DRUG INFORMATION ASSOCIATION

44TH Annual Meeting

Boston 2008

Metadata-Driven Technology for Implementing CDISC SDTM

Michael Todd

President

Nth Analytics

Quality to the Nth Degree

Nth Analytics

CDISC Implementation Goals

- Design a strategy such that:
  - No knowledge needed of system that originally produced the legacy data
  - Applicable to files from any system
  - Implementation is flexible enough to adapt to different study designs
  - Minimal programming support required for maintenance
  - Reasonable cost

Implementing an ETL Process

- Programs read table-driven metadata to translate the analysis data into SDTM formats
  - Tells the SAS code which analysis variables populate the SDTM variables
  - Indicates when specialized code is required
- All code is developed to be generic using the metadata to indicate when variations are required
- New studies only require changes to metadata

Commercial ETL Product

- Commercial products
  - Obvious advantages:
    - Already built!
    - Validation is simplified
    - Vendor training available
  - Downside
    - Expensive!
  - Products
    - SAS Data Integration (DI) Studio
    - Oracle Warehouse Builder (OWB)

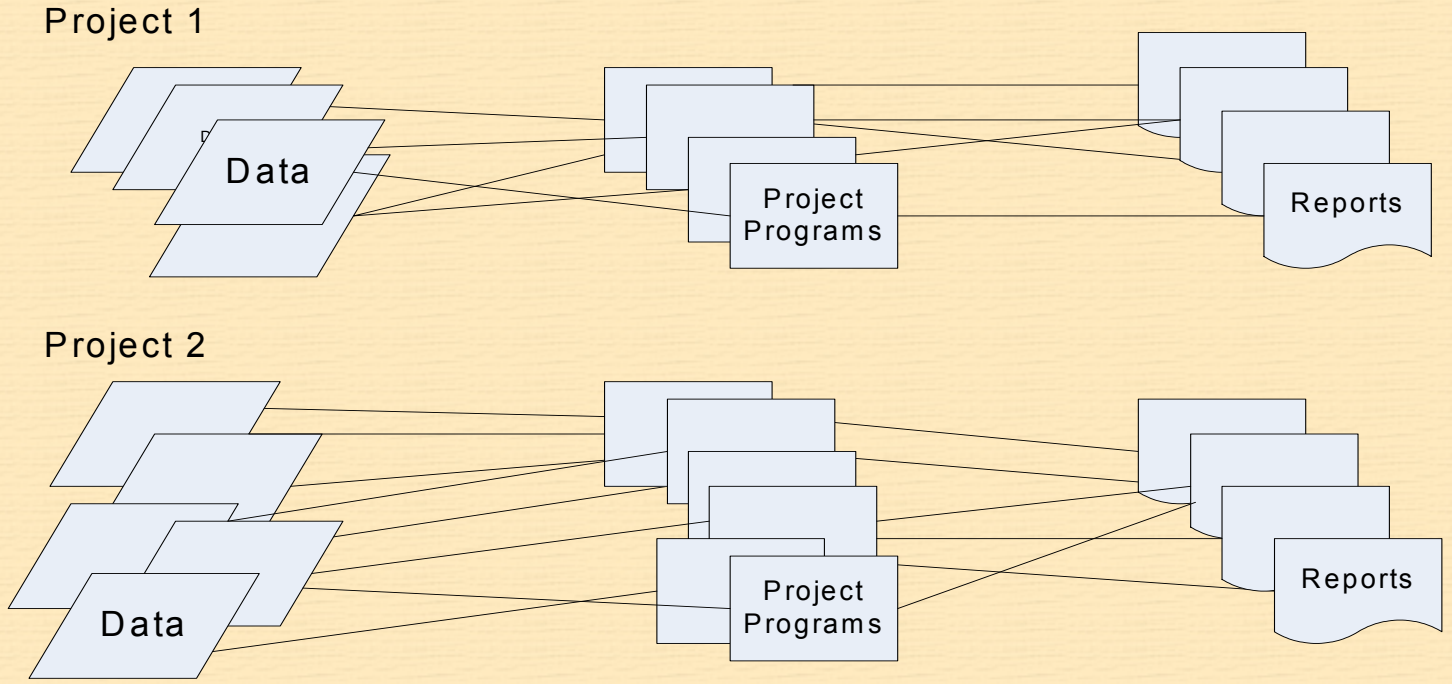
In-House ETL Product

- While in-house products have obvious downsides, they represent the only feasible solution for smaller companies, due to cost of commercial products
- Advantages:
  - Incremental development
  - Cost
  - Achieves same result as commercial products
- Disadvantages
  - Requires SDLC validation
  - Usual in-house development problems

