

ALSC

**DH09
AUTOMATED CREATION OF
SUBMISSION-READY ARTIFACTS**

**SILAS MCKEE
NOVEMBER 2018**

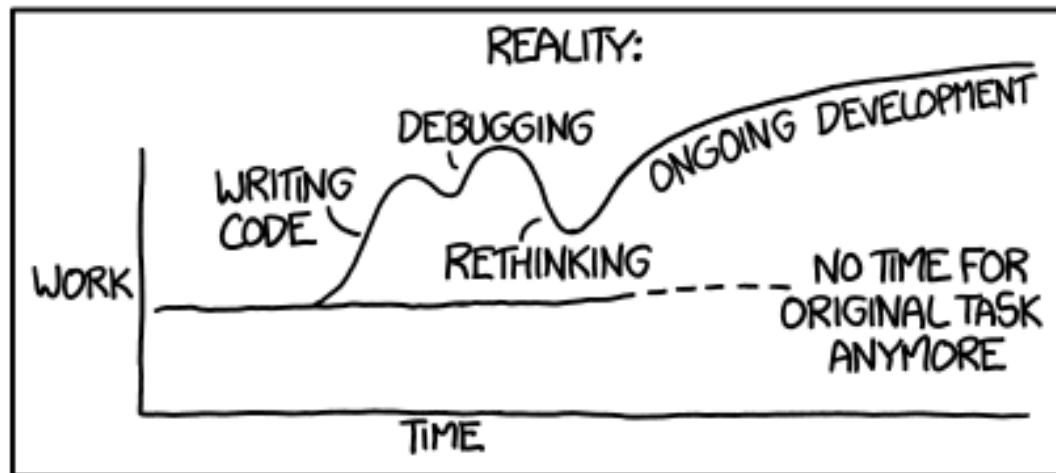
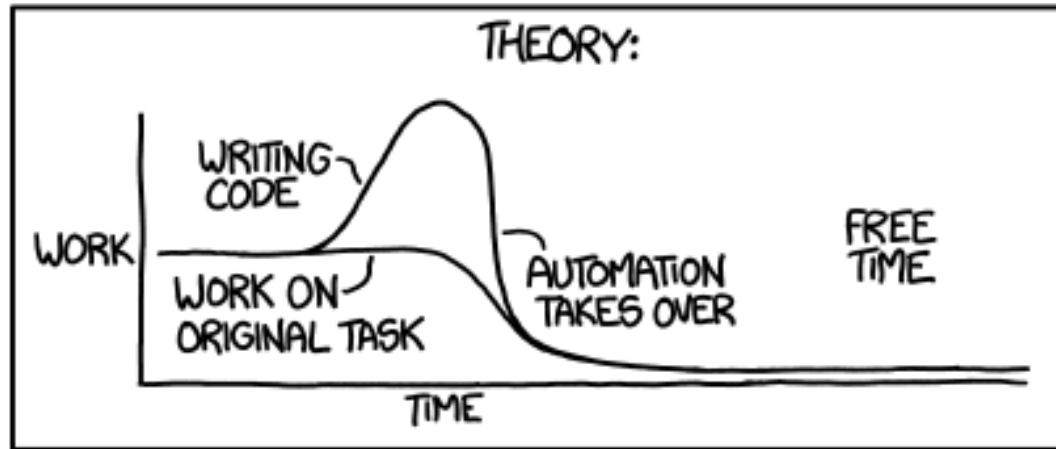


AGENDA

- 1. Motivation**
- 2. Automation Overview**
- 3. Architecture**
- 4. Validating the System**
- 5. Pilot Study Results**
- 6. Future State**

WHY AUTOMATE?

