PhUSE 2016

Paper DH05

How a Metadata Repository enables dynamism and automation in SDTM-like dataset generation

Judith Goud, Akana, Bennekom, The Netherlands
Priya Shetty, Intelent, Princeton, USA

ABSTRACT

The traditional clinical data lifecycle (CDLC), programming tasks involve data integration from disparate sources and late-stage legacy data conversions, which are inefficient and complicate full data traceability. In addition to study level programming, standards management and governance processes need to be put in place to manage the compliance to regulatory guidance.
However, if managed well, standards and metadata can be used to drive automation of business processes like database creation, CRF annotation and dataset generation.

While a centrally managed metadata repository can accelerate the implementation of standards and facilitate regulatory compliance, it also supports reusability of standard algorithms, automation of business processes and full data traceability.

This paper describes a metadata-driven data flow use case for a dynamic and automated platform that generates SDTM-like datasets from raw datasets based on metadata managed in a repository.

The advantage of this approach is that all metadata, including transformational metadata, can be managed, validated and governed centrally, while facilitating faster, more consistent dataset generation.

INTRODUCTION

The Biopharma industry has long been seeking solutions for dynamism and automation of data warehouses/hubs and business/operational processes as a means to accelerate drug discovery and development. The key trends and drivers for dynamism and automation are outlined in Figure 1 below:

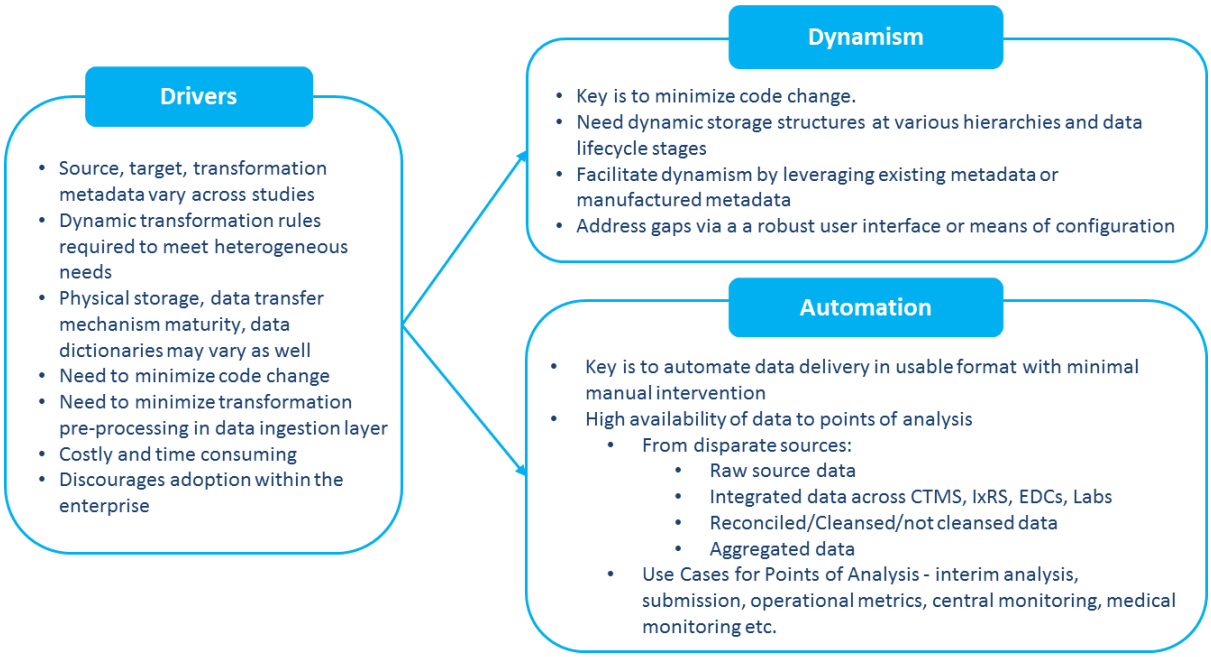

Figure 1. Trends – Dynamism and Automation in Integration

Several challenges continue to persist for this vision, such as disparate data systems/sources and formats, reacting to changes in standards (such as CDISC SDTM/ADaM) and other regulatory guidance, transforming the data quickly as per different business needs and with minimal change control, while maintaining full traceability. Therefore, the overall clinical development lifecycle and enhancement of scientific capabilities still has significant opportunities for further growth and improvement[1].