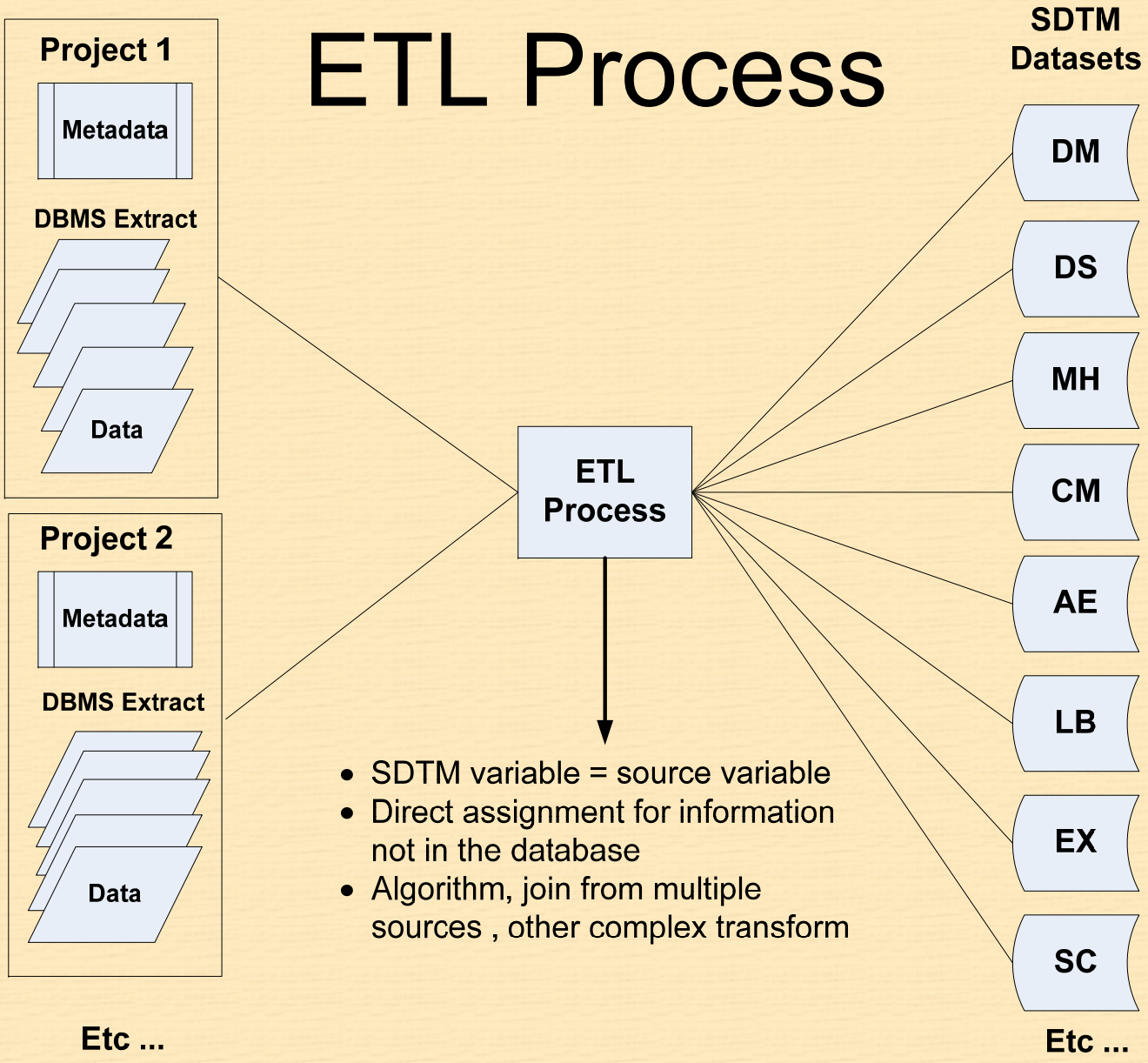
ETL Transformation Process

- Define how raw/analysis data fits into SDTM domains and variables
- Match data to required, permitted and expected SDTM data when possible
- Provide an automated mechanism for specifying the data sources and algorithms
  - Metadata for the SDTM files
  - Basis for the FDA-mandated “DEFINE.XML” documentation

Process Without Automation



ETL Process



- SDTM variable = source variable
- Direct assignment for information not in the database
- Algorithm, join from multiple sources , other complex transform

Why it Works

- Role of standards
  - Standards drive the process. Target has standard structure so it makes standardizing tools economically worthwhile.
  - While source variables differ, there are certain commonalities that can be exploited
- Knowledge required
  - CDISC Standards
  - Understanding of raw data issues
  - Study design
  - Limited derivation

Elements

- Data
- People
- Dataflow
- Software
- Validation

The Data

Source

- ADVEVNT
- DEMOG
- MEDHIST
- PHYSEXAM
- INCLEXCL
- PROTDEV
- LABS
- VITSIGN
- CMED
- STUDYMED

… etc.

- Lab Normals
- Protocol-specific dictionary
- Abnormality Criteria
- Protocol
- Annotated CRF
- Statistical Analysis Plan

