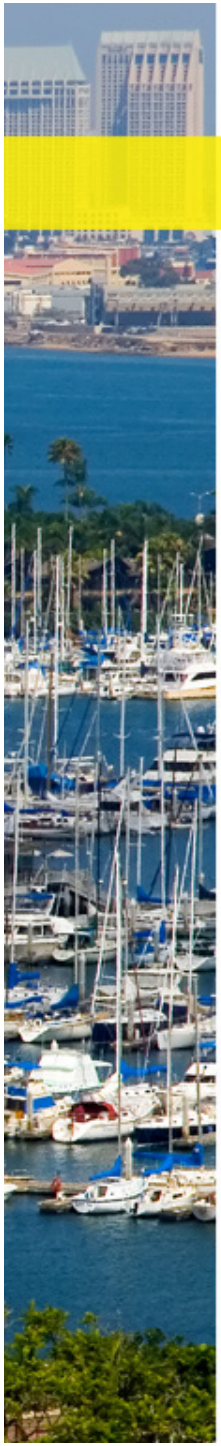


San Diego 2009

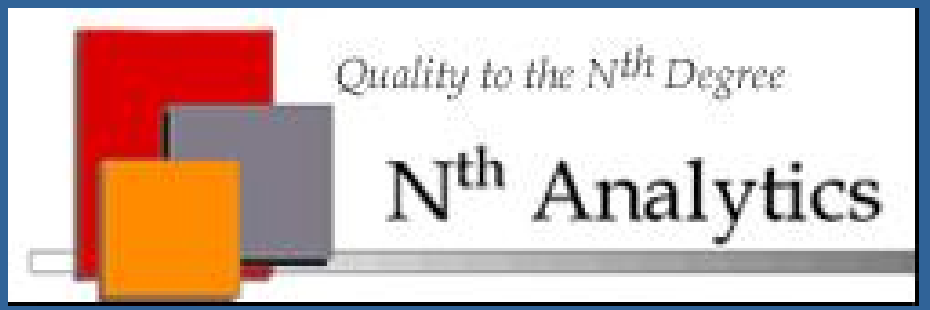
45th Annual Meeting

DIA
DRUG INFORMATION ASSOCIATION

SDTM Mapping: Current Technology and Expert Systems

Michael Todd and Thomas Jablonski

Nth Analytics



Disclaimer

- The views and opinions expressed in the following PowerPoint slides are those of the individual presenter and should not be attributed to Drug Information Association, Inc. ("DIA"), its directors, officers, employees, volunteers, members, chapters, councils, Special Interest Area Communities or affiliates, or any organization with which the presenter is employed or affiliated.

-

Introduction

- Current SDTM mapping methodology is well-established but limited
  - Many companies use some version of metadata-driven ETL mapping system
  - However, it requires a mapping expert to define the metadata
    - the number of experts is limited
- We need a fully automated expert system to convert clinical trial data to SDTM on a massive scale

