

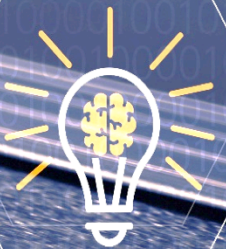


Strategy | Consulting | Digital | Technology | Operations

DH08: Efficiency Comes From Reusability and Repeatability

Hanming Tu, Accenture, Berwyn, USA

High performance. Delivered.



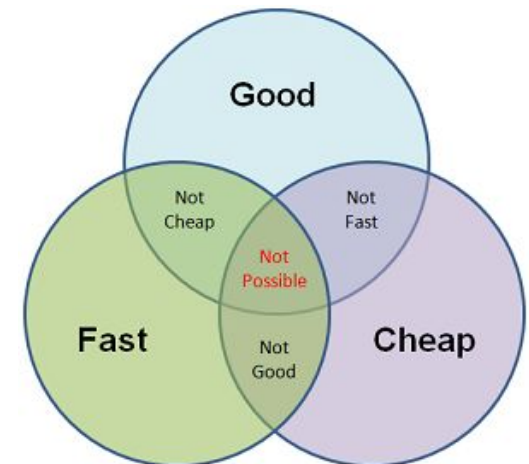
Agenda

- Issue Statement
- Create Mapping
- Build Reusable Codes
- Define Repeatable Workflow
- Check compliance
- Conclusion

Issue Statement

Project Triangle

- CDISC SDTM Data Conversion is part of an ETL (extraction, transformation and loading) process
- It requires to develop new ETL packages for each study and produces the same data structure and maintain correct relationships for all the studies
- How could we generate ETL packages that produce the quality SDTM data sets from vastly heterogeneous sources with high efficiency?



Issue Statement

Moving targets

- The CDISC SDTM model is the target but it is still evolving
- How do you manage the change of your target?
- How do you reuse the map, codes and process that you have developed?
- How to maintain or increase the quality while the target (scope), schedule (time) and budget (cost) may be still in flux.



Four Major Steps

- Create Mapping
- Build Reusable Codes
- Define Repeatable Workflow
- Check compliance

Create Mapping

Map Source to Target

- Mapping data sets to domains
- Mapping variables to columns
- Mapping specification

Mapping at Domain Level

Four types of SDTM Mapping

Simple match: one to one

Single dataset matches with single SDTM domain

- Simple but very rare
- Only if clinical data management system (CDMS) is SDTM compliance but still not 100%
- Example domains
 - Trial Arms -- TA
 - Trial Elements -- TE
 - Trial Visits -- TV
 - Trial Inclusion/Exclusion Criteria -- TI
 - Trial Summary -- TS

Converging match: many to one

Many to one: records from many different datasets merged into one SDTM domain

- Complex and common
- Represent dependent relationship
- Example SDTM domains or classes
 - RELREC: AE, CM, LB, MB, MS through studyid, rdomain, usubjid, idvar, idvarval, reltype, relid columns
 - SUPPQUAL: SuppAE, SuppCM, SuppDM, SuppEG, SuppEX, SuppMH, SuppPE, etc

Topic based match: domain aggregation

- Topic based match: Datasets with topic variables will be consolidated into SDTM observation classes
 - The TRT topic domains to Interventions
 - The TERM topic domains to Events
 - The TESTCD topic domains to Findings
- This has to be done for every study; then all the studies are pooled into one in the multi-study project.

Diverging match: one to many

One to many: records from one dataset split into many SDTM domains

- Common and complex
- Example SDTM domains
 - Demographics – DM → DM, DS, SC, SUPPDM
 - Laboratory Test Results – LB → LB, CO, SUPPLB

Mapping variables to columns

- Decoding or encoding: translating, controlled terms, lookup – value to code or vice versa
- Combining or splitting: many fields to a column or vice versa
- Transposing or pivoting: rows to columns or vice versa
- Selecting or filtering: variables or records
- Aggregating or deriving: new variables
- Generating: surrogate-keys

- Name: matched or renamed
- Data type: matched or casted
- Length: contained or split
- Label: matched or changed