Target

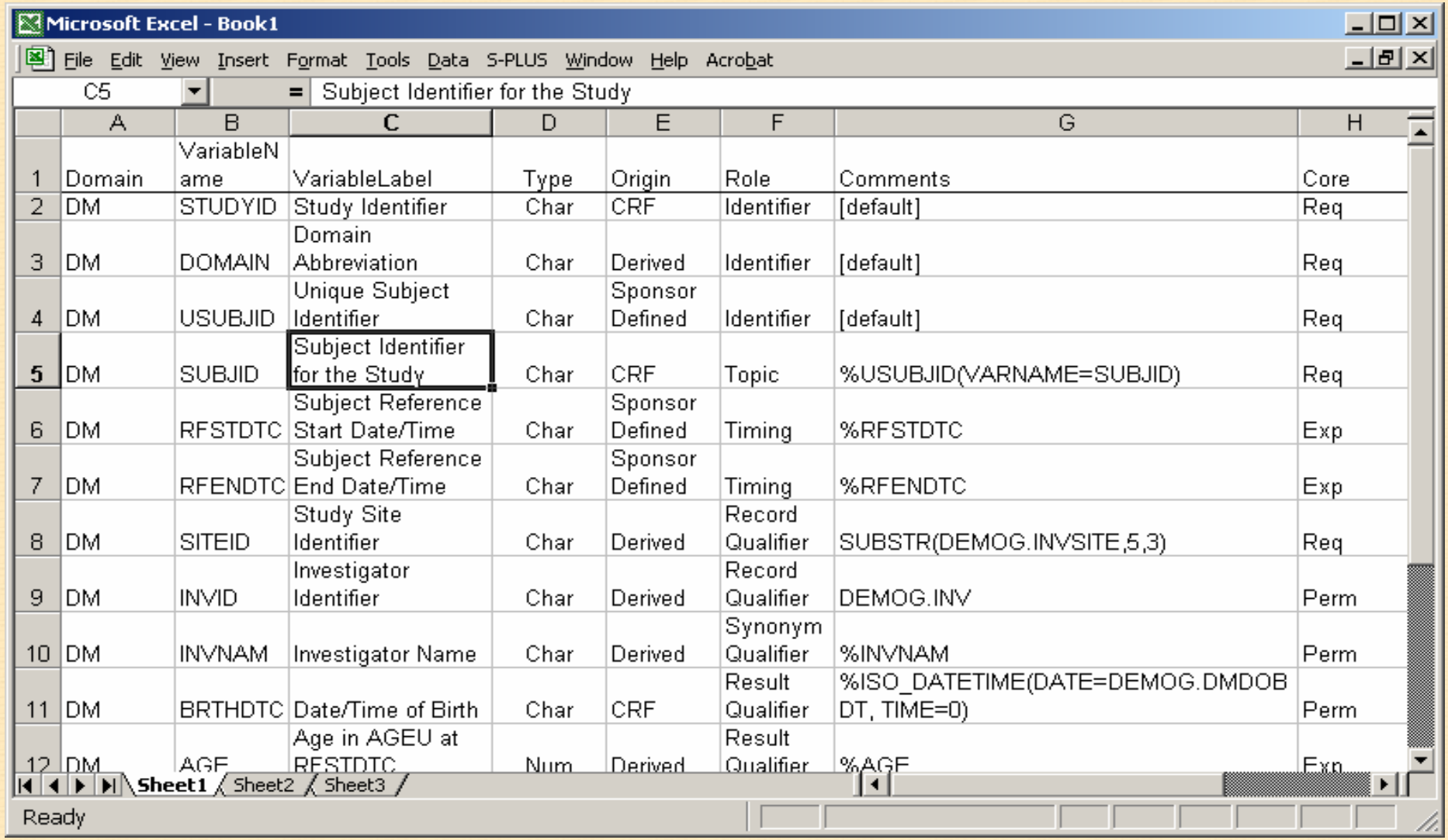
- DEFINE.XML
- Trial Design Datasets TE, TA, TV, TI, TS
- Subject Datasets SV, SE
- AE
- SUPPAE
- CM
- SUPPCM
- DM
- DS
- EX
- SUPPEX

… etc.