In addition, traditional CDLC programming tasks involve integrating data from disparate sources and late-stage legacy data conversions based on constantly changing standards, which pose several challenges in terms of efficiency, data traceability, and change control. Global and study level standards management and governance processes are needed to help address these challenges, along with managing compliance to regulatory guidance.

To enable efficient data exchange, validation, and process automation, attention needs to be given to the data about the data – metadata. A platform that will allow data integration from any system in any format and transform this data into any desired target format can drive further process efficiency.

A centralized metadata repository and governance platform to manage ever changing standards and a flexible source agnostic data integration platform can be used to drive automation of processes like collection database creation, data integration from disparate sources, CRF annotation and dataset generation.

This approach provides immense value for bringing in dynamism and automation for various downstream functions such as analysis ready dataset generation, operations reporting, and risk based monitoring. These benefits include accelerating standards adoption, promoting reusability of mapping algorithms across studies, providing full data traceability, and supporting regulatory compliance.

This paper specifically describes a metadata-driven data flow use case for a dynamic and automated platform and process that integrates clinical data from multiple sources and generates SDTM-like datasets from raw datasets based on metadata managed and governed in a repository. The advantage of this approach is that all metadata (structural and transformational) can be validated and governed centrally, while accelerating the delivery of analysis ready datasets to downstream systems.

DATA STANDARDS AND END-TO-END IMPLEMENTATION

The traditional drug development lifecycle has been fairly linear and brittle, leading to longer timelines and limited reuse of information. To help make the CDLC more iterative and adaptable to change, standards and metadata for artifacts created and used throughout this lifecycle should be managed in order to facilitate their automated reuse. The Biopharma industry can make strides to reduce clinical development time and costs by implementing standards via introduction of agile processes into the CDLC and pairing those processes with a parallel Metadata Lifecycle as illustrated in Figure 2[1]. Instilling this flexibility within the lifecycle facilitates ongoing learning throughout a study and an increased ability to respond to change.

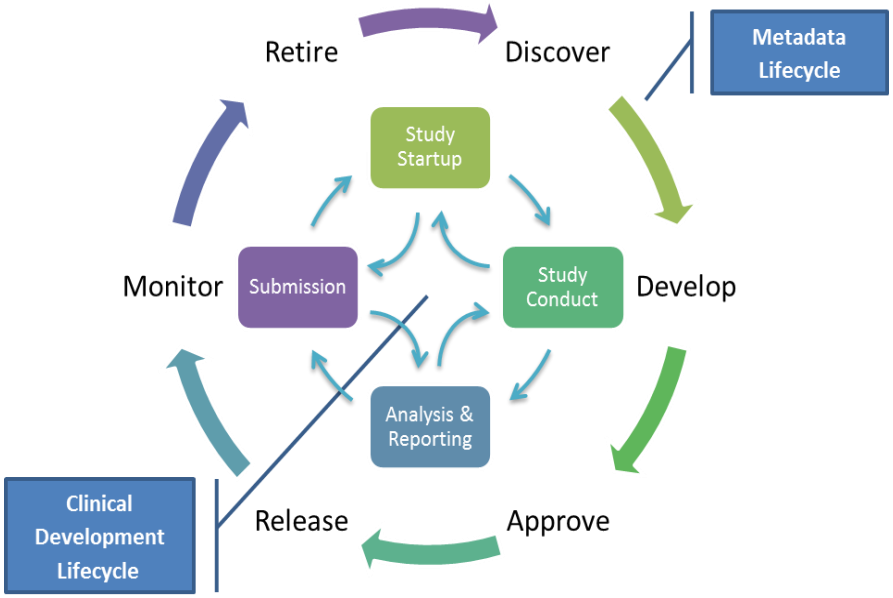

Figure 2. Clinical Development Lifecycle (CDLC) with Metadata Lifecycle

When there is a need to create a clinical artifact such as a dataset specification, a process can be followed to **discover** any existing data, metadata elements, or artifacts that can be reused instead of creating them from scratch. When needed metadata or data is not available, new artifacts can be **developed**. Once an artifact is complete, all of its new components go through a review and **approval** process to ensure it is compliant with appropriate standards and policies, and **released** for use and reuse by other studies or systems.
Monitoring should be done for curation purposes (e.g. change requests or outside influences that may impact metadata) and at some point metadata content may no longer be actively used and should be **retired**.