MAP103			Study	Project	User Admin	Unlock			
Source Datas	Variable	Type	Format	Label	SDTM Domain	SDTM Variable	Mapping Comments	NOTES	Mapping Questions
DEMO	studid	char		Study Identifier	DM	SITEID	DO NOT MAP		
DEMO	sited	num		Unique Site Identifier	DM	COUNTRY	USA'		
DEMO	subjid	num		Unique Subject Identifier	DM	STUDYID	Refer to General Conventions tab.		
DEMO	subjid				DM	USUBJID	Refer to General Conventions tab.		
DEMO	subjid				DS	STUDYID	Refer to General Conventions tab.		
DEMO	subjid				DS	USUBJID	Refer to General Conventions tab.		
DEMO	subjid				SC	STUDYID	Refer to General Conventions tab.		
DEMO	subjid				SC	USUBJID	Refer to General Conventions tab.		
DEMO	visit	char		Clinical Planned Event			DO NOT MAP		
DEMO	visitrn	num		Visit Number			DO NOT MAP		
DEMO	page	char		Page No.			DO NOT MAP		
DEMO	visdt	num	DATE8	Visit date	DM	DMDTC			
DEMO	consdate	num	DATE8	Date of informed consent	DS	DSSTOTC			
DEMO	consdate				DS	DSTERM	INFORMED CONSENT'		
DEMO	consdate				DS	DSDECD	INFORMED CONSENT'		
DEMO	consdate				DS	OSCAT	PROTOCOL MILESTONE'		
DEMO	BIRTHDAT	num	DATE9	Date of Birth	DM	BRTHDTC			
DEMO	AGE	num		Age (years)	DM	AGE		AGE is equal to age at informed consent.	
DEMO	AGE				DM	AGU	YEARS'		
DEMO	GENDER	char	\$SEX.	Subjects Gender	DM	SEX	When GENDER=1 then SEX='M' When GENDER=2 then SEX='F' When AIAN='I' then HALE='American Indian/Alaskan Native'		
							When AIAN='I' and WHITE='I' then RACE='MULTIPLE' and AIAN is mapped to QVAL where QJAM='RACET' and WHITE is mapped to QVAL where QJAM='RACE2' (Only occurs once, for SUBJID:102012)		
DEMO	AIAN	char	\$YESONLY.	American Indian/Alaskan Native	DM	RACE	Else do not migrate		
DEMO	AIAN				SUPPDM	ROOMAIN	DM'		
DEMO	AIAN				SUPPDM	IDVAR	DMSEQ'		
DEMO	AIAN				SUPPDM	IDVARVAL	DMSEQ		

Build Reusable Codes

Important considerations for increasing the code reusability

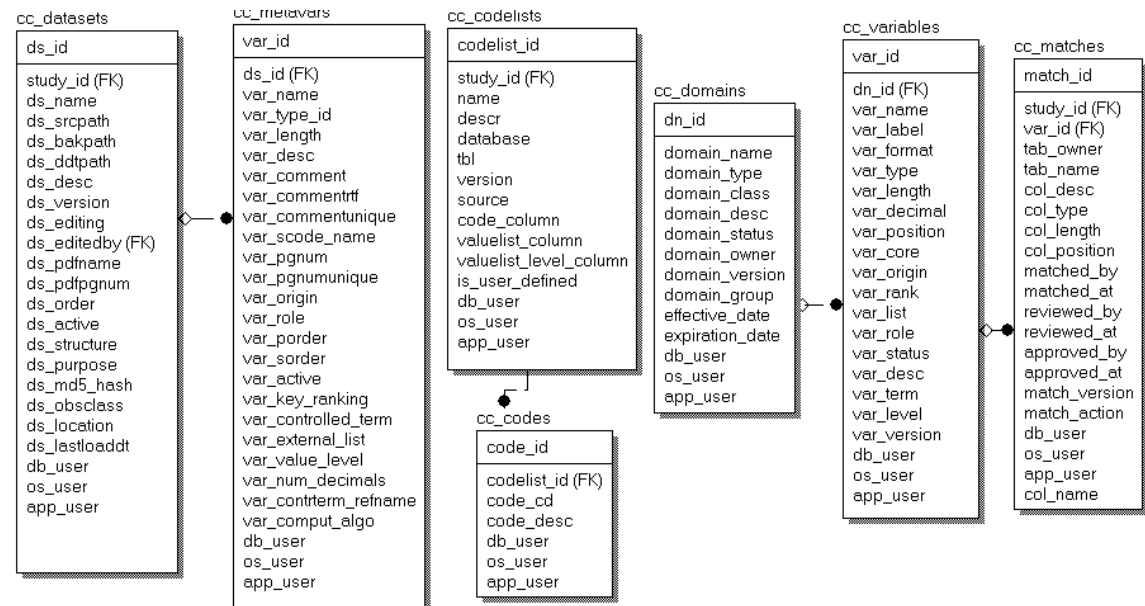
- **Standard adoption** is the key for code reusability
 - Train people to understand the standards
 - Define standard templates
 - Build public libraries for code snippets and public transformation: Custom functions, procedures and packages; public data rules; and public Experts
 - Group code snippets and functional transformation into modular mapping and transformation: pluggable maps
 - Define workflow to govern the process: Workflow Manager and Process Flows