"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"



$$\frac{\text{Return (benefit)}}{\text{Investment (cost)}} = ROI \text{ (Return on Investment)}$$

- Reduce the per trial cost of submission artifact creation
- Reduce the cycle time for submission artifact creation
- Increase data quality

Source: <https://xkcd.com/1319/>

SUBMISSION ARTIFACT CREATION

CHALLENGES

1. Insufficient standardization of data sources and targets

- a. Significant progress has been made with the evolution of CDISC and sponsor standards to ensure a consistent approach to data collection and mapping to submission formats

2. Diverse data needs of clinical trials

- a. Science is evolving and new data must be collected to support the science, which results in a constant evolution of standards to support clinical trial data

3. Rigorous regulatory requirements for submission artifacts

- a. 'Good enough' is not enough - in many fields machine learning can reduce the effort to map and consolidate data, but the requirements for clinical trial data are more precise and exacting.

SUBMISSION ARTIFACT CREATION

MOVING FROM TRADITIONAL PROGRAMMING TO FULL AUTOMATION

No Automation

Full
Automation

What It Is 

Why Choose It 

Relative Efficiency 



Organic Programming Methodology

Traditional programming

Smallest initial investment

Most Flexible

- ✓ Low Efficiency
- ✓ Cycle times measured in Months



Custom Macro Library

- Traditional programming
- Standard macro library for code re-use

Supports junior level resources

Enables code re-use

- ✓ Moderate Efficiency
- ✓ Cycle times measured in Weeks



Automated Metadata Driven Processing (AMDP)

- Metadata-driven code generating engine

Repeatable outputs

Reduced cycle times

Reduced cost

- ✓ High Efficiency
- ✓ Cycle times measured in Days

SUBMISSION ARTIFACT CREATION

OUR APPROACH

1

Standardize data collection and target structures

As a pre-requisite, we ensure conformance of collected data to CDASH-based collection standards

2

Define precise mapping specifications

For each study, we develop machine-readable specifications that define source to target mappings, target metadata, and the sequencing programming steps

3

Develop a macro library

Our team created and continues to expand a library of SAS 'utility' macros that perform discrete tasks

4

Create a code-generating engine

Our team created a framework for generating SAS code based on the mapping specifications and the macro library

AUTOMATION WITH METADATA

CDISC METADATA

- Provides structural metadata to inform dataset structure and variable creation statements
- Ensures compliance with data standards

Vital Signs Dataset (VS)						
Variable	Label	Type	Controlled Terms or Format	Origin	Role	Comment
STUDYID	STUDY IDENTIFIER	text		CRF Page	Identifier	Demographics CRF Page 4
DOMAIN	DOMAIN ABBREVIATION	text		CRF Page	Identifier	DOMAIN ABBREVIATION
USUBJID	UNIQUE SUBJECT IDENTIFIER	text		CRF Page	Identifier	Demographics CRF Page 4
VSSEQ	SEQUENCE NUMBER	integer		CRF Page	Identifier	SEQUENCE NUMBER
VSTESTCD	VITAL SIGNS TEST SHORT NAME	text		CRF Page	Topic	Vital Signs CRF Page 4, CRF Page 7
VSTEST	VITAL SIGNS TEST NAME	text		CRF Page	Synonym Qualifier	VITALS SIGNS TEST NAME
VSPOS	VITAL SIGNS POSITION OF SUBJECT	text		CRF Page	Record Qualifier	POSITION OF THE SUBJECT DURING A MEASUREMENT OR EXAMINATION.