As an example, in the traditional CDLC, the creation of clinical artifacts is generally slow, error prone, resource intensive and costly. Organizations develop each artifact from scratch through manual review and interpretation of standards and artifacts created earlier in the lifecycle and/or from other studies.

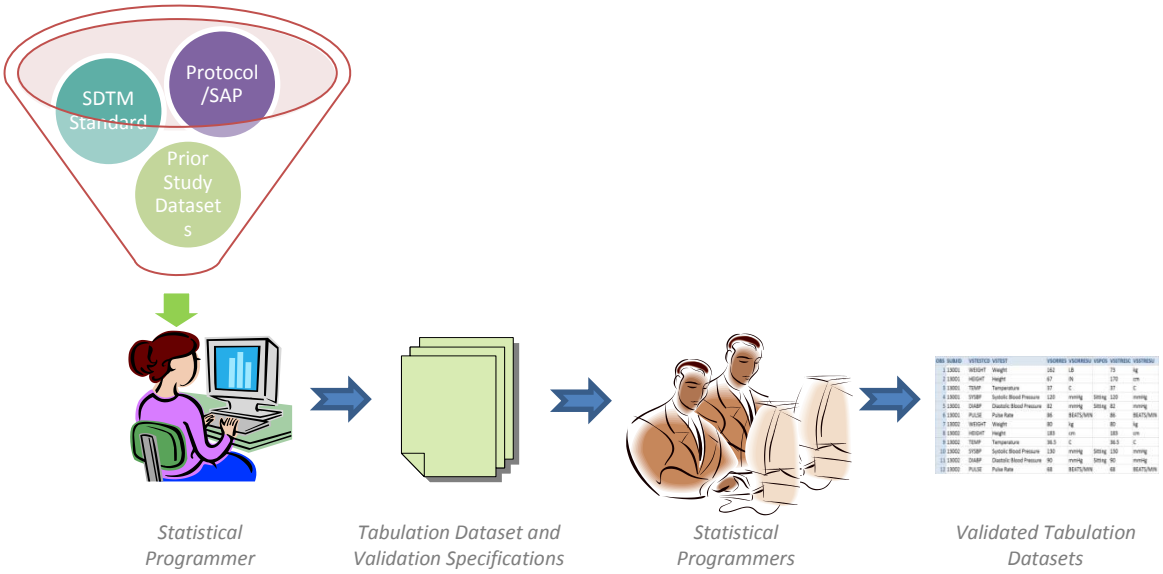

Figure 3 Traditional Approach to Validated Tabulation Datasets Development

For example, as illustrated in Figure 3, to create study specific datasets, a statistical programmer and/or statistician starts with reviewing dataset standards from previous similar studies and the given study protocol and statistical analysis plan. S/he manually reconciles differences and drafts the study dataset specification, which is sent for team review and approval. In addition, validation specifications are drafted to indicate how the target datasets will be validated and sent for review and approval. The approved dataset and validation specifications are sent to statistical programmers who then develop the study dataset programs. These programs are tested and approved before being released with supplemental documentation, such as an annotated CRF and Data Reviewers Guide.

These metadata artifacts are created and managed in a loosely coupled and often unstructured manner - standards maintained in Word/PDF formats, specifications maintained in Excel, programs and macros developed in SAS™ or other technologies, and different artifact versions are tracked in different systems/networks. These challenges are amplified for studies that run for several months and years, with protocol amendments, new versions of standards and dictionaries, and changes in study team resources.