Build Reusable Codes

Key consideration: Metadata

- **Metadata-driven** process is the key for automation
 - Metadata makes data meaningful
 - Metadata is machine readable
 - Metadata is the base for automation

Metadata model
used in the
system



Define Workflows: Process Repeatability

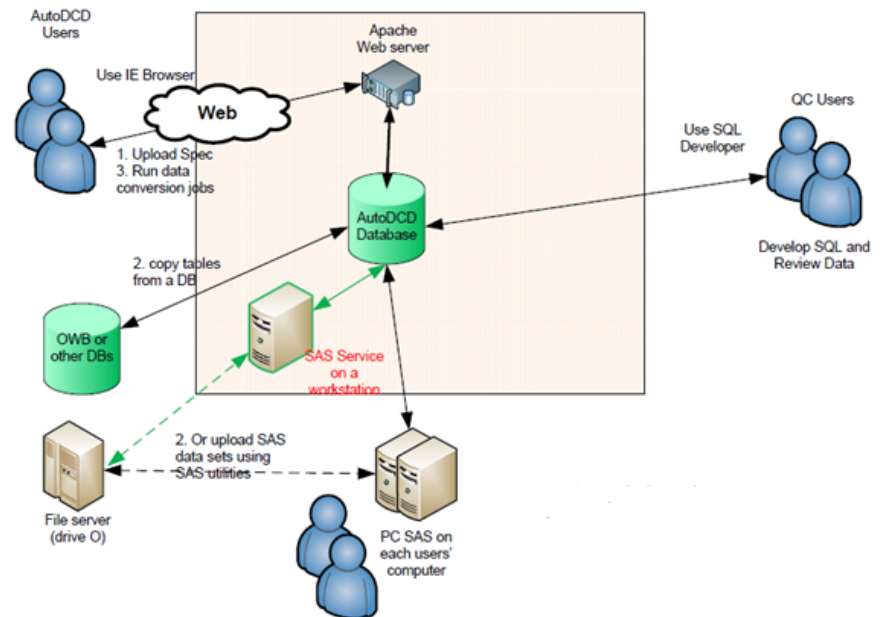
How to link reusable silo codes into a repeatable process?

- Built an automatic data conversion system to link the pieces in data integration and standardization process.
- Provides the workflow to link maps to form a controlled data flow, from vertical code reusability to horizontal process repeatability.
- The data conversion portal provides a web-based software tool for the Data Integration and Standardization (DIS) Department to use.
- The portal improves and accelerates the Data Conversion Development (DCD) process.