Current Technology

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

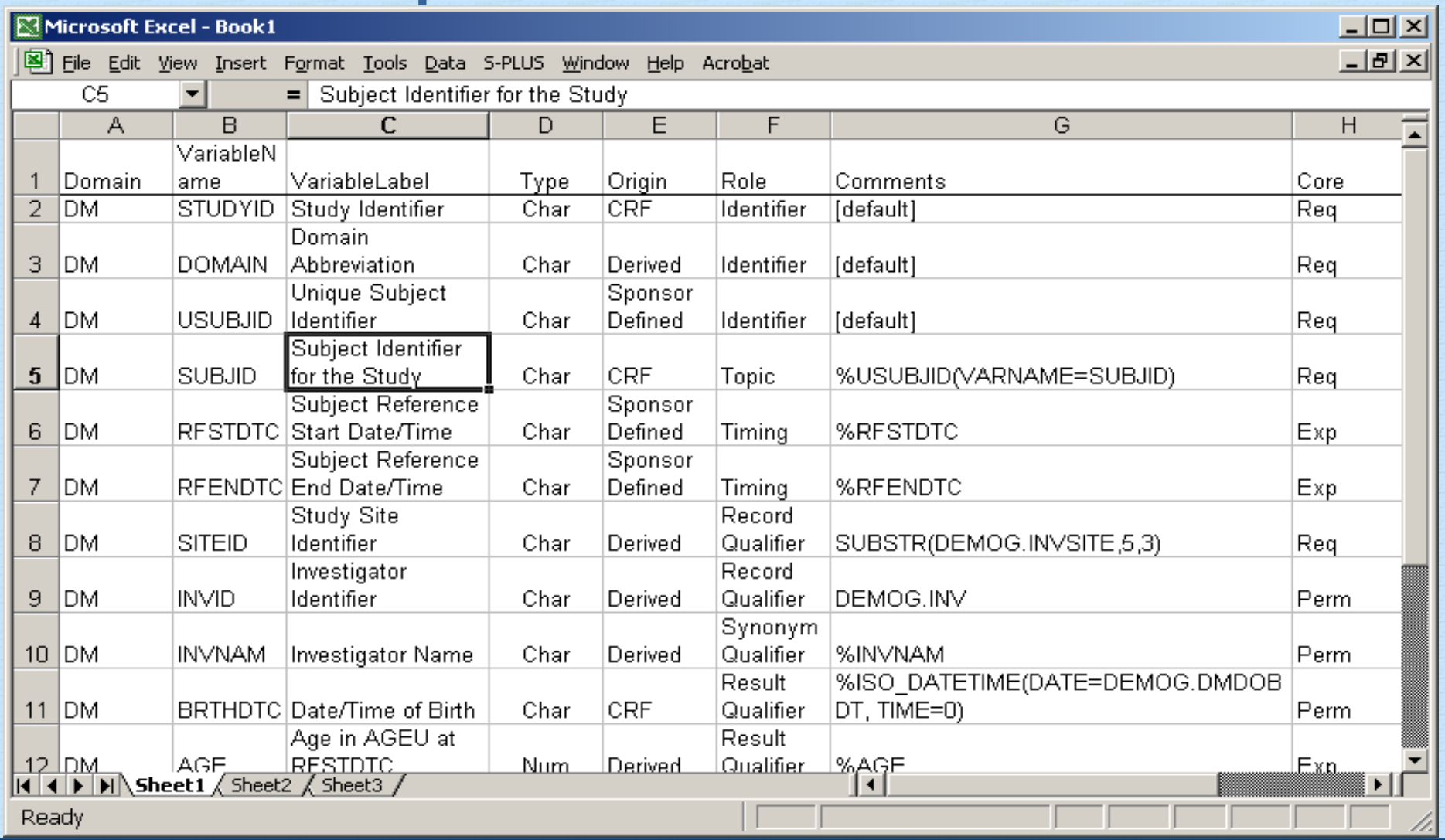
Implementing an ETL Process

- 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

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

Sample SDTM Metadata


Microsoft Excel - Book1

File Edit View Insert Format Tools Data S-PLUS Window Help Acrobat

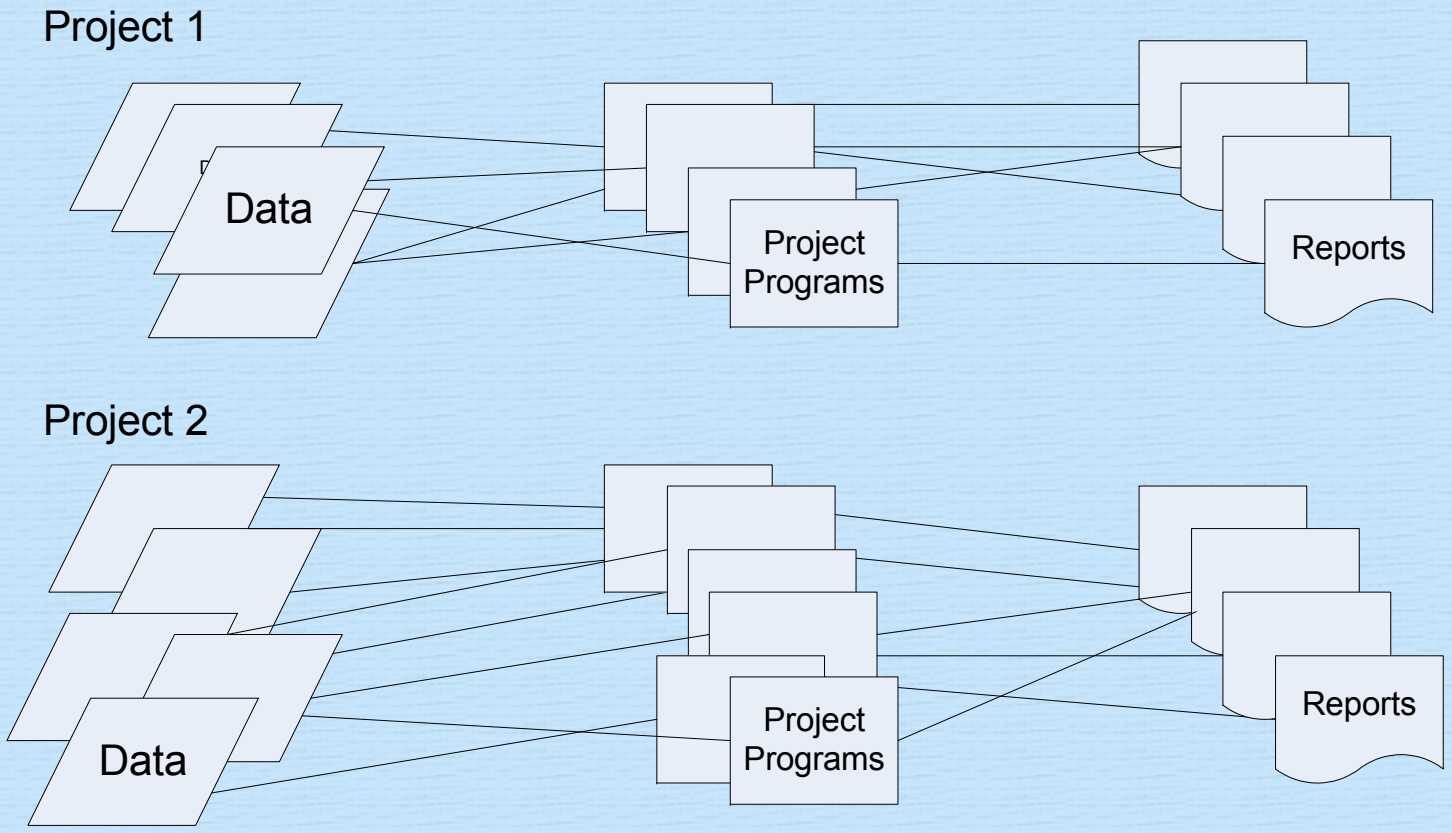
C5 = Subject Identifier for the Study

	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

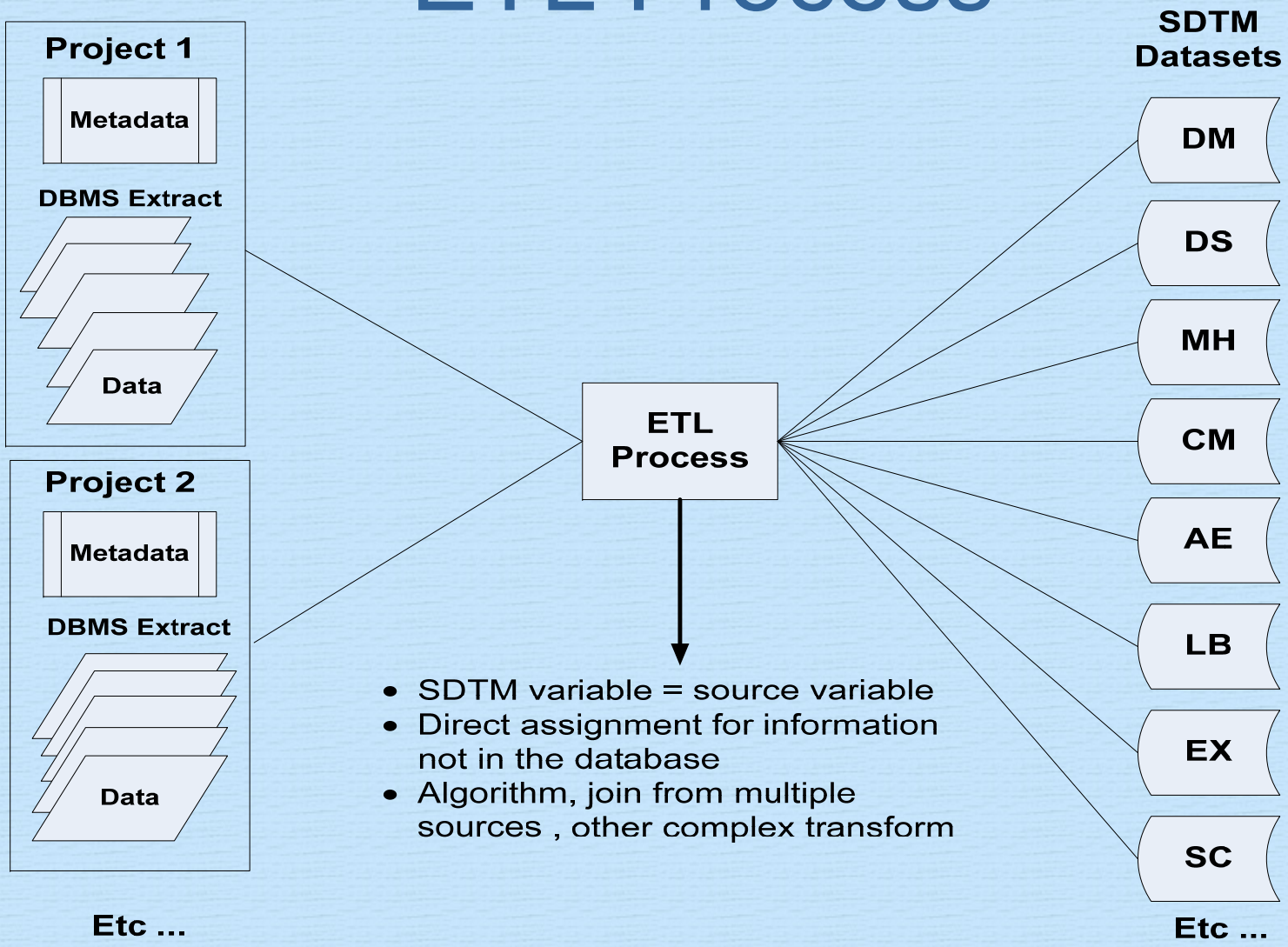
Sheet1 Sheet2 Sheet3

Ready

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

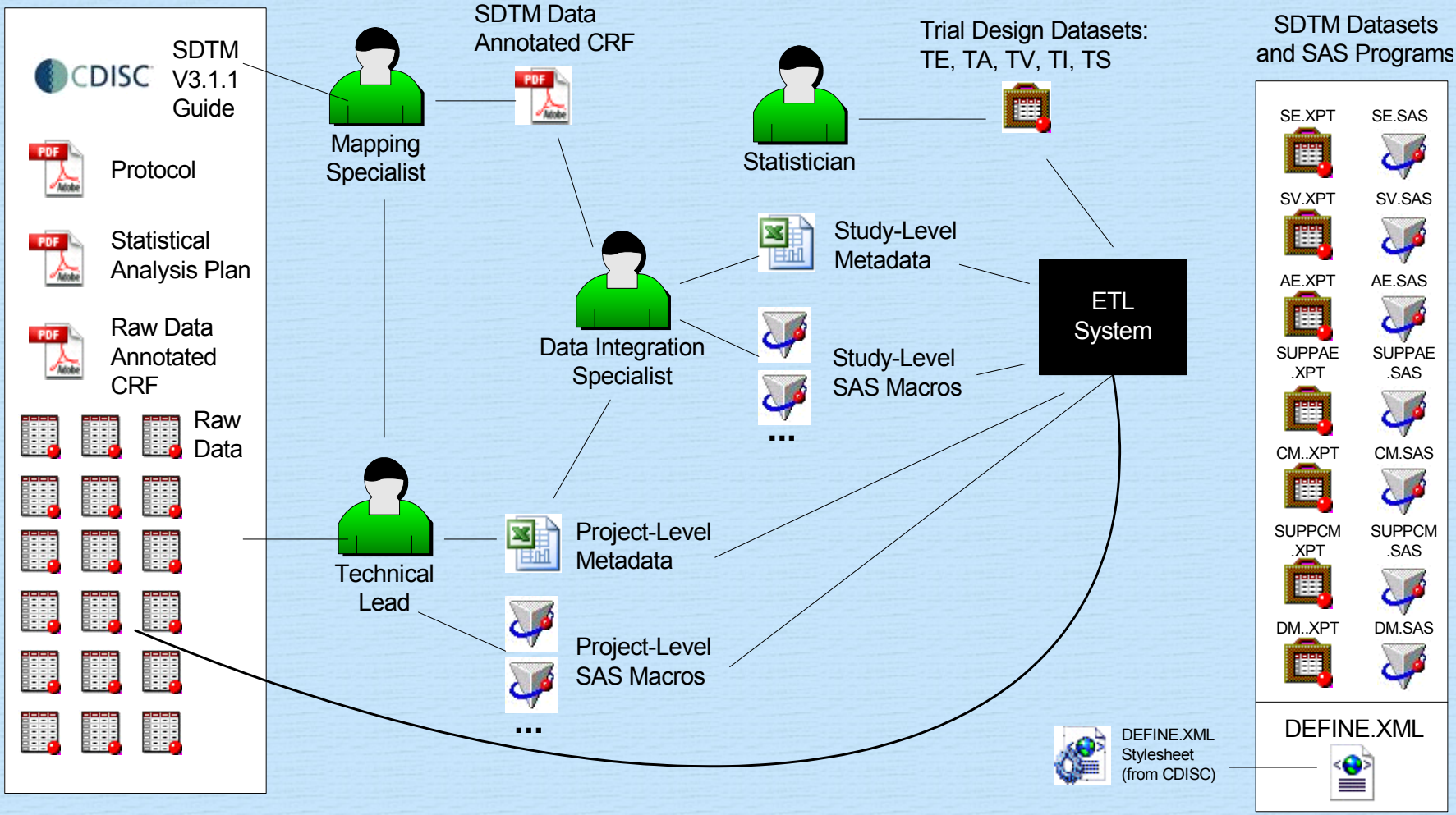
Why it Works

- Role of standards
  - Standards drive the process. Target has standard structure so can be standardized.
  - While source variables differ, commonalities can be exploited
- Knowledge required
  - CDISC Standards
  - Understanding of raw data issues
  - Study design
