating motivations for investigating a replacement for SAS V5 transport format as well as cover some of the prior investigations, pilots and outcomes.

In order to represent the issues with the SAS V5 transport format, we have gathered and summarised currently held opinions and observations by people who are using the format. We have categorised the observations into four main categories; these categories will be used to group the suggested criteria for a replacement for SAS V5 transport format.

Issues with the Current Transport format

Following are an examination of the issues with the current SAS V5 transport format that prompted a reevaluation of the transport mechanism adopted across the industry. In general, there are 4 types of issues that can be identified as part of the existing standard; Data File Format, Storage, Content and Extensibility.

Data File Format

- Limited Variable Types; the current data formats supported are limited to US ASCII (for Character formats) and IBM INTEGER and DOUBLE (for Numeric formats).
- Only supports US ASCII Character encoding. No multibyte characters are possible; this requires translation and/or transcription from the source data.
- Field names are restricted in terms of width and format. Field names must be alphanumeric, Variable names are limited to a maximum length of 8 characters, Variable labels are restricted to a maximum length of 40 characters.
- Character field widths are limited to 200 characters.

Storage

- Does not make efficient use of storage space. There is often empty space for columns allocated, but not used by data and this can lead up to 70% wasted space.
- The inability to compress datasets leads to significant file logistical issues, due to the requirement that the maximum size of the files is 5 Gigabytes or smaller.

Content

- The format is only suited for transporting and storing two-dimensional data structures. This restricts the structure of the content that can be transported.
- The two-dimensional nature of the transport format has led to sub-optimal designs in the structures of the content models that are used to store and transport clinical and nonclinical data.
- Lacks a robust metadata layer, relying on external files such as the define.xml to provide the missing data for comprehensive study data review. This requires that multiple files are kept synchronised, often in different locations.

- There is no concept of user tracking, such as an audit trail within the format itself.

Extensibility

- SAS V5 transport format is not an extensible modern technology.
- Creating SAS V5 transport format files from SAS datasets is a standard part of sponsor/CRO workflows using common industry tools.

FDA Public Meeting on Study Data Exchange

In November 2012, FDA held a public meeting entitled: "Regulatory Drug Review: Solutions for Study Data Exchange Standards". The purpose of the meeting was to solicit input from external stakeholders regarding the advantages and disadvantages of current and emerging open, consensus-based standards for the exchange of regulated study data. The author of this summary attended that meeting and reviewed relevant background materials accessible on the FDA website.[1] The purpose of this summary is to inform the ongoing effort within PhUSE to issue a white paper with recommendations on this topic.

The meeting took place at the FDA White Oak campus, 11/05/2012 10am-4pm. The agenda had 4 parts:

- Meeting Introduction and Overview
- FDA Review of Current Environment and Challenges
- Study Data Exchange Solutions
- Discussion.

Mary Ann Slack from the CDER Office of Special Programs opened the meeting. She stated that the current exchange format, SAS Transport version 5 is an old format, has known limitations and is not extensible. In addition to well-known technical limitations, important relationships between data points are not well captured in a tabular data structure. The FDA seeks public comment on alternatives, recognizing that any replacement solution will take time to implement. Any solution will need to meet FDA requirements.

To illustrate FDA data needs, draft scenarios were provided. Additional presentations from Doug Warfield, Armando Oliva, M.D., and Chuck Cooper M.D. further described current limitations. Functional requirements focused on the need for [1] audit trail (i.e. provenance) information to better understand the data and any changes, transformations, etc. as the data progress through the data lifecycle from collection to submission, [2] greater flexibility to implement new content requirements, including minimizing costs and time to implement new versions, [3] better support for data integration downstream, realizing the increasing need to integrate data from multiple sources, [4] robust metadata exchange to improve understanding the data. The last two points indicate the need for increased computable semantic interoperability.

Overall, five proposed solutions emerged:

1. SAS Transport V5 Extensions: Bill Gibson (SAS)
2. CDISC Operational Data Model (ODM): David Gemzik (Medidata) and Wayne Kubick (CDISC) and Fred Wood (Octagon)
3. HL7 version 3 including Clinical Document Architecture (CDA): Armando Oliva, M.D. (FDA)

---

[1] See http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm332003.htm and http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM332947.pdf (last accessed 2016-03-28)

4. Semantic Web Standards, including the Resource Description Framework (RDF) and Web Ontology Language (OWL): Mathias Brochhausen (University of Arkansas), Charlie Mead (W3C)
5. Analytical Information Markup Language (AnIML): Gary Kramer (ASTM)

Highlights of the discussion were the following:

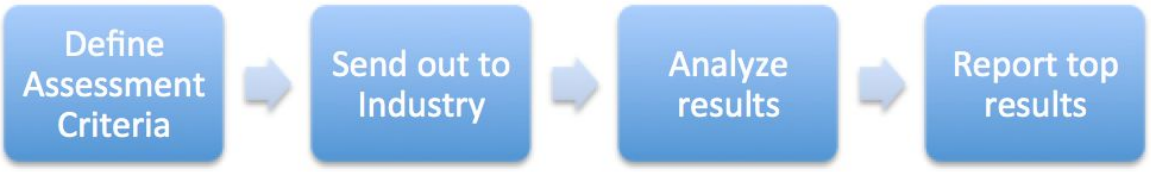
- SAS V8 Transport Format: There was interest in exploring an extended SAS V5 format which would allow for SAS Version 8 datasets to be transported; this would be a short term solution for technical limitations to the current format. It's clear that it would not solve the structural limitations and a longer-term solution would also need to be identified. Attendees discussed that this would be a lower level of effort to assess as a short term solution but more information is needed.
- ODM: There was strong support from the audience to see the FDA conduct an ODM pilot. Attendees discussed ODM as good option but it was clear that there are known challenges to be addressed (e.g., lack of information model, not ISO 21090 compliant, problem with relationships).
- Requirements: Clear business requirements, particularly from the FDA, are needed so that alternatives can be assessed objectively and a decision can be made. There was interest in knowing what pilots are ongoing at the FDA. Metrics from a pilot or a comparative pilot are also important to assess the economic impact of a change.
- HL7: An attendee discussed the challenges in identifying available resources with the necessary expertise for HL7 implementations.
    - One attendee expressed that we should harmonize with EHR standards (to leverage the investment in HL7 in healthcare and enrich the Study Data Tabulation Model (SDTM) content), and minimize data management and exchange burden with investigators and electronic health record (EHR) systems/vendors.
    - Numerous speakers advocated not using HL7, referencing its complexity and lack of experience with the standard, and the ongoing investment made with CDISC implementations.
- RDF/OWL: There was general interest in semantic web, but more information is needed to better understand its potential use.
- BRIDG: There was feedback that Biomedical Research Integrated Domain Group (BRIDG) was not discussed enough and its role in a solution.

SAS V8 transport format

The SAS V8 transport format was created to address some of the issues raised as part of the FDA Public Meeting on Study Data Exchange. The macros to generate the expanded format were released in 2012 and are supported in versions of SAS 8.2 and above. Some of the currently held observations by those using SAS V5 transport format have been addressed in the SAS V8 transport format, e.g. longer character fields. (Note: the SAS V8 transport format has sometimes been referred to as “SAS Transport (V5) Expanded”.)

Methodology

The group needed to be able to objectively evaluate any new and existing transport formats according to a specific set of use cases contextualised by the industry. In order to avoid any bias, the use cases prepared by the agency could not be shared with the working group, so a different approach was adopted.



The team elected to take the following approach:

- Define a set of criteria that could be used to evaluate the suitability of any given format.
- In order to make the criteria as useful as possible it was decided that there was a need to quantify the coherence of a proposed format with the criteria.
- In order to further contextualise the results, the collected criteria were included as part of a questionnaire which was submitted to the industry for prioritisation.
- The responses to the questionnaire were analysed and used to develop a weighted index of assessment criteria that can be included as part of a formal evaluation of alternative formats for the transport of clinical datasets.

Evaluation criteria for new transport format

Method

The criteria were developed by a cross functional team representing a cross spectrum of stakeholders; including regulators, biopharmaceutical companies, software vendors and contract research organisations.

The team recognised the importance of semantic precision and defined a set of principles to be used for developing the criterion. The principles were defined as follows:

- It should be possible to unambiguously define the criterion in one or two sentences

- It should be possible to use the definition in such a way to define a quantitative or qualitative measure with which to assess relative compliance of a given format to the criteria.
- It should be possible to provide at least one example where there is a conformance with the criteria in a real world example.

Criteria

Definitions

- Encapsulated Data - implies that the transmitted data is manipulated in a non-destructive manner with the necessary metadata in a header of a given package for sender and receiver to understand and process.
- Data Provenance - refers to the ability to trace and verify the origin of data, as well as how and by what systems it has been altered since its origination.