Define Workflows: Features

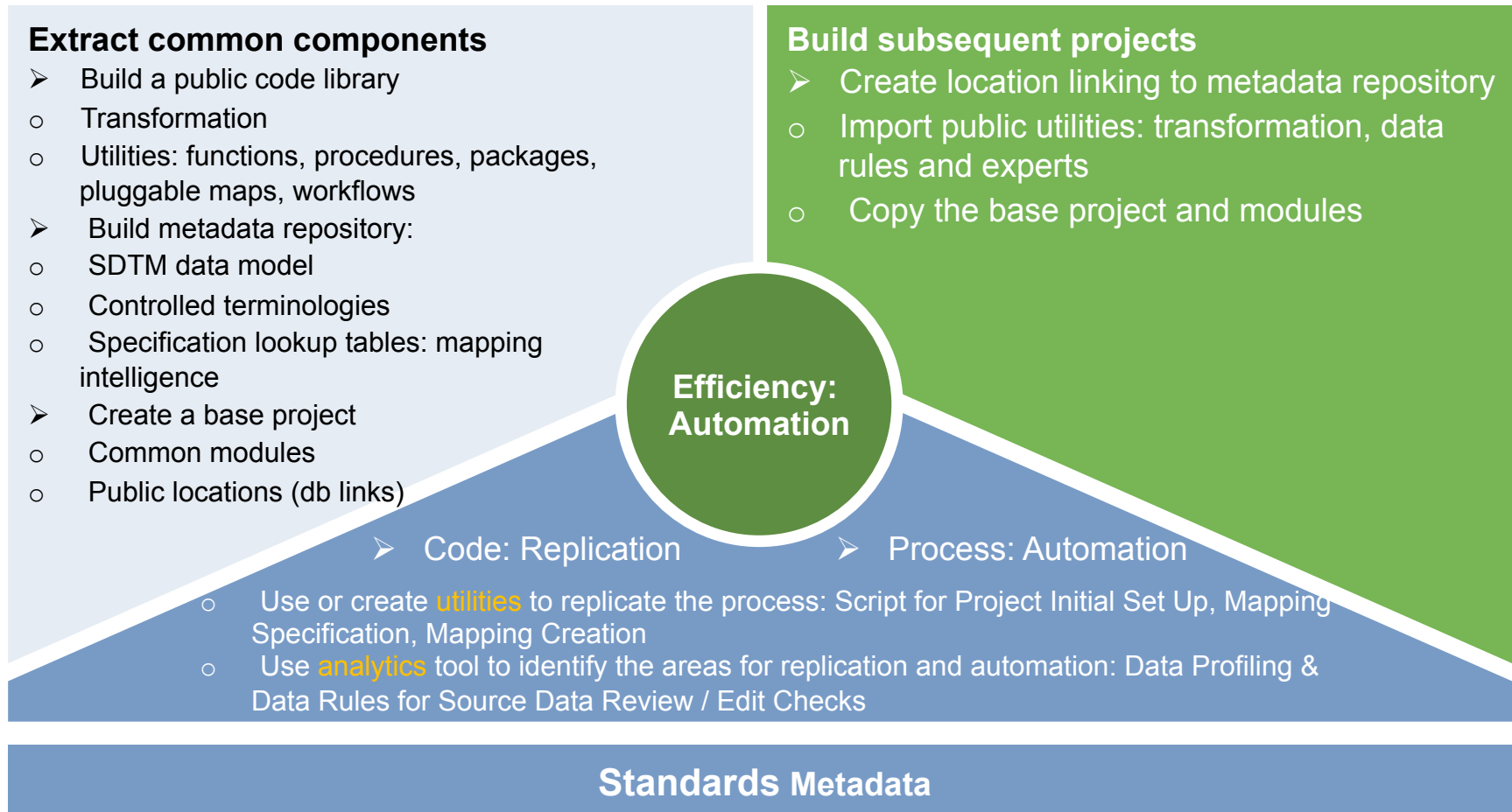
Features implemented in the system

- Load and store map specification in relational database
- Manage workspace with client, project, study, and specification hierarchy
- Allow users to create and delete intermittent views and tables that we used in mapping
- Run data conversion jobs by domain or by a group of domains or all
- Link and copy tables from an Oracle database or use the tables created and loaded through SAS upload utility
- Keep audit trails for each job
- Track the performance of each job



Build Repeatable Process

Intelligence based automation: is it **fast**?



Approach Comparison

Comparison among the approaches and tools

Traditional Approach: No standard/No metadata	OWB Approach: Standards/No metadata	Accenture Approach: Standards/Metadata
ETL using custom programming such as SAS, PL/SQL, JAVA, Perl, etc.	ETL with User Interface	Web-based User Interface
High paid programmers	Users do not need to know the programming language – PL/SQL	No PL/SQL programming is needed
No audit trail	In an audited environment	In an audited environment with validated products
No security	Built-in security: database and OWB security	Authenticated and authorized users only with audit trails
No consistence among coding	Consistence with all the users	Automatic and consistent coding
Difficult to manage and support	Easy to manage and support	Easy to manage and support
Scalability: silo and not scale; through adding more manpower	Scalability through hardware and software	Very scalable

Check the Compliance

The quality questions: is it **good**?