People: Tasks and Job Roles

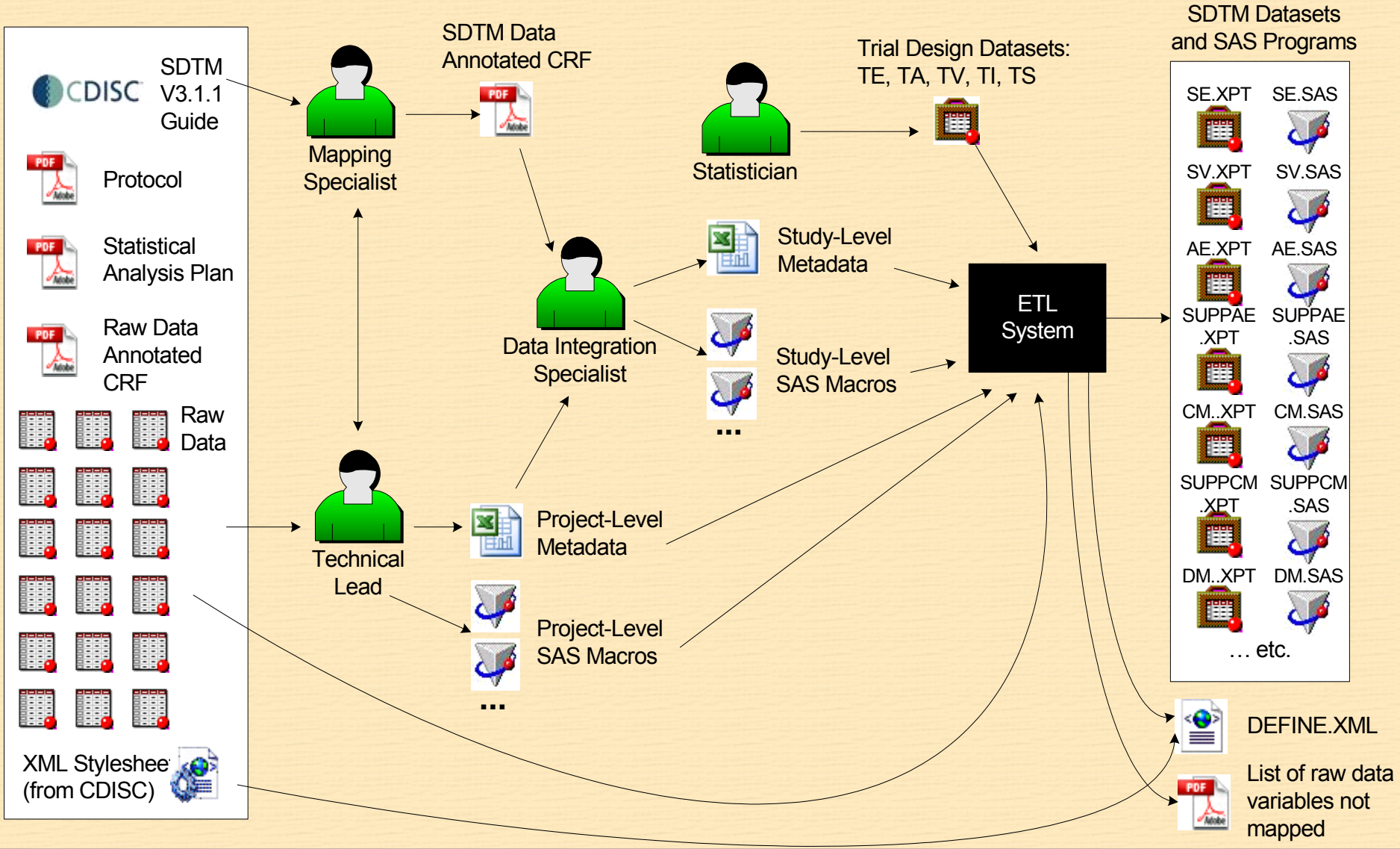
Task	Job Role	Requires
Development of the Annotated CRF	Mapping Specialist	Knowledge of SDTM V3.1.1 Implementation Guide Data Management expertise
Development of mapping specifications		
Create trial design datasets	Statistician	Ability to translate abstract concepts into datasets
Development of conversion jobs in ETL Environment	Data Integration Specialist	SAS programming Knowledge of ETL tool
QC of the SDTM files	QC Specialist	All of the above

