PhUSE US Connect 2018

Paper DH06

Automated Creation of Submission-Ready Artifacts

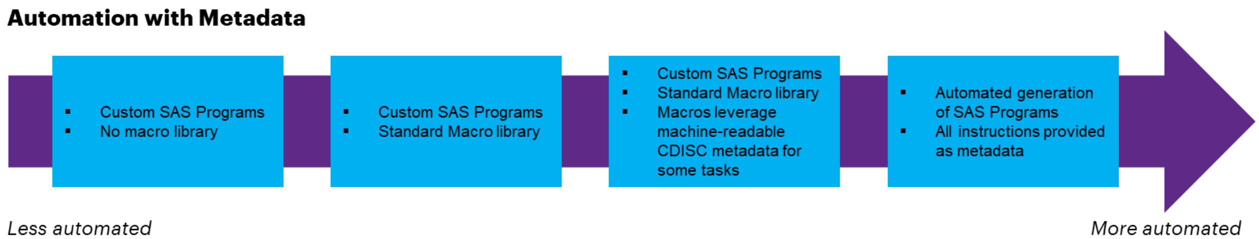
Silas McKee, Accenture, Pennsylvania, USA
Lourdes Devenney, Accenture, Pennsylvania, USA

ABSTRACT

Despite significant progress towards the standardization of clinical data acquisition, tabulation, and analysis, the creation of submission-ready artifacts (datasets, tables, figures, and listings) remains a largely manual process. We suggest that generation of submission-ready artifacts can be efficiently automated by leveraging metadata standards. Our solution relies upon a two-step process wherein a code-generating engine interprets standard metadata libraries to generate executable SAS® programs. The resultant SAS programs are then executed to generate submission-ready artifacts. This approach has several distinct advantages. First, it drastically reduces the need for custom programming during the generation of standard artifacts. Second, the generated SAS programs are intelligible to both statistical programming teams responsible for generating the artifacts and health authorities responsible for reviewing the submitted artifacts. Finally, the solution reduces the time to produce and validate submission-ready artifacts. We describe herein our approach to automation, and provide early results from pilot studies.

INTRODUCTION

Automated generation of submission-ready artifacts remains a challenge. The approach to automation taken in this paper is generally referred to as 'metadata-driven' processing. The concept has been around for some time, and there are differing interpretations. We believe it is useful to think of metadata use in the creation of SAS programs along a continuum. On the left side of the continuum lies complete custom development; a programmer writes a complete SAS® program to produce the required data deliverable. On the right lies a 'true' metadata driven process; the programmer provides all instructions through metadata, and there is no custom programming.



In this paper, we describe AMDP (Automated Metadata-Driven Processor), an automation tool that lies to the right on the metadata-driven continuum. AMDP can generate SDTM Datasets, ADaM Datasets, Tables, and Listings based solely on metadata. There are a few requirements to enable this automation. The first is adherence to a data acquisition standard. Metadata-driven automation requires consistently defined eCRF metadata as an input. Second, the solution depends on a set of a robust metadata defined for each artifact to be produced. This paper explores our approach to metadata-driven processing, details the architecture and capabilities of our current implementation, reports initial results from a pilot study, and discusses future areas for development.

METADATA DRIVEN PROCESSING APPROACH

Before delving too deeply into our approach to metadata-driven processing, we need to briefly explore the term metadata. Metadata, at the most basic level, means 'data about data.' Industry professionals are generally familiar with the CDISC metadata: Name, Label, Comment, Origin, Length, and Data Type. Unfortunately, this metadata is useful primarily as a means of creating variable statements and validating that the resultant dataset conforms to the standard. Far more is required to drive automation of submission-ready artifacts.

The metadata required to support automation is complex, and has not been standardized. To differentiate it from the more familiar CDISC metadata, we introduce the term 'operational metadata' to mean the metadata necessary to support automation. There is a natural tradeoff in flexibility versus complexity. Simpler metadata will generally be more limited, as the code generating engine will need to make more assumptions on how to handle different cases.

Consider for a moment a TFL (Tables, Figures, and Listings) generating program that reads from three possible configurations of metadata (detailed in the table below). The first gives only the title for the table. Everything else must be hard coded. This simplifies the metadata, but also reduces the flexibility since any other changes require updates to the code. The second example shows some information stored as study metadata. This study metadata can be changed without touching the code, but must be the same for all tables in the study. The third example results in the most complex metadata and the most content defined per table. It is also the most flexible because the alignment, font, and size of the title are all configurable per table.