- How could we verify whether the data sets compliance to the standards or the defined quality?
- How do we measure the quality of the work?
- We need to have some sort of governance in place to ensure the quality of data sets and set of metadata to describe the quality of the data.
- Clinical data quality information is semantic information about clinical data quality (CDQ), including how a clinical trial is conducted and how the data elements are collected, entered, processed, and analyzed.
- CDQ concerns not just data accuracy, but also data traceability and compliance against common standards.

Check the Compliance

Compliance Standards

System-based approach

- Paper-based (DDE):
 - Source-to-database error rates: 976 per 10,000 fields
 - CRF-to-database: 14 errors per 10,000 fields
- Fax-Based (OCR)
 - Re-fax Rate: 5~20%
- EDC systems
 - Source-to-database: 50 errors per 10,000 fields

Standard-based approach

- Process standards: GxP where x=L, C, M, etc.
- Data standards:
 - Structure: CDISC ODM, CDASH
 - Value (metadata and controlled terminology):
 - ISO 11179 – IT -- Metadata registries (MDR) – Organizations
 - ISO/IEC – Software System
 - ISO 20943 - IT -- Procedures for Achieving Metadata Registry Content Consistency
 - ISO 21090 - Healthcare Data types
 - ISO 23081 – Records management
 - CWM – Data warehousing; RDF – Web resources; DIF – Scientific data sets
 - CDISC CSHARE
 - Content: CDISC SDTM
- Verification standards:
 - Validation checks: WebSDM checks
 - JANUS checks: FDA checks for data loading
 - Severity Definition: Low, Medium and High

Run through compliance checks

- **Consistency:** Checks data in 2 or more columns to ensure data correspond in cross-column (visit number without visit description, age unit without age), cross-domain (SUBJID in a domain but not in DM) or cross-system (external dictionary).
- **Format:** Checks if data are in an allowable format such as ISO 8601; Leading and trailing spaces; and missing value “.” in character column
- **Limit:** Checks if data are within range such as start/end time, and toxicity grade.
- **Metadata:** Checks if tables and columns have valid metadata
- **Presence:** Checks data that are missing or present
- **Referential:** Checks if a described table/record data relation are valid; SDTM is not 3NF, referential information is stored as data (RDOMAIN, USUBJID, IDVAR, IDVARVAL) , in Supplemental Qualifiers (SUPPQUAL), related records (RELREC) and comments(CO).
- **Value:** Checks data against valid values such as code lists, Illegal values

Check Compliance

Intelligent automation: is it **smart**?

Categories	Standard FDA	Enhanced
Consistency	46	38
Format	6	53
Limit	10	2
Metadata	8	45
Presence	4	33
Referential	8	4
Value	27	42
Σ	109	217

- Ensure data quality
- Provide insights for risk-based monitoring
- Provide insights for redesigning CRFs
- Provide insights for transformation automation

$$DPMO = 1000000 * \sum_{t=1}^n \sum_{i=1}^m \frac{e_i}{d_i}$$

Where i is a domain, a class or a study in SDTM; m is total number of domain, classes or studies. The t is an investigator site, a project, a client or a compliance check category; n is total number of sites, projects, clients or compliance categories.

$$e_i = \sum_{c=1}^C error_c$$

Where c is each compliance check; C is total compliance checks being run for a domain, a class, or a study in SDTM. Error is number of issues being found.