Data File Format

- File Size
    - Definition: For a given dataset the comparative size of the transport package (in bytes)
- Compression
    - Definition: Does the transport format support compression and decompression of encapsulated data using standard open compression formats?
- Encryption
    - Definition: Does the transport format support the encryption of encapsulated data using industry standard algorithms (including PKI)?
- Digital Signature
    - Definition: The transport format will support the application of one or more digital signatures on encapsulated dataset
- Data integrity
    - Definition: The transport format will support a hash or checksum function to mitigate unexpected data changes
- Schema driven
    - Definition: The transport format should support a schema to ensure that a data transport file will be well-formed and valid.
- Well defined Metadata
    - Definition: The transport format will support a set of well-defined metadata tags that allow effective communication of encapsulated data between sender and receiver.
    - Examples: Encryption used, record number, subject UUID, etc.
        - A use case for this would be partitioning study datasets for an interim subject transfer and having enough metadata to reconstitute the original study

        - Sending partial datasets for a subject
        - Incremental or cumulative data transfers
- Wide Payload Support
    - Definition: The transport format may support transfer of a wide range of well-defined payloads over and above data currently well-described using tabular data structures.
    - Examples:
        - Image data, DICOM data, WAVEform, RDF [Ask Armando]
        - Protocol (electronic)
        - Statistical Analysis Plan (electronic)
- Relationship Data
    - Definition: The transport format will support meaningful relationships between data.
    - Example: Replace RELREC with metadata laden links for relationship between clinical observations and histopathology findings.
- Partial Data Transfers
    - Definition: The transfer format should support the transmission of subsets of data in a meaningful fashion.
    - Examples: (this should be linked with the well defined metadata)
        - Transmitting data on a subject level
        - Transmitting all data for a given time period across multiple subjects on request
        - Transmitting incremental datasets
- Must be an Open Standard
    - Definition: The full transport format specification is freely available, well documented and allowed for free use without license. All supporting materials (eg schemas, documents) will be available without cost.
- Should support multibyte character encodings
    - Definition: The transport format supports the fidelity of captured source data in transmission without requiring translation or transcoding. The encoding of a transport file should be declared by the format. Restrict support to UTF-8 encoding.
    - Example:
        - Should support submissions in Kanji for Japanese Studies
- Audit records
    - Definition: The transport format should support the transport of audit data/metadata.
    - Example:
        - The CRF-level audit trail should be able to be transported as part of an end-to-end submission
        - Something similar to the capability present in the ODM
- Traceability and Provenance
    - Definition: The transport format should support the transport of traceability data and metadata to establish data provenance.
    - Example:

        - For a given data value in a submission analysis dataset it will be possible to trace back to the original source of data including transformations and/or computations (eg. age is based on birthdate and study start date).
- Transmit data and metadata
    - Definition: It will be possible to transfer both data and metadata in the same transport file.
    - Example:
        - In a given data transfer incorporate both the metadata and data, and link from data elements to corresponding metadata.

Value

Costs of adoption

- Definition: The transport format should represent a net positive return on investment for adoption
- Resource costs - cognitive load for personnel
    - Definition: The transport format should be sufficiently familiar to not require large costs of training and utilisation
    - Example:
        - The transport format should support transport of tabular datasets
        - The transport format should be simple to build (e.g. PROC ALTRANS)
- Resource costs - storage/transport
    - Definition: The choice of the new transport format should not incur large increases in costs for processing, sending and storing data held in the format.
    - Example:
        - A substantial increase in file size would increase costs of hard drive space and bandwidth for transmitting.
        - Complex encryption mechanisms might incur a processing cost for unencrypting at each stage of data creation and review
- Resource costs - software
    - Definition: The adoption of the alternative transport format will not require a large capital outlay for software to build and manipulate the data format. It should be supportable using existing data management systems
    - Examples:
        - Compatible with ODM systems (e.g. XML based or similar)
- Cost of Format adoption for generation and processing of clinical data.
    - Definition: The time taken to get a submission to regulators and for regulators to be able to initiate and complete review should not be impacted by the adoption of the new transport format.
    - Example:
        - There will be a minimal cost in time for generation of data in the new format, relative to the existing standard.
- Value of adoption of new transport format
    - Definition: Time to review of submission should decrease because of better expressivity and improved quality of datasets

    - Example:
        - The format should support self-validation for identification of common submission issues
        - Time spent recreating full context datasets should decrease
- Validation of capability of new format
    - Definition: Tools exist that are capable of validating the content of transport files against CDISC implementation guide rules. These rules include data format standard rules and data domain context rules. Any new transport format would need tools to product similar validation.
    - Example:
        - Value not found in non-extensible code list
        - Missing data for *--STRESC* when *--ORRES* is provided.