AUTOMATION WITH METADATA

'OPERATIONAL' METADATA

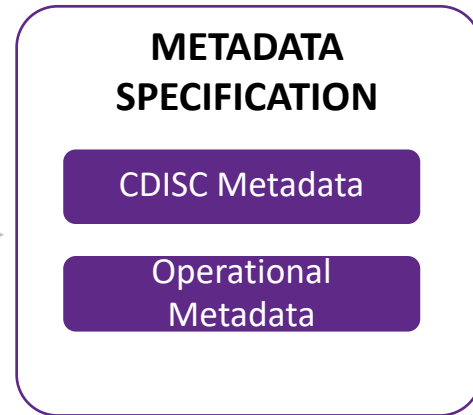
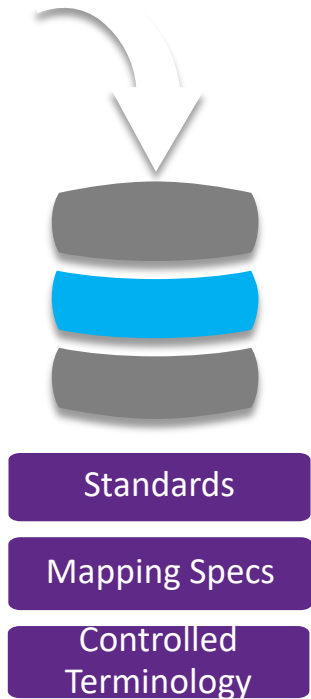
- Provides instructions for join / merge operations
- Provides instructions for macro calls including parameters values
- Provides instructions on the algorithms used to transform, derive, or compute data.

SOURCE_DATASET_NAME	SOURCE_VAR_NAME	RULE_TYPE	DERIVATION_TYPE	DERIVATION_ORDER	RULE_DEFINITION	TARGET_DATASET_NAME	TARGET_VAR_NAME
VS_INPUT_TMP2	VSPULSE_STD VSPULSE_RAW	S	SC	75		VS_PULSE	VSSTRESN
VS_INPUT_TMP2	VSSTRESN	S	SC	76		VS_PULSE	VSSTRESC
VS_INPUT_TMP2	VSPULSE_STD_UN	D	AS	77	'beats/min'	VS_PULSE	VSSTRESU
VS_INPUT_TMP2	VSPOS_PULSE_STD	S		78		VS_PULSE	VSPOS
VS_INPUT_TMP2	VSMETHOD_PULSE_STD	S		79		VS_PULSE	VSMETHOD
VS_INPUT_TMP2(WHERE=(VSHR	VSHR_RAW	D	AS	80	'HR'	VS_HR	VSTESTCD
VS_INPUT_TMP2	VSHR_RAW	D	AS	81	'Heart Rate'	VS_HR	VSTEST
VS_INPUT_TMP2	VSHR_RAW	S		82		VS_HR	VSORRES
VS_INPUT_TMP2	VSHR_UN	D	AS	83	'beats/min'	VS_HR	VSORRESU
VS_INPUT_TMP2	VSHR_STD VSHR_RAW	S	SC	84		VS_HR	VSSTRESN
VS_INPUT_TMP2	VSSTRESN	S	SC	85		VS_HR	VSSTRESC
VS_INPUT_TMP2	VSHR_STD_UN	D	AS	86	'beats/min'	VS_HR	VSSTRESU
VS_INPUT_TMP2	VSPOS_HR_STD	S		87		VS_HR	VSPOS
VS_INPUT_TMP2	VSMETHOD_HR_STD	S		88		VS_HR	VSMETHOD
VS_INPUT_TMP2(WHERE=(VSBS	VSBSA_RAW	D	AS	89	'BSA'	VS_BSA	VSTESTCD
VS_INPUT_TMP2	VSBSA_RAW	D	AS	90	'Body Surface Area'	VS_BSA	VSTEST
VS_INPUT_TMP2	VSBSA_RAW	S		91		VS_BSA	VSORRES
VS_INPUT_TMP2	VSBSA UN	S		92		VS_BSA	VSORRESU