One of the approaches to address these CDLC challenges is the metadata driven approach – both structural and transformational metadata – and reusability of this metadata from one study to another. Standards and metadata for artifacts created and used throughout this lifecycle must then be managed in a centralized repository in order to facilitate their automated reuse. There is a big difference in how information is managed and used as opposed to traditional processes.

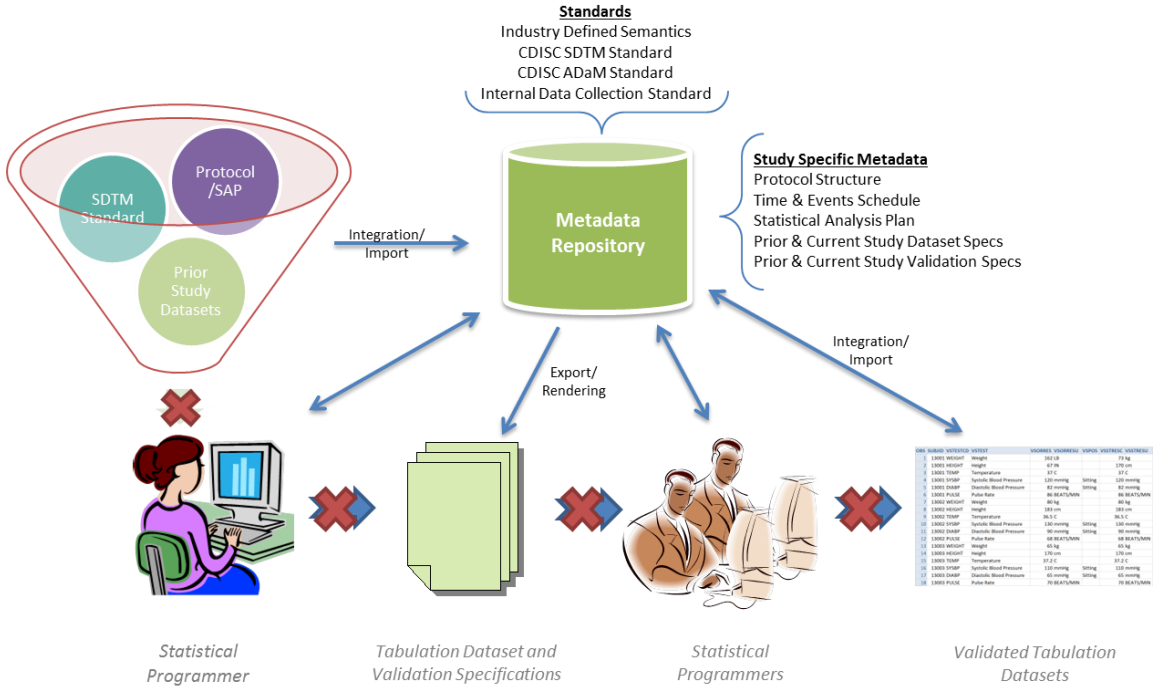

Figure 4. Data Tabulation process using a central metadata repository

As illustrated in Figure 4, this approach places a metadata driven flow as the driving force. Metadata related to source and target data structures, and transformational metadata related to the source to target mapping or derivations should be stored and governed in a centralized repository. There is always a certain percentage of source to target mappings that can be applied as is from one study to another. For example in Safety domains of studies for a given therapeutic area, there is a limited percentage of variation to be customized from one study to another. In Efficacy domains, the level of custom programming required may be higher than standard domains, however there is still a fair amount of reuse expected to be leveraged for studies within the same therapeutic area.

This reuse promotes considerable time, effort and eventually cost savings. Mapping libraries can be maintained at various levels such as, Global, Therapeutic Area and study specific.

Thus, rather than the statistical programmer or statistician manually interpreting the protocol, statistical analysis plan, data tabulation standards and prior study datasets, s/he has all of those definitions already in a centralized repository. S/he can re-use applicable existing content to create a new set of metadata to represent the study specific datasets and validation specifications, only adding items for structural and transformational metadata that are required to be customized for that specific study.

As depicted in Figure 5, a centrally managed metadata repository helps accelerate and maintain the implementation of new versions of standards and regulatory requirements. It further supports reusability of standard algorithms, automation of business processes and full data traceability. Taking advantage of the standards and previously defined artifacts and all of their components, the clinical development lifecycle can be significantly streamlined and much more flexible and adaptable to change, resulting in higher quality data, business process automation, and more efficient use of resources.