Content

- Definition: Changes to the content model that will deliver benefits for adopters.
- Able to represent relationships in the data without requiring duplication within a single data transfer
    - Definition: The ability to indicate relationships between elements within a encapsulated dataset. The relationship should also be able to be annotated (e.g. reason for ascribing relationship)
    - Examples:
        - Represent the causality for a given concomitant medication with respective to one or more adverse events (and vice versa)
        - Actions taken on Adverse event, for example hospitalisation
        - Represent the connection between the administration of an intervention and the subsequent timed observations of the subject
        - Refine model to avoid duplication of data, context, metadata
- Able to represent relationships to external resources
    - Definition: It will be possible to link encapsulated content to external resources such as standard controlled terminology
    - Examples:
        - Link to Controlled Terminology using resource URI
- Tabular Data Representation
    - Definition: Encapsulated content should support tabular representations of data
    - Example:
        - It will be possible to represent legacy datasets.
        - Transform data into tabular data structure.
- No field width restrictions
    - Definition: The transport format will support arbitrary width fields. Format should allow declaration of width for the purposes of content validation.
    - Example:
        - Data should not need to be truncated for transport
        - Data should only occupy as much space as needed (not fixed width)
- More discrete datatype definitions

    - Definition: Transfer format will support additional datatypes than existing than CHAR/NUM, eg XML Schema Definitions.
    - Examples:
        - Date
        - Time
        - Datetime
        - Datetime with timezone
        - Integer
        - Float
        - Bool
- Transactional Data Model
    - Definition: The content model will support the expression of transactional data for a data submission if requested.
    - Examples:
        - Reflect changes to data to reflect findings of a data safety monitoring board
- null Flavour support
    - Definition: The content model should support something similar to the null flavour in ISO21090 datetypes
    - Examples:
        - A missing value should have a qualifier to indicate reason for absences (eg not given, refused)
        - This is currently absent from the SDTM model

Compatibility/Extensibility

- Backward compatibility
    - Definition: New transport format will be capable of being transformed to and from existing transport format
    - Examples:
        - Decompose defined data types to CHAR/NUM
        - Truncate variable length fields to fixed length fields and SUPPQUAL
        - Translate discrete relationships to RELREC where possible
        - Note that this would not accommodate loss through UTF-8 -> US ASCII
- Compatibility with existing Health data standards
    - Definition: It should be be compatible with existing standard healthcare formats
    - Examples:
        - Transform to and from ODM (including dataset-XML)
        - Transform to and from HL7 C-CDA
        - Transform to and from BIMO
- Projected Lifespan of Standard Support
    - Definition: The transport format should be supported by a non-commercial industry body with a mandate for a minimum length of time of full support for the transport format. This may depend on the age of the existing standard
    - Example:

        - Consider CDA vs FHIR, will both standards continue to exist in active development or will one supplant the other? Will the standard owner continue to maintain support and development?

- Extensibility
    - Definition: It should be able to accommodate new content requirements easily, cost-effectively, and retain backwards compatibility (i.e. no or minimal need to modify data management tools or processes). This implies support for namespaces.
    - Example:
        - Addition of custom attributes peculiar to a system adopting the standard
        - Systems naive to an extension will not be affected by use of extension

Questionnaire Design

In order to get a fair balance of the responses, the criteria were partitioned into three top-level categories:

- Implementation details - for this section every criteria needed an adjudication of Important or Not-Important
- Future Capabilities - for this category the responders needed to identify up to five criteria that would be important to them
- Process and Change - for this category the responders needed to identify up to four criteria that were important to them in their job roles

For the purposes of categorisation of responses, there were additional questions around industry segment, relevant experience and interest in further participation.

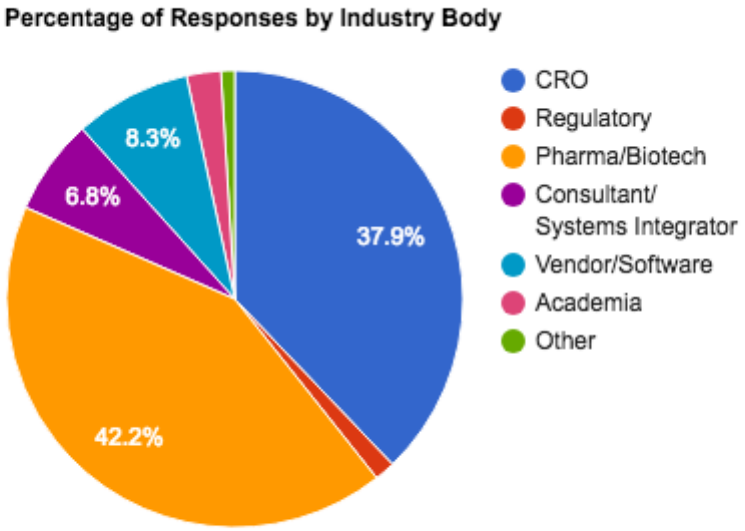
Questionnaire Response

The questionnaire was prepared using SurveyMonkey (www.surveymonkey.com) and publicised using direct mail requests through both the PhUSE membership list, direct mail via CDISC membership mailing lists, as well as via posts on the social networking sites; Twitter and LinkedIn.