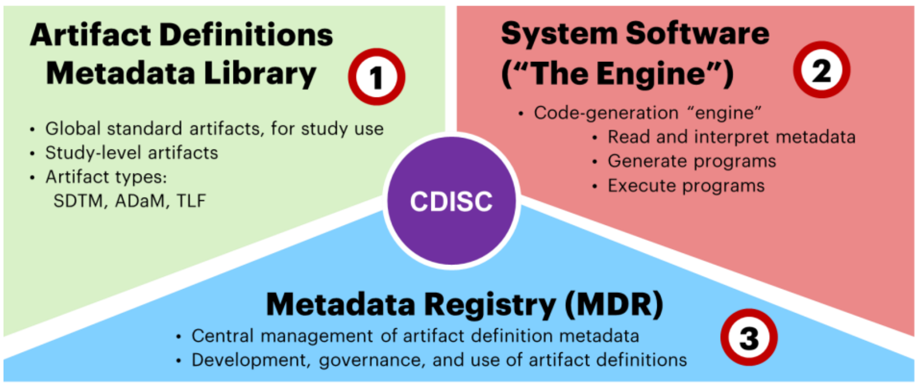
Example	Study Metadata	Table Metadata	Metadata Complexity	Flexibility
1	NA	**Title:** 'Summary Table for Vital Signs'	Simple	Low
2	**Alignment:** Left **Font:** Times New Roman **Text Size:** 12	**Title:** 'Summary Table for Vital Signs'	Medium	Medium
3	NA	**Title:** 'Summary Table for Vital Signs' **Alignment:** Left **Font:** Times New Roman **Text Size:** 12	Complex	High

These are trivial examples intended to convey the general principle: more flexibility comes at the cost of increased metadata complexity. When we consider the different ways that data can be manipulated to generate SDTM, ADaM and TFLs, the issue becomes much more pronounced. The principle used in development of our system was to minimize the complexity of the metadata as much as possible, while retaining the flexibility to support automation of SDTM, ADaM, and TFLs.

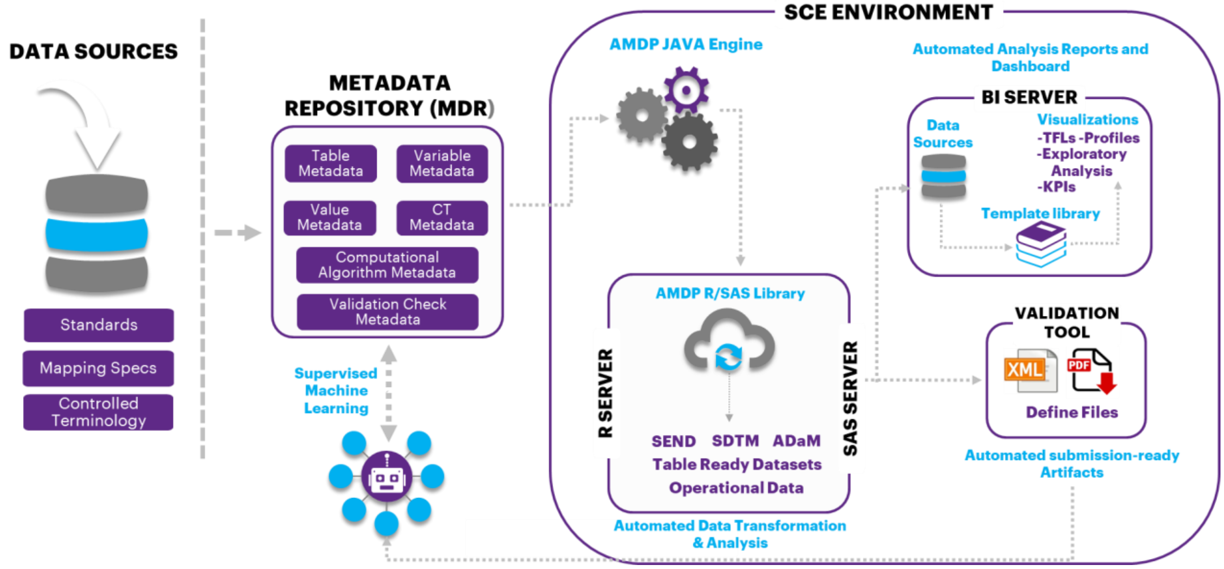
SYSTEM ARCHITECTURE

The AMDP system has three main components:

1. The metadata definitions for artifacts to be created
2. A metadata processing engine that both generates and executes SAS programs based on artifact definitions
3. A metadata registry system to store and manage the definitions



These components can exist in a variety of formats without changing the foundational architecture. The system version described in this paper uses a set of standard artifacts developed in collaboration with the Accenture Operations team. The engine is written in SAS and the application references metadata stored and maintained in an Excel-based MDR. The proposed future state re-imagines these components to promote the ability to efficiently scale the solution and is depicted in the diagram below. The future state will be discussed in more detail later in this paper.