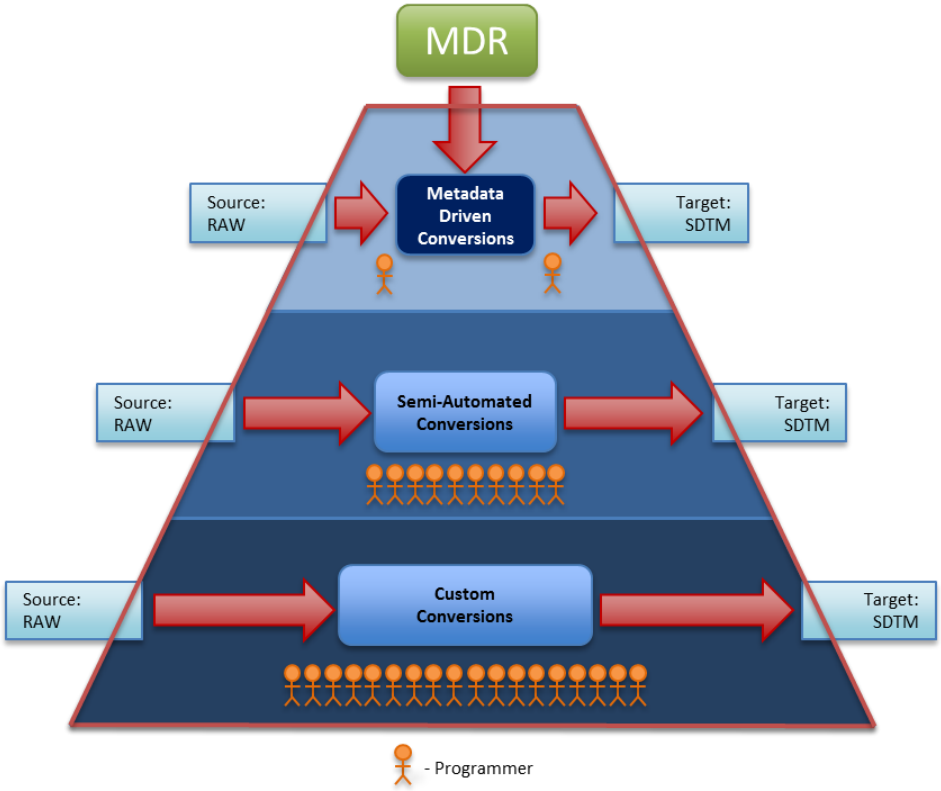

Figure 5. Metadata Driven Conversions

The following section depicts the use case flow of managing metadata in a centralized repository, Semantics Manager™ (SM), and leveraging this metadata to enable a dynamic automated flow from various sources to target analysis ready datasets (SDTM+/-). SM is an object-oriented metadata repository solution that enables consistent metadata management and governance, as well as business process automation.

SAS SDTM-LIKE DATASET GENERATION USE CASE

To help alleviate ongoing challenges in traditional data integration and transformation approaches described in the previous section, Intelent and Akana collaborated to present a metadata-driven data flow approach that dynamically generates target datasets from raw datasets based on metadata managed in the repository.