Sample SDTM Metadata

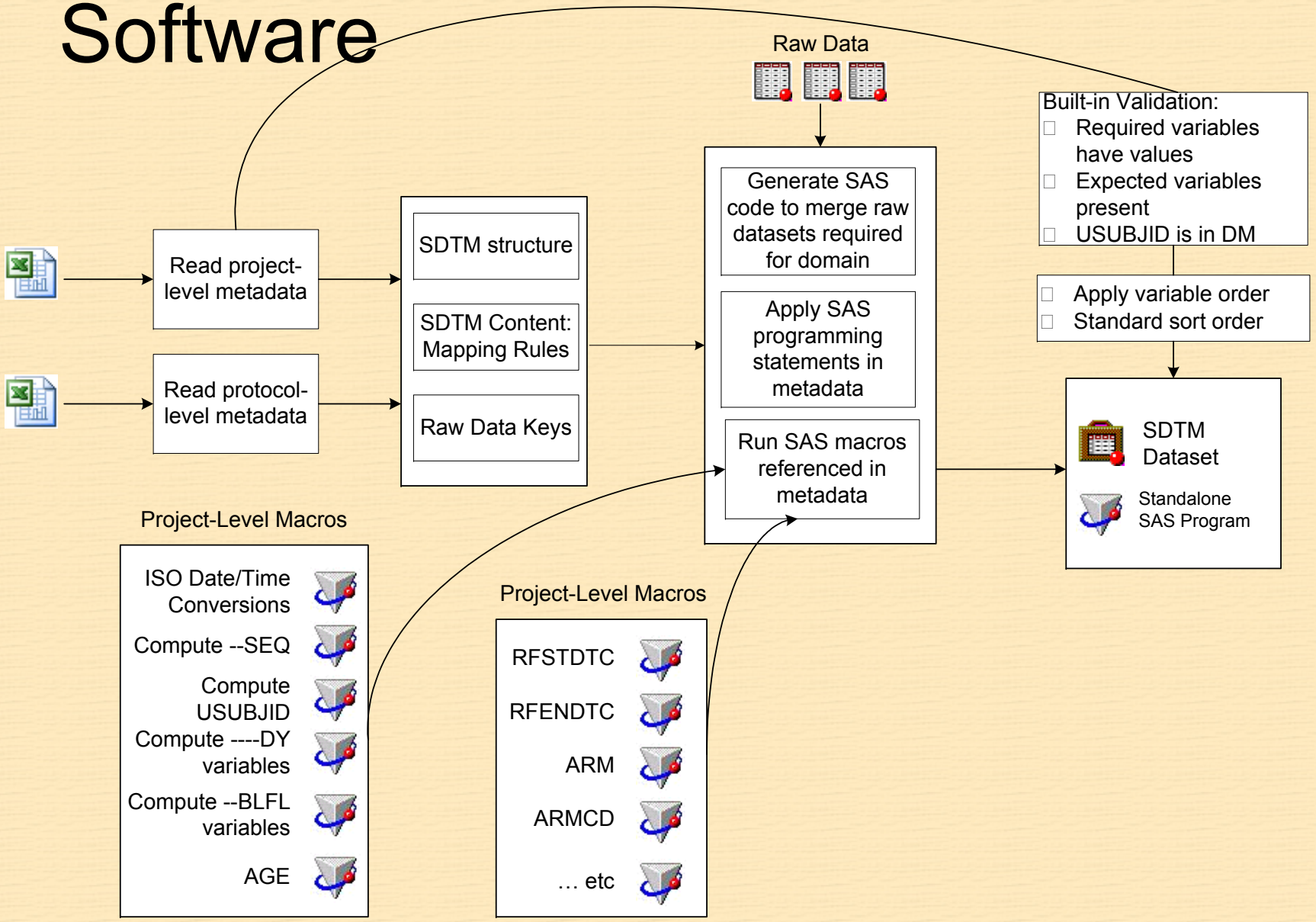


	A	B	C	D	E	F	G	H
1	Domain	VariableName	VariableLabel	Type	Origin	Role	Comments	Core
2	DM	STUDYID	Study Identifier	Char	CRF	Identifier	[default]	Req
3	DM	DOMAIN	Domain Abbreviation	Char	Derived	Identifier	[default]	Req
4	DM	USUBJID	Unique Subject Identifier	Char	Sponsor Defined	Identifier	[default]	Req
5	DM	SUBJID	Subject Identifier for the Study	Char	CRF	Topic	%USUBJID(VARNAME=SUBJID)	Req
6	DM	RFSTDTC	Subject Reference Start Date/Time	Char	Sponsor Defined	Timing	%RFSTDTC	Exp
7	DM	RFENDTC	Subject Reference End Date/Time	Char	Sponsor Defined	Timing	%RFENDTC	Exp
8	DM	SITEID	Study Site Identifier	Char	Derived	Record Qualifier	SUBSTR(DEMOG.INVSITE,5,3)	Req
9	DM	INVID	Investigator Identifier	Char	Derived	Record Qualifier	DEMOG.INV	Perm
10	DM	INVNAM	Investigator Name	Char	Derived	Synonym Qualifier	%INVNAM	Perm
11	DM	BRTHDTC	Date/Time of Birth	Char	CRF	Result Qualifier	%ISO_DATETIME(DATE=DEMOG.DMDOBDT, TIME=0)	Perm
12	DM	AGE	Age in AGEU at RFSTDTC	Num	Derived	Result Qualifier	%AGE	Exp