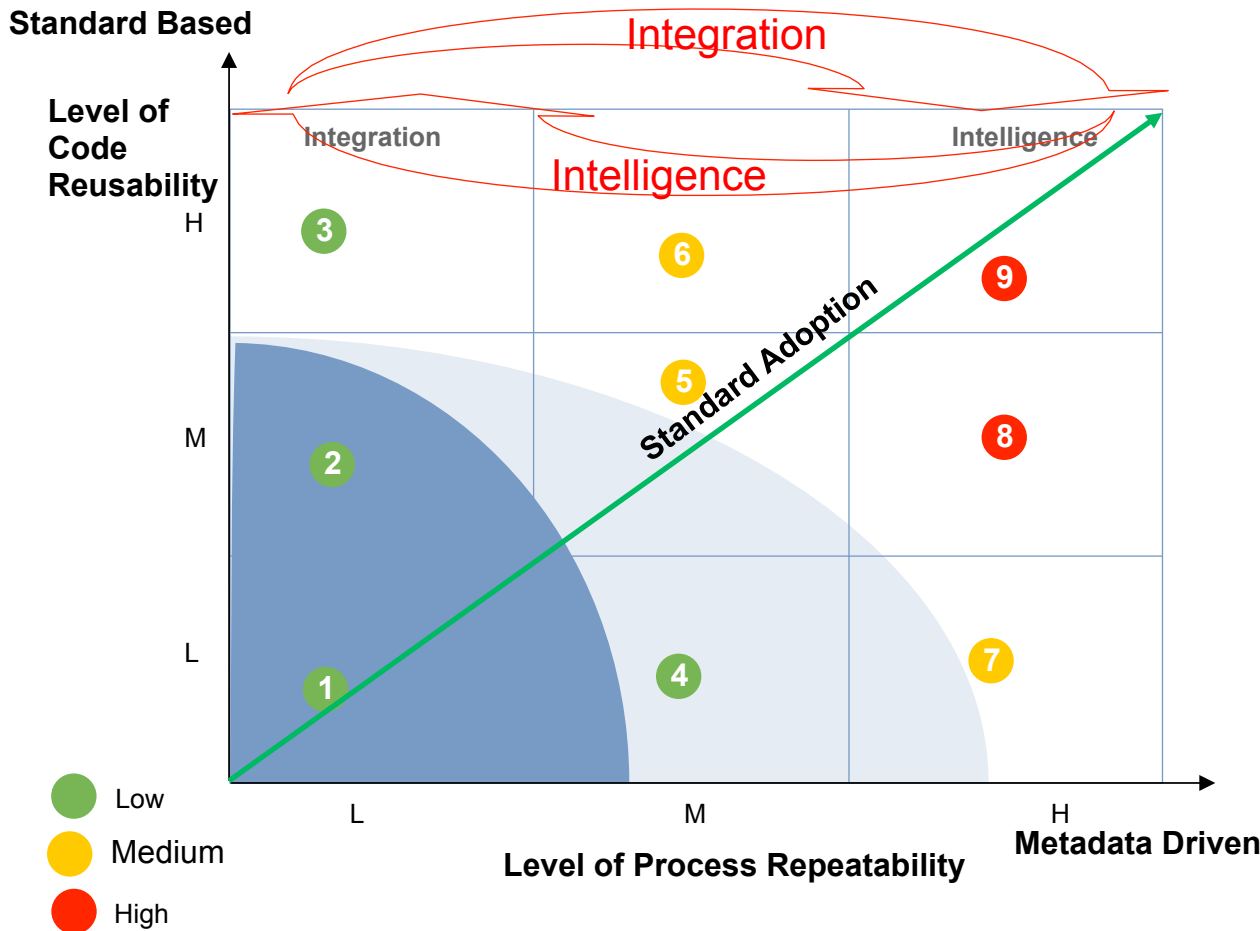
$$d_i = \sum_{j=1}^J C_j R_j$$

Where j is an observation or row of data; J is total number of observations or records being run through a compliance check. C is number of columns or variables; R is number of rows or observations.