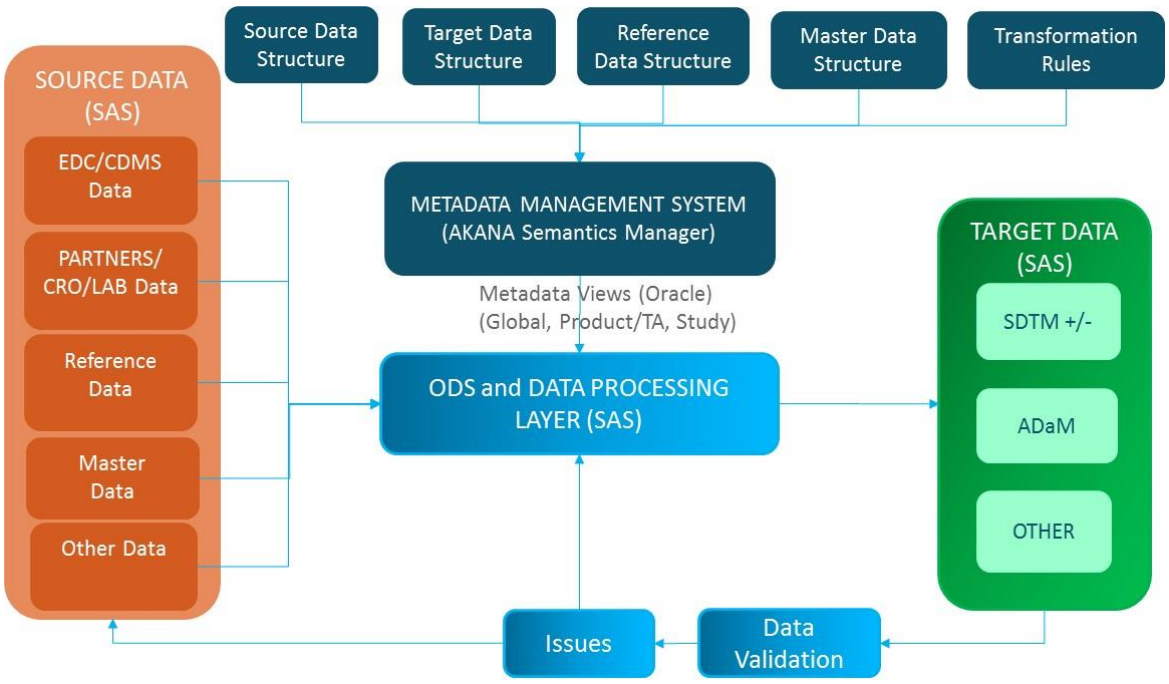

Figure 6. Metadata Driven Clinical Data Flow

As depicted in Figure 6, the goal is to maintain and manage structural and transformational metadata in SM, extract the metadata into a custom view, build a dynamic and automated process which would integrate data from disparate sources and apply the SM metadata to dynamically build the target datasets. This process was applied and demonstrated for a few key representative domains such as DM, DS, EX, AE, LB, and VS.

For these domains, metadata was imported into SM for the source and target dataset structures and the transformations for the variable and domain level source to target mappings. A custom view was compiled in SM to summarize the structural and transformational metadata. This view categorized the types of derivations/transformations and also included the order in which these derivations have to be executed across variables and domains for a given study. An excerpt of this view is displayed in Figure 7.