  - Limited derivation

Tasks and Job Roles

Task	Job Role	Requires
Annotated CRF	Mapping Specialist	•In-depth knowledge of SDTM V3.1.1 IG •Expert knowledge of clinical data
SDTM 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	•In-depth knowledge of SDTM V3.1.1 IG •Expert knowledge of clinical data
System maintainance Project-level metadata/macros	Technical Lead	•In-depth knowledge of system, SDTM, and clinical data

Dataflow


CDISC
SDTM V3.1.1 Guide
Protocol
Statistical Analysis Plan
Raw Data Annotated CRF
Raw Data
Mapping Specialist
SDTM Data Annotated CRF
Statistician
Trial Design Datasets: TE, TA, TV, TI, TS
Data Integration Specialist
Study-Level Metadata
Study-Level SAS Macros
...
ETL System
Technical Lead
Project-Level Metadata
Project-Level SAS Macros
...
DEFINE.XML Stylesheet (from CDISC)
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
DEFINE.XML

Limitations

- Requires experts
- Severely limited throughput, relative to amount of clinical trial data
- Converting legacy data on a systemic scale is infeasible

Future Directions

Requirements Going Forward

- Without legacy data, goals of meta analyses, etc. will be limited and incomplete
- In order to effectively use SDTM, the FDA warehouse must include all data for a compound, not just new data going forward
- Converting legacy on this scale is simply infeasible with current techniques.

Challenge

- Convert unstructured information such as text into relational tables that can be used to generate code to create SDTM & DEFINE.XML
- To create this system, imagine thinking like a computer.
  - You have sources of information
  - You have a set of rules
  - You have a storage of knowledge available.
  - Apply heuristics to create SDTM datasets with a certain probability of accuracy.

Sources of Information

- Data
  - Main source of information
  - Can assume data exists, while protocol & CRF may not for legacy studies.
- CRF
  - Usually this is an image, can it be processed?
- Protocol and Study Report
  - Possibly use text-mining techniques to extract information to help organize data

Why SDTM is Amenable to an Expert-System Approach

- SDTM represents a well-defined, rule-based structure
  - Expert system assumptions rely on well-defined structure and meaning
  - Assume that data have organization and meaning, however hard to determine
- Certain things make it easier
  - Assume characteristics for clinical data, as opposed to exponentially more possibilities for any arbitrary data
  - Limited set of target SDTM domains

Role of the SDTM Expert

- Recognize the kinds of data coming in
- Redistribute data to SDTM
  - Have to recognize the type of data without being told what it is
  - How do you know something is lab data, if you can't rely on variable names and labels?
  - How would a machine recognize the type of data just by the structure and values?