Conclusion: Efficiency Level Matrix

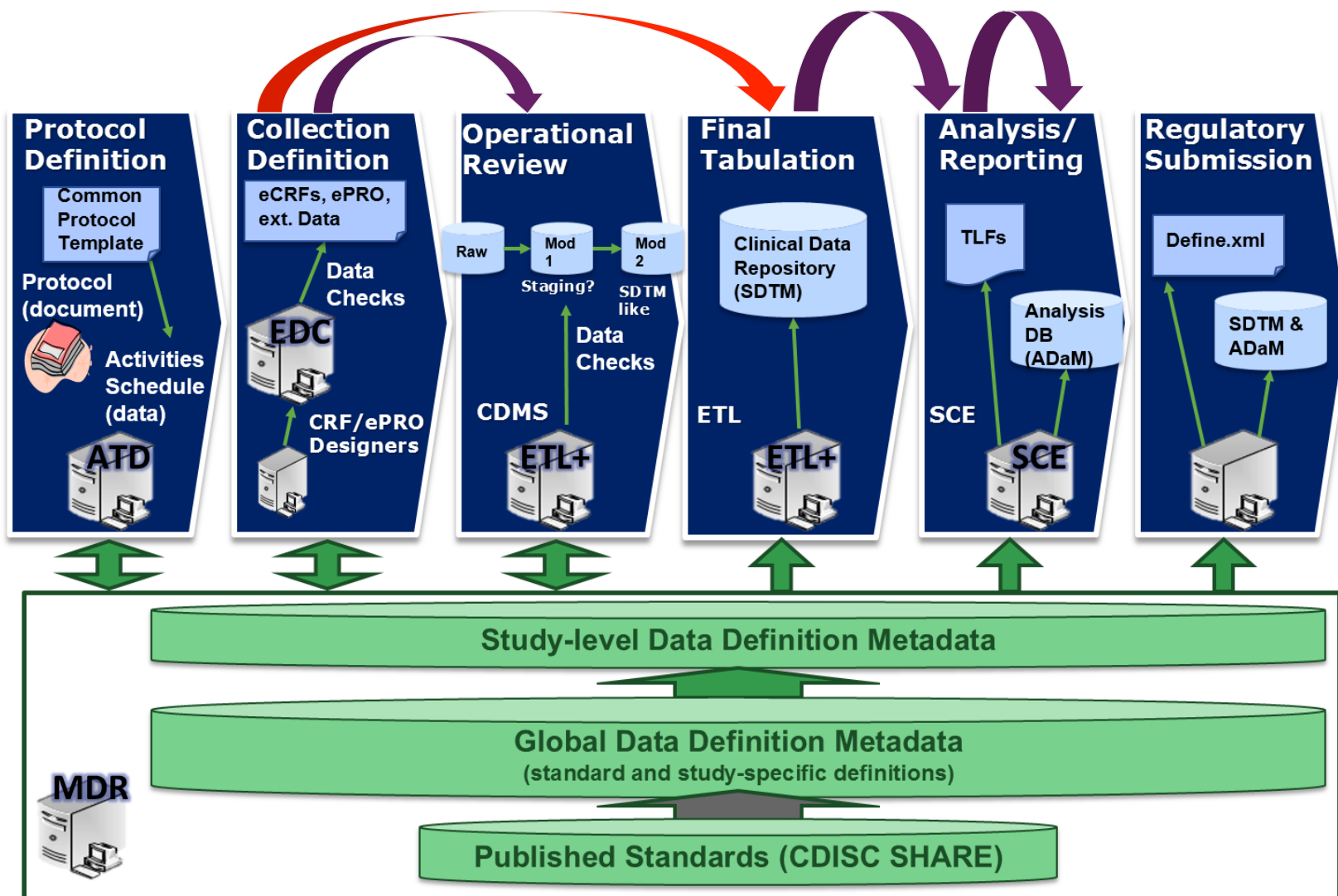
Standard-based systems allow for integration while metadata-driven systems enable automation;
The more intelligence collected about the clinical data, the more integration could be;



List of Efficient Levels:

- 1 No code reuse – double programming for every study
- 2 Code reuse at function level
- 3 Code reuse at module level and company code standard and reusable code library exists
- 4 Adopted standards in part of the process in silo systems
- 5 Adopted standards in some part of the process with some code reusability
- 6 Adopted standard enable the high reusability in the process
- 7 Adopted standard but have not build up metadata and no intelligence
- 8 Adopted standards and start learning the insight of the process
- 9 Adopted standard in all parts and with learned intelligence applied to the process

Further automation through “Intelligent Data Flow”



Q & A

Questions and Answers



Contact Information

Hanming Tu	Dave Evans
P: 610-407-1817; C: 484-881-2384 E: hanming.h.tu@accenture.com	P: 484-881-2411 E: dave.a.evans@accenture.com
Address: 1160 West Swedesford Road, Berwyn, PA 19312, USA	
Web: www.accenture.com	
Fax: 610-535-6615	