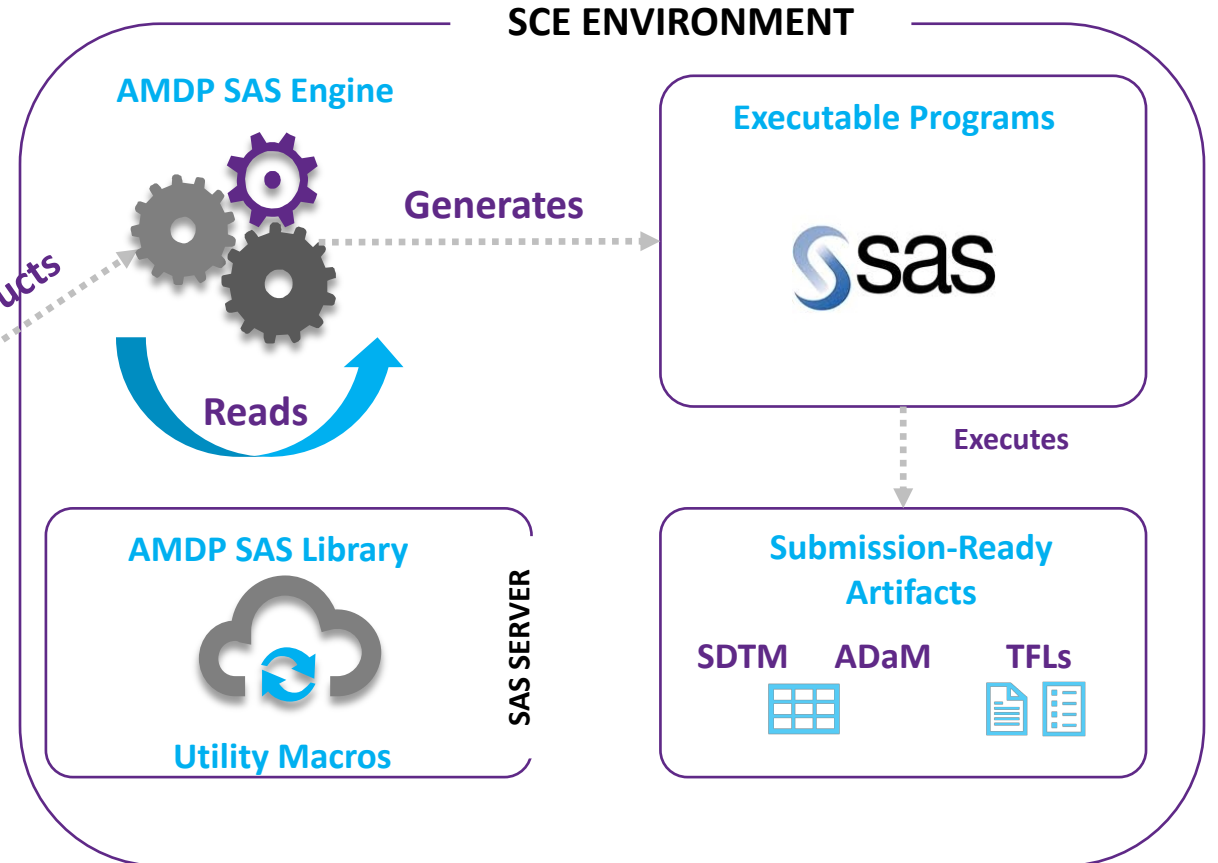
AMDP ARCHITECTURE

Version 1.0

METADATA SOURCES



Instructs



VALIDATING THE SYSTEM

The Artifacts

1. Created a set of standard artifacts that included representation from the following categories across SDTM, ADaM, and TFLs

SDTM

General Observation Classes

- Findings
- Interventions
- Events

Special Purpose

Special Purpose Relationship

ADaM

ADaM Subject-Level Analysis Dataset (ADSL)

Occurrence Data (OCCDS)

Basic Data Structure (BDS)

TFL

Basic Listings

Summary Tables

Change from Baseline Tables

Shift Tables

2. Validated the generation of these standard artifacts against data from two different studies

PILOT STUDY RESULTS

Description

- The team received a proprietary global metadata library and source data for the study
- The team then developed study specifications (both human and machine-readable) to define the conversion of the source data into 48 SDTM domains.
- Successfully completed SDTM conversion with high quality in 31 business days.