Role of the Expert (continued)

- Expert can recognize tests even if the data are not labeled
  - Experts can differentiate the data
  - Under a well-defined set of rules, data has logical and mathematical place to be.
- System must handle tests may not exist today, but would still fit into findings, events, or interventions.

Examples of Reliable Assumptions

- Data are in English
- Each dataset contains the same type of data
  - AEs and conmeds are not in the same dataset
- Each dataset contains keys: variable(s) that enable datasets to be joined together
- Dates and times have a sequence
  - Discoverable by sorting

How to Think Like an Expert (Machine)

Identifying Dataset HMZ11

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

- What is this?
- How do we know?
  - No obvious visit or timing variable, other than C3
  - C4 and C5 may be controlled terminology
  - C1 and C2 look like keys

Identifying C1

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

- Left-most column often is a protocol
- Mixture of letters, numbers, and special characters: probably a code
- No hits for dictionary lookup for meaningful terms
- If the sponsor is known, there may be a list of protocols for lookup

Identifying C2

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

- The same things we noted for C1 also apply to C2.
- 'FAB-10' is as likely to be a protocol number as 'X312.
- It is only because C1 has the same value for all records in the dataset that we can conclude with a high probability that 'FAB-10' is a protocol number.

Protocol and Subject Numbers

C1
FAB-10

C2
X3121
X3122
X3123
X3124
X3125
X3126
X3127
X3128
X3129
X3130
X3131
Y2221
Y2222
Y2223
Y2224

- Assume if a possible protocol number has only one value in a dataset, it very likely is a protocol number.
- If there are two values (FAB-10, FAB-11), possibly the dataset contains results from two protocols.
- If it contains only one value in multiple datasets, this boosts our confidence in it being a protocol number

- If the list of subject numbers is consistent across datasets, we can assume with more certainty they are subject numbers
- We assume datasets contain the same subjects, for the most part

Identifying C3

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

Is it a sequence number?

- C3 contains only integers
- Increasing series from 1 to n with some gaps and some ties
- Most subjects have the same number of records
- Implies series of checkboxes on CRF, preprinted choices
- If we select distinct C2, C3, there should only be one record for each combination..

Identifying C4

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

Is this a list of body systems?

- Terms would match known systems in a dictionary lookup for body systems
- We would expect most subjects to have the same terms
- Should be controlled terminology
- Usually corresponds to the sequence number (C3), although not always

Identifying C5

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

If C4 is the body system, is C5 the status?

- We assume that if there is a body system, there should be a result for that system
- Terms appear to be a finite set, implying controlled terminology
- "HISTORY / NOT ACTIVE" and "CURRENTLY ACTIVE" suggest medical history