Despite the wide coverage, the response rate was low with a total response size of **205**. This raises a very interesting question as to what the actual significance of the perceived shortcomings of the existing format is to the industry as a whole at this time. Hypothetically, if this was a mission critical issue then the number of responses would be in the range 40-60% rather than a ~ 0.05% response observed.

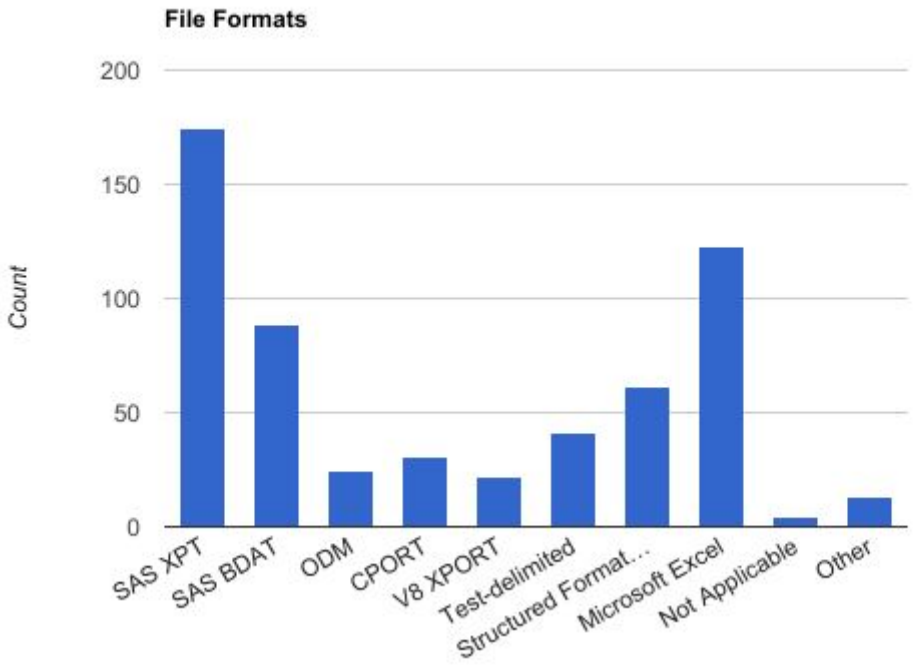
With that taken as read; the directionality of the results may be useful for the industry. With the low response rate as a caveat, the group shall present what results were garnered.

In terms of the distribution of responses, the organisational breakdown of respondents is as follows:



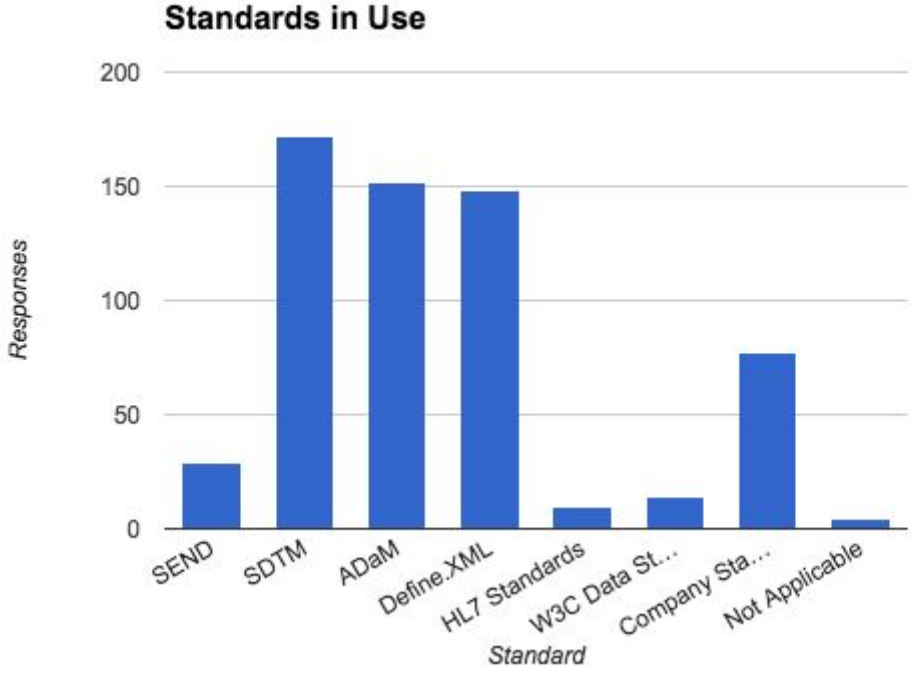
Unsurprisingly, the dominant responders are Pharma/Biotech and CRO. There were some responses from individuals identifying themselves as Regulatory; the actual count was three. The sample size is too small to draw any significant conclusions based on this for the wishes of the regulatory bodies.