Domain Prefix	Variable Name	Seq. For Order	Variable Label	Type	Sort Order	SOURCE DOMAIN	SOURCE FIELD	TRANFORMATION (SAS)	Derivation	Order of Execution	Derivation Parameters	Length	Used for the study
AE	AEYN	1	AE Occurred?	Char		AE	AEYN		SUBSET	2	AEYN="Y"	1	Y
AE	STUDYID	2	Study Identifier	Char	1	AE	STUDYID		COPY	1		20	Y
AE	DOMAIN	3	Domain Abbreviation	Char					ASSIGN	3	"AE"	2	Y
AE	USUBJID	4	Unique Subject Identifier	Char	2	AE	STUDYID, SUBJID		CONCAT	4	"-"	20	Y
AE	AESEQ	5	Sequence Number	Num		AE			SEQUENCE	99	Concatenate using separator	8	Y
AE	AEGRPID	6	Group ID	Char		AE	AEGRPID		COPY	1		60	N
AE	AEREFID	7	Reference ID	Char								20	N
AE	AESPID	8	Sponsor-Defined Identifier	Char	5	AE	AESPID		COPY	1		10	Y
AE	AEPRESP	9	Pre-Specified Adverse Event	Char		AE	AEOCCUR	IF AEOCCUR IN("Y", "N") THEN AEPRESP="Y"	DERIVE	5		1	Y
AE	AEOCCUR	10	specific adverse event	Char		AE	AEOCCUR	IF AEOCCUR NOT IN("Y", "N") THEN AEOCCUR=""	DERIVE	6		1	Y
AE	AETERM	11	Reported Term for the Adverse Event	Char	3	AE	AETERM		COPY	1		200	Y
AE	AEMODIFY	12	Modified Reported Term	Char			AEMODIFY		COPY	1		200	Y
AE	AELLT	13	Lowest Level Term	Char		AE,reflib.Meddra11_0	LLT_NAME		MERGE	14	AELLTCD#LLT_CODE	200	Y
AE	AELLTCD	14	Lowest Level Term Code	Num		AE	AELLTCD		COPY	1		8	Y
AE	AEDECOD	15	Dictionary-Derived Term	Char		AE,reflib.Meddra11_0	PT_NAME		MERGE	15	AELLTCD#LLT_CODE	200	Y
AE	AEPTCD	16	Preferred Term Code	Num		AE,reflib.Meddra11_0	PT_CODE		MERGE	16	AELLTCD#LLT_CODE	8	Y
AE	AEHLT	17	High Level Term	Char		AE,reflib.Meddra11_0	HLT_NAME		MERGE	17	AELLTCD#LLT_CODE	200	Y
AE	AEHLTCD	18	High Level Term Code	Num		AE,reflib.Meddra11_0	HLT_CODE		MERGE	18	AELLTCD#LLT_CODE	8	Y
AE	AEHLGT	19	High Level Group Term	Char		AE,reflib.Meddra11_0	HLGT_NAME		MERGE	19	AELLTCD#LLT_CODE	200	Y
AE	AEHLGTCD	20	High Level Group Term Code	Num		AE,reflib.Meddra11_0	HLGT_CODE		MERGE	20	AELLTCD#LLT_CODE	8	Y
AE	AECAT	21	Category for Adverse Event	Char								60	N
AE	AESCAT	22	Subcategory for Adverse Event	Char								60	N
AE	AEBODSYS	23	Body System or Organ Class	Char		AE,reflib.Meddra11_0	SOC_NAME		MERGE	21	AELLTCD#LLT_CODE	200	Y
AE	AEBDSYCD	24	Body System or Organ Class Code	Num		AE,reflib.Meddra11_0	SOC_CODE		MERGE	22	AELLTCD#LLT_CODE	8	Y

Structural and Transformational Metadata extracted from SM drives dynamic SDTM generation program

Derivation Types

Derivation order of execution

Figure 7. Source and Target Metadata extracted from Semantics Manager

As described in Figure 8, a dynamic process was developed in SAS, including data from an EDC system for representative domains such as DM, DS, EX, AE and VS, integrating external data from flat files for LB, and adding reference data pertaining to MedDRA dictionaries and controlled terminologies/code lists.

This dynamic process then applies the transformational metadata from the custom view extracted from SM, and generates SAS datasets automatically for a given study. This process can be executed in batch mode for scheduled loads, or manual mode for an on-demand load. The target datasets are validated using validation scripts to identify any anomalies or issues from expected mapping rules. Issues identified may result in source data updates (to be managed by source system owners) or metadata updates to be managed in SM (addition or modification of transformation rules, or source or target structure).