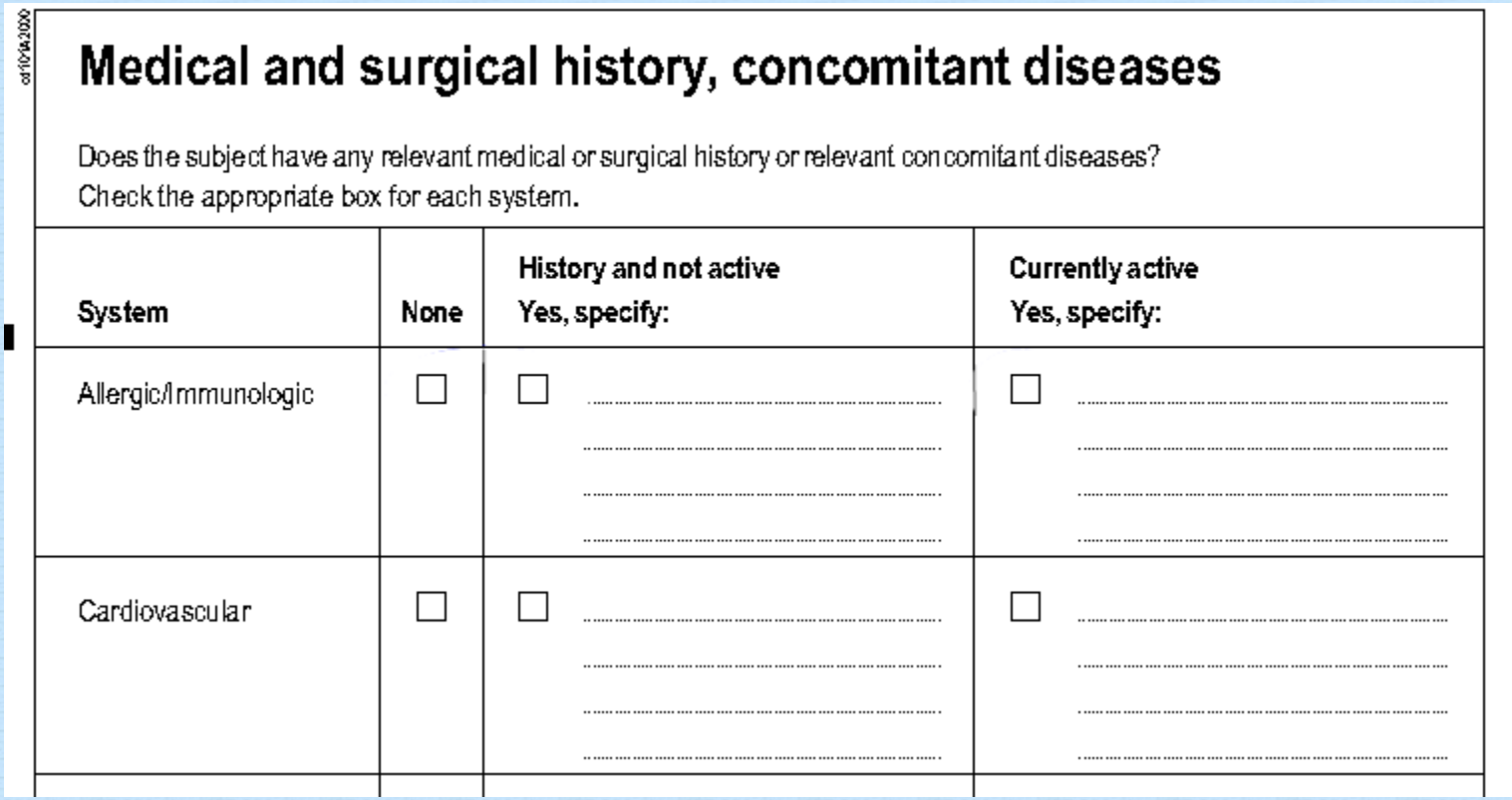
Identifying C6 and C7

C1	C2	C3	C4	C5	C6	C7
FAB-10	X3121	1	ALLERGIC/IMMUNOLOGIC	NONE		
FAB-10	X3121	2	CARDIOVASCULAR	HISTORY / NOT ACTIVE	CORONARY ATHEROSCLERIC DISEASE	
FAB-10	X3121	3	DERMATOLOGICAL	NONE		
FAB-10	X3121	4	EARS, NOSE, THROAT	NONE		
FAB-10	X3121	5	ENDOCRINE	HISTORY / NOT ACTIVE	ELEVATED LIVER ENZYMES	
FAB-10	X3121	6	EYES	HISTORY / NOT ACTIVE	BILATERAL CATARACT REMOVAL	
FAB-10	X3121	7	GASTROINTESTINAL	HISTORY / NOT ACTIVE	DIVERTICULOSIS	
FAB-10	X3121	8	GENITO-URINARY	HISTORY / NOT ACTIVE	BPH. RNAL STONES	MILD RENAL INSUFFICIENCY
FAB-10	X3121	9	HAEMATOLOGICAL	NONE		
FAB-10	X3121	10	MUSCULOSKELETAL	CURRENTLY ACTIVE	MUSCLE CRAMPS	

If C4 is the body system and C5 the status, the remaining columns probably are verbatim descriptions

- There are several disease-related words
- Appears to be verbatim text
- Unclear why there are multiple columns of information.
  - Probably legacy data structure with each description in a separate column.

Definitive Identification



Medical and surgical history, concomitant diseases

Does the subject have any relevant medical or surgical history or relevant concomitant diseases?
Check the appropriate box for each system.

System	None	History and not active Yes, specify:	Currently active Yes, specify:
Allergic/Immunologic	☐	☐	☐
Cardiovascular	☐	☐	☐

- If we cheat and look at the CRF, it is obviously Medical History

Summary

- Current SDTM mapping technology depends on experts
- Severely limits throughput relative to all legacy data needed for a comprehensive clinical trial database
- A fully automated expert system that can perform SDTM conversions with a high probability of accuracy is a promising approach.