The questionnaire aimed to identify what file formats the responders had experience with and used on a regular basis. In terms of responders current experiences with file formats the following results were identified:



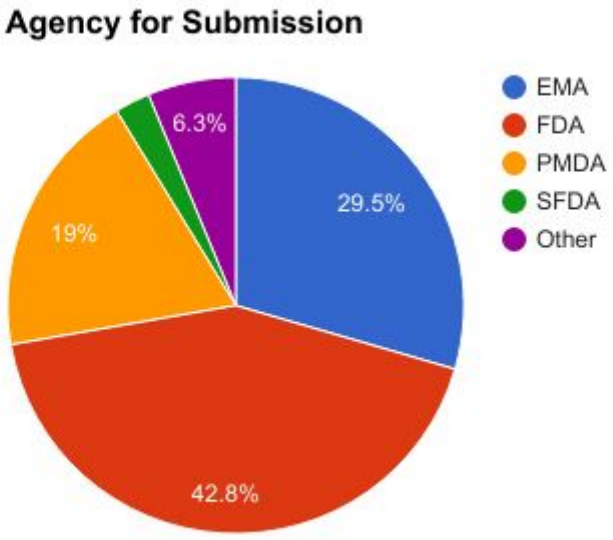
The responses are fit nicely with the commonly held views with respect to the processes and technologies used in industry. There is perhaps a smaller than expected level of adoption of the V8 XPORT format - however as discussed previously there can be some ambiguity in individual's understanding of what version of SAS Transport is in use.

It was also considered important to identify what areas of experience responders had with respect to data standards (both global and company); the following results were obtained:



Unsurprisingly SDTM was the most common standard in use. Interestingly, a large proportion of the responders stated that they were using ADaM, which seems counter to commonly held views on ADaM adoption.

When considering the agency to which submissions are routinely being made, the following breakdown was observed:



There was a good distribution of agencies for submission. Unsurprisingly, the FDA was the dominant agency in this target population. This underlies the importance of the FDA in taking a lead in this topic.

Analysis of the Criteria

The analysis was normalised to take the top ten categories across each of the partitioned response set. Some of the criteria were replicated across the different categories; the following top responses were observed for the three categories for the entire response group

Implementation	**Future Capabilities**	**Process and Change**
File Size	Transport format metadata	Transmit data/metadata
Compression	Transmit Data/Metadata	Traceability/Provenance
Encryption	Traceability/Provenance	Extensible Standard
Digital Signature	Compatible with existing data standard	Supported by Software
Data I ntegrity	Built in support for encryption	Support multibyte character encodings
Schema Driven	Represent Data relationships	Backward Compatible
Transport Format Metadata	Partial Data Transfers	Low transition cost
Must be an Open Standard	Built in support for data auditing	Internal Data Relationships
Should support multibyte character encodings	Digital Signing of Datasets	Shorter review time
Better Data Types	Diverse Payload Support	Transmit existing data structures

There are some deeper analyses of the responses included in the appendices. The authors stress to remind the reader that the sample size is 205.

Conclusions

Based on the small response rate, it could be concluded that the issues identified herein are not sufficiently impacting on business efficiency to prompt an immediate need for a replacement file format for the SAS V5 transport format. The only reason we could see for a new file format is if one or more of the criteria identified were deemed essential by the regulatory agencies; this would then create a suitably motivating force to justify the expending of the effort required.

Based on this, the steering committee do not recommend a formal extension of the project to encompass the goals detailed as part of Phase II at this time. This white paper will exist as a record of the domain analysis for future teams to reference.

Based on this we have the following high-level recommendation:

Adopt to V8 XPT as a standard:
This resolves the following issues with V5 XPT

1. Resolves 200 byte/character limit
2. Allows for long variable names and labels
3. Existing solution for SAS data set read/write (production since 2012) with limited changes needed for other packages (maybe none for some)
4. Fits in with other submission artifacts like Define-XML and eCTD structures (i.e. direct replacement for V5 XPT)
5. Files are usable without additional metadata (limited data set-level tabular metadata)
6. File size will be the same as existing solution
7. Is still an open format

It will not address the following:

1. Is not schema driven
2. It will not directly support digital signatures.
3. It is restricted to a US-ASCII format and will not address support for multibyte characters.