Dataflow


SDTM V3.1.1 Guide
Protocol
Statistical Analysis Plan
Raw Data Annotated CRF
Raw Data
XML Stylesheet (from CDISC)
Mapping Specialist
SDTM Data Annotated CRF
Technical Lead
Data Integration Specialist
Statistician
Trial Design Datasets: TE, TA, TV, TI, TS
Study-Level Metadata
Study-Level SAS Macros
Project-Level Metadata
Project-Level SAS Macros
ETL System
SDTM Datasets and SAS Programs
SE.XPT
SE.SAS
SV.XPT
SV.SAS
AE.XPT
AE.SAS
SUPPAE .XPT
SUPPAE .SAS
CM..XPT
CM.SAS
SUPPCM .XPT
SUPPCM .SAS
DM..XPT
DM.SAS
… etc.
DEFINE.XML
List of raw data variables not mapped

Software


Raw Data
Built-in Validation:
Required variables have values
Expected variables present
USUBJID is in DM
Apply variable order
Standard sort order
Read project-level metadata
Read protocol-level metadata
SDTM structure
SDTM Content: Mapping Rules
Raw Data Keys
Generate SAS code to merge raw datasets required for domain
Apply SAS programming statements in metadata
Run SAS macros referenced in metadata
SDTM Dataset
Standalone SAS Program
Project-Level Macros
ISO Date/Time Conversions
Compute --SEQ
Compute USUBJID
Compute ----DY variables
Compute --BLFL variables
AGE
Project-Level Macros
RFSTDTC
RFENDTC
ARM
ARMCD
… etc

Sources of Error

Given a validated system, there are still several sources of error in the process:

1. CRF SDTM Annotations
2. Trial design datasets
3. Metadata
4. SAS macros

CRF SDTM Annotations

- Source of error:
  - CRF annotations are not automated
  - Dependent on expert knowledge of SDTM Implementation Guide
  - System cannot use metadata to write annotations on the CRF
- Solution requires:
  - Better knowledge of XML

Trial Design Datasets

- Source of error:
  - Manual process
  - Depends on detail-oriented statistician with ability to translate study design into abstract concepts
  - System cannot read protocol to generate the trial design datasets
- Solution requires:
  - "Study Designer" tool to enable clinicians or statisticians to generate the trial design datasets
  - XML and machine-readable protocol

Metadata

- Source of error:
  - No automatic link between annotated CRF and metadata
  - System cannot read annotations from CRF to write metadata
  - Contains SAS programming statements
- Solution requires:
  - XML would enable system to read annotated CRF, but derivations would still be a source of error

SAS Macros

- Source of error due to:
  - Traditional SAS programming approach to handle complex derivations
- Solution requires:
  - Error rate can be minimized, but not eliminated, through good programming practices
  - Use of Stored Processes (black boxes) without access to source code

ETL System: Advantages

- Does not disrupt existing clinical trial systems
- It works for all legacy data
- Reduces cost since only metadata changes for each new study
- Only new study specific situations cause additional coding
- Changes to CDISC standards are made to the metadata avoiding costly programming revisions

Advantages (continued)

- Self-documenting
- Metadata easily converted to DEFINE.XML
  - If you follow the system, metadata are guaranteed to provide complete and accurate documentation
- Generates submission-ready SAS programs
  - System macros create standalone code
  - Code looks good because it is machine-written
  - Use PROC COMPARE to verify that standalone programs accuracy create the SDTM datasets

Summary

- Metadata is the key for successful automation
- Off-the-shelf tools can provide a powerful ETL solution