Quality Metrics

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

$$Quality = 1 - \left(\frac{Error\ Count}{Total\ Variables} \right)$$

PILOT STUDY RESULTS

Work Breakdown

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%

Takeaways

- Time to perform most tasks expected to decrease with study volume
- Very little Code Development
- 55% of effort devoted to specification development

WHAT'S NEXT?

1

Specification Development

Streamline the user experience for creating metadata specifications by creating an intuitive user interface

2

Metadata Model

Simplify the metadata needed to support generation of outputs, while retaining needed functionality

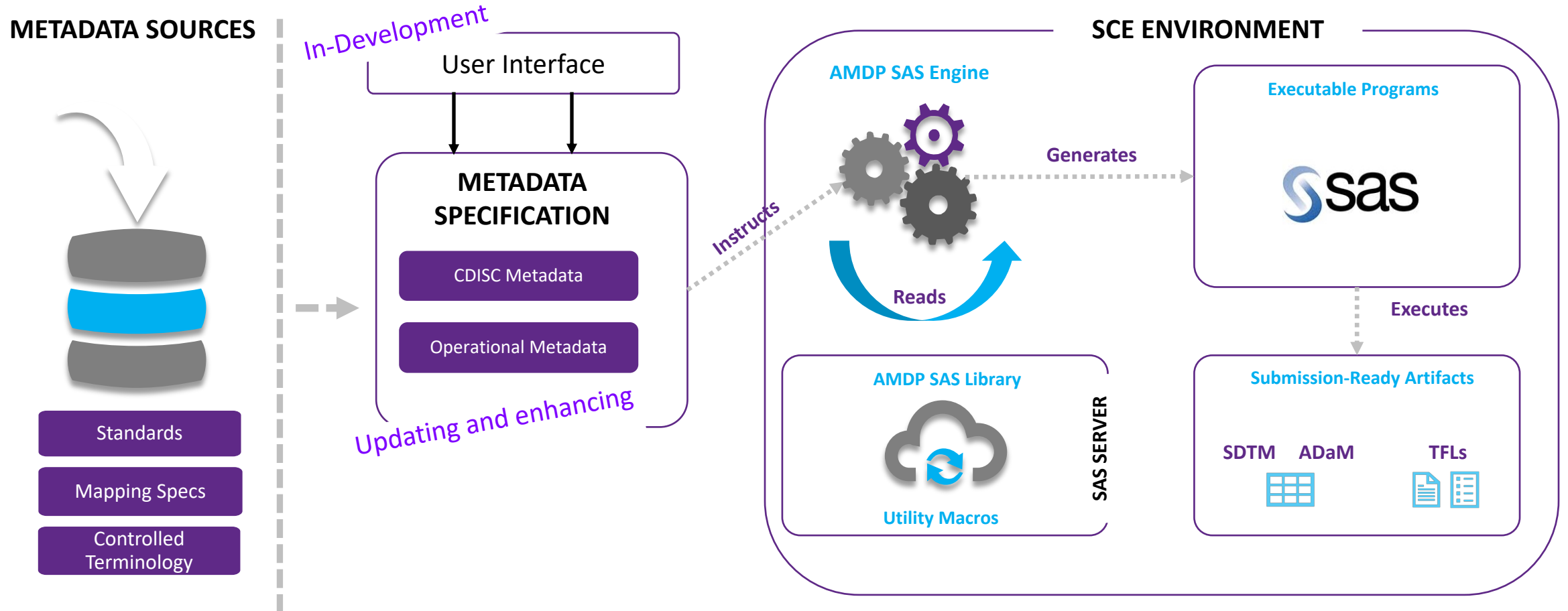
3

Code Generating Engine

Re-write the engine in a general-purpose language such as JAVA to support the scalability and maintainability of the system

AMDP ARCHITECTURE

Current State



Questions?

Silas McKee
silas.a.mckee@accenture.com