SCOPE OF ARTIFACTS CREATED

The system can be used to generate and execute SAS programs that produce SDTM, ADaM and TFLs. AMDP has targeted comprehensive production of SDTM with little or no custom programming. The system has been validated to produce domains in each of the 3 general observation classes (Findings, Interventions, and Events), the Demographics Special Purpose Domain, and all corresponding SUPPQUAL Domains. Significant automation with ADaM is also possible. AMDP can generate the ADSL dataset, as well as datasets in the OCCDS and BDS structures. Finally, Tables and Listings that adhere to pre-defined standard templates can be produced. Currently, producing Figures is a limitation of the software.

The current capabilities of the system to produce artifacts in each of these categories are depicted in the table below. The capabilities as described are intentionally conservative in that they refer to demonstrated results during validation of the system.

Artifact Type	Template or Class	Artifact(s) Produced during Validation
SDTM	Findings	DM, DS, EX, EG, VS, LB, IE, PE,
	Interventions	CM,
	Events	AE, HO
	Special Purpose	DM
ADaM	ADSL	ADSL
	OCCDS	ADCM
	BDS	ADEG
TFL	Summary Table – Continuous and Categorical Variables Statistics	“Demography Summary Table”
	Summary table – Frequency Count and Percent by Single or Double Level	“Concomitant Medications [by CMDECOD]”
	Summary table – Frequency Count and Percent by Single or Double Level	“Concomitant Medications within Classes of Interest [CMCLAS*CMDECOD]”
	Change from Baseline Table	“Summary of ECG Parameters and Change from Baseline by Time by Treatment”
	Shift Table	“Summary of Post-Baseline Worst-Case ECG Parameters Shift from Baseline by Treatment”
	Basic Listing	“Listing of Concomitant Medications”

PILOT STUDY RESULTS

The AMDP system was used to convert a pilot study of data to SDTM. The client provided a proprietary global metadata library and two data cuts of source data for the study (56 & 59 datasets respectively). The Accenture team then developed study specifications (both human and machine-readable) to define the conversion of the source data into 48 SDTM domains. The entire effort took 31 business days. It is important to note that a single pilot study does not fully represent the value in SDTM automation. Subsequent studies would build upon the work done in the first study by leveraging re-usable content.

The breakdown of effort required to convert the study is represented in the table below. We expect that significant reductions in effort would be possible in most areas as a library of re-usable study specifications are developed and the team gains familiarity with the application.

Work Type	Reduction Expected?	Relative Effort (%)
Global Specification Interpretation	Yes	10%
Client-specific Team Training	Yes	5%
Human-readable Specification Development	Yes	15%
Machine-readable Specification Development	Yes	40%
AMDP Code Development	Yes	5%
QC and Compliance Reviews	No	15%
Client and Project Management	No	10%
Total		**100%**

The project team also evaluated the quality of the SDTM datasets produced using the system. In general, very high-fidelity datasets were produced with a small number of errors noted. Overall quality was 98.7% and was defined as 1 – (# errors / # variables).

Error Category	Error Count	Total Variables	Quality
Major	3	700	99.6%
Minor	6	700	99.1%
Total	**9**	**700**	**98.7%**

CONCLUSION AND FUTURE VISION

We have demonstrated the effectiveness of a metadata driven application in generating a wide variety of outputs including SDTM Datasets, ADaM Datasets, Tables, and Listings. The preliminary results are encouraging, and we intend to pursue development in three main areas:

1. **Code-generating Engine -** we are exploring the possibility of re-writing the engine in a general-purpose software development language such as JAVA to support the scalability and maintainability of the system. Additionally, we would like to extend support include support for R, so that either SAS or R programs can be generated by the system. This would reduce the reliance on SAS licenses thereby reduce the cost of the system.

2. **Metadata model -** during both validation and pilot studies, numerous enhancements to the metadata model have been proposed. These enhancements will simplify the metadata necessary to support generation of outputs, while retaining needed functionality.

3. **Specification development –** A user interface for specification development was outside the initial scope for AMDP. Going forwards, we understand the importance of an effective user interface for specification development to ensure maximum efficiency gains. In our pilot study, 55% of the effort was spent developing the machine readable and human readable study specifications. We intend to provide a streamlined user experience that guides the programmer through specifying the content for new artifacts while ensuring that the metadata created results in valid programs. We believe this will dramatically improve the efficiency gains in using this automated system.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of George Stuts and Anil Karunakaran in the creation of this paper. Both have provided important content, valuable guidance on the current system, as well as input into the future vision. We would also like to acknowledge the AMDP development and validation teams who made this system a reality.

CONTACT INFORMATION

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

Silas McKee
Accenture
1160 W. Swedesford Road
Southpoint One
Berwyn, PA 19312
Work Phone: 251-267-1743
Email: silas.a.mckee@accenture.com