The working group still feels that given this is the second investigation into replacing the XPT that there is some unmet needs that are motivating initiatives to identify a format that will continue to support requirements of an increasingly technologically oriented industry. As such, it is expected that more extensible submission formats may be investigated as part of other working groups looking at new formats and data structures, such as Linked Data.

Consideration of a future replacement or augmentation of either SAS transport format should consider the full list of criteria, focusing on those that are not addressed by the existing transport format; such as schema-driven (extensibility), file size/compression, digital signature/encryption/data integrity and partial datasets.

It is also likely that this future work will encompass more than a direct replacement of tabular data movement. Other parts of the submission package like eCTD, Define-XML, and supporting documentation will likely be impacted by any new proposal.

Acknowledgements

The PhUSE Alternative Transport team will like to acknowledge the valuable contributions of the following people in the scoping, development of content, analysis, authoring and review of this document:

- Scott Bahlavooni - d-Wise/PhUSE
- Daniel Christen - SAS
- Joris Derks - OCS Consulting
- Dirk Engfer
- Ian Fleming - d-Wise/PhUSE
- William Frank - HHS
- Bill Gibson - SAS
- Bob Friedman - Xybion
- Dave Iberson-Hurst - Assero
- Wayne Kubick - HL7
- Geoff Low - Medidiata Solutions/PhUSE
- Andrew Newbigging - TrialGrid
- Armando Oliva - Semantica LLC
- David Pressley - United Therapeutics
- Eric Simms - GSK
- Ian Sparks - TrialGrid

We would like to acknowledge the support of both the PhUSE CSS Steering Committee and PhUSE CSS Project management without whom this paper would not have been possible.

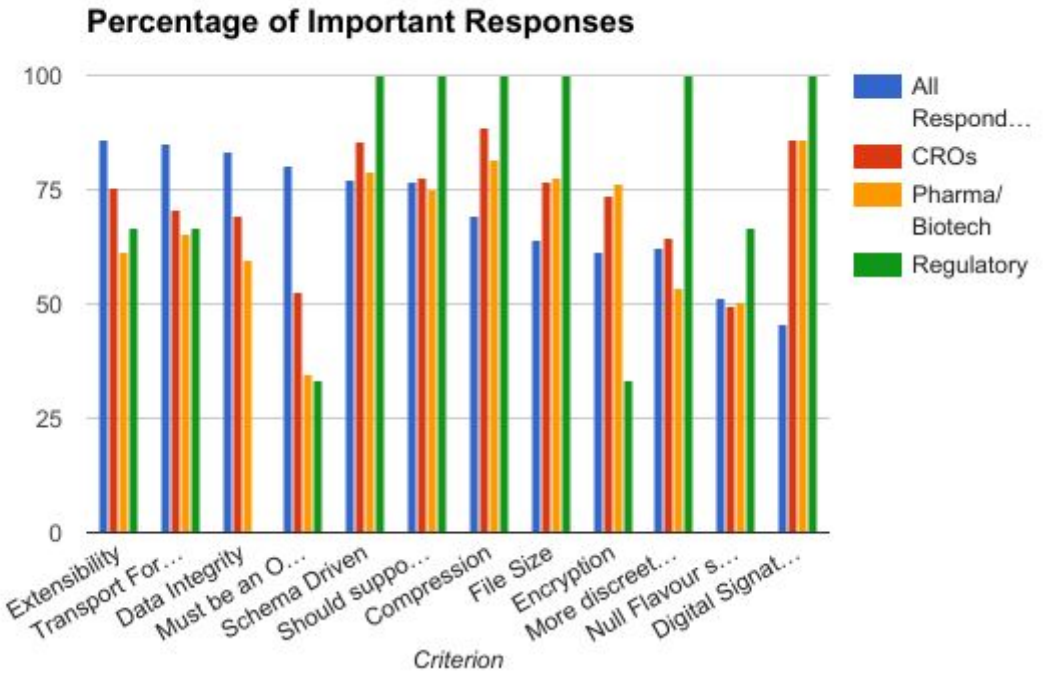
Appendices

We present some analysis of the questionnaire response dataset for review and comment.

Analysis - Implementation Characteristics

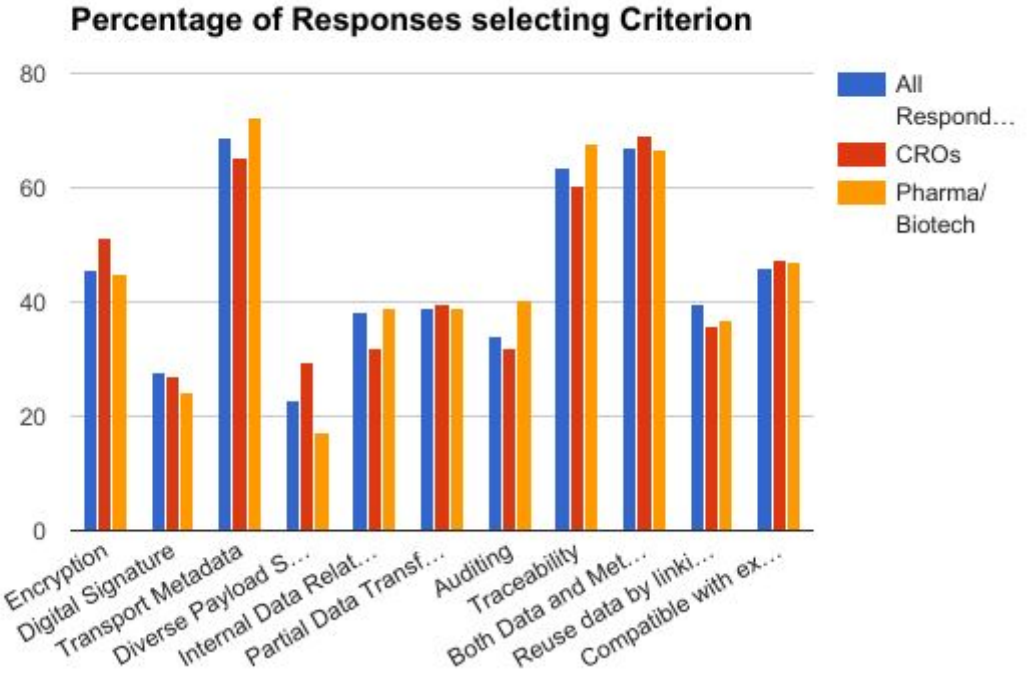
When considering the Implementation aspects, the following comparative results are seen:

Criterion	**All Responders**	**CROs**	**Pharma/Biotech**	**Regulatory**
Extensibility	85.8	75.6	61.6	66.7
Transport Format Metadata	85.3	70.5	65.5	66.7
Data Integrity	83.3	69.2	59.5	0
Must be an Open Standard	80.2	52.6	34.5	33.3
Schema Driven	77.1	85.5	79.1	100
Should support multibyte character encodings	76.6	77.6	75	100
Compression	69.3	88.5	81.4	100
File Size	64.2	76.6	77.6	100
Encryption	61.6	73.7	76.5	33.3
More discrete datatype definitions	62.5	64.5	53.6	100
Null Flavour support	51.2	49.4	50.6	66.7
Digital Signature	45.8	85.9	86	100



Analysis - Future Capabilities

	All Responders	CROs	Pharma/Biotech	Regulatory
Total Responses	206	78	87	3
Encryption	45.6	51.3	44.8	0
Digital Signature	27.7	26.9	24.1	33.3
Transport Metadata	68.9	65.4	72.4	33.3
Diverse Payload Support	22.8	29.5	17.2	0
Internal Data Relations	38.3	32.1	39.1	66.7
Partial Data Transfers	38.8	39.7	39.1	0
Auditing	34	32.1	40.2	33.3
Traceability	63.6	60.3	67.8	33.3
Both Data and Metadata	67	69.2	66.7	33.3
Reuse data by linking	39.8	35.9	36.8	66.7
Compatible with existing standards	46.1	47.4	47.1	66.7



Analysis - Process and Process Change

	All Responders	CROs	Pharma/Biotech	Regulatory
Internal Data Relationships	32.5	29.5	29.9	66.7
Support multibyte character encodings	40.8	35.9	39.1	33.3
Audit records	26.2	21.8	29.9	33.3
Traceability/Provenance	57.3	56.4	60.9	66.7
Transmit data/metadata	59.2	60.3	52.9	33.3
Low retraining cost	23.3	28.2	21.8	33.3
Supported by Software	44.2	48.7	46	33.3
Low transition cost	34	29.5	41.4	33.3
Shorter review time	31.6	33.3	27.6	33.3
Support Tables	28.2	20.5	34.5	0
Backward Compatible	35.4	43.6	39.1	0
Long term Support	24.8	26.9	21.8	0
Extensible Standard	55.3	55.1	50.6	0