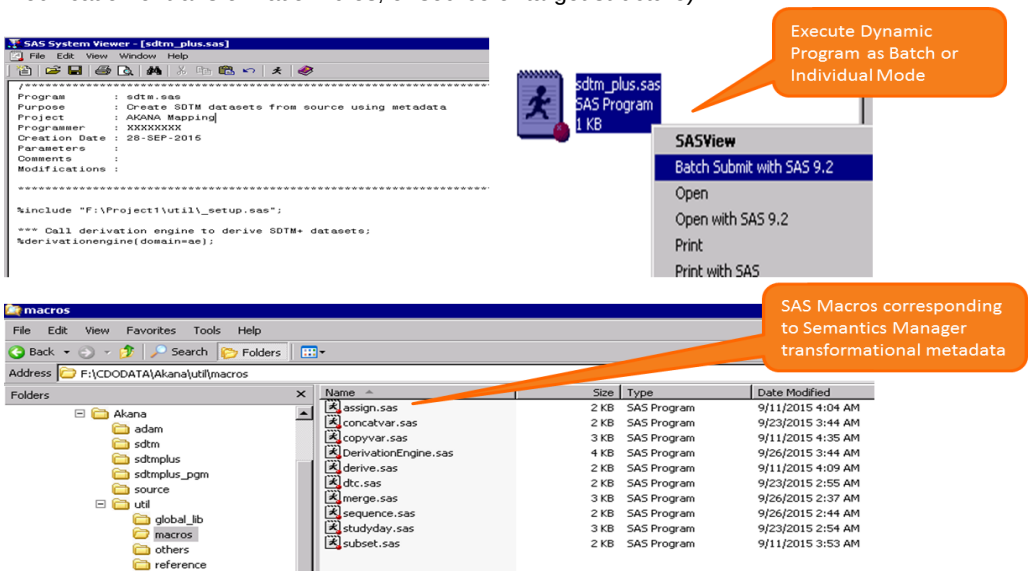

Figure 8. Dynamic Program Execution to generate Target Datasets

Figure 9 below depicts some examples to demonstrate the dynamism and automation realized with this approach. If there is a change in the transformation rule, such as updating the USUBJID derivation rule by introducing additional characters, the modification is done in SM, and the SAS process is executed on all applicable studies to apply the change in the target real-time. The second scenario depicted is the merge of AE source data with MedDRA for populating all the required variables for AE coded terms as recommended by the SDTM standard. In the third scenario, for events where there is an update in source data such as date, when the process is run either in scheduled or on-demand load, the target date is also automatically updated.

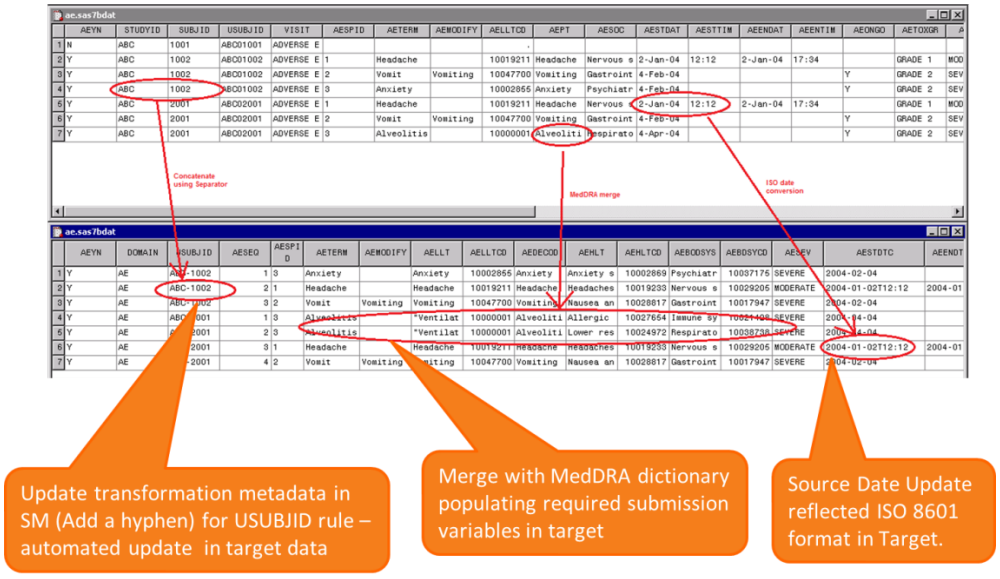

Figure 9. Metadata driven Source to Target Transformation Examples

For each of these scenarios, there is no manual programming required. The metadata is updated in SM, and the load for the dynamic process is either run in scheduled or on demand form to automatically apply the required updates in the target SDTM like structures. Not only is the data delivered quicker to the downstream systems, but this approach also demonstrates limited need of change control from one study to another, reuse of algorithms across studies and from standards, and enables efficiency for validation cycles, especially for standard safety domains. This helps efficiently address the challenges that had been described in the earlier sections of this paper.

This use case thus demonstrates tangible benefits such as

- High availability of data (source, integrated, and standardized) facilitates rapid insights and accelerated timelines for submissions
- Dynamic and automated generation of datasets across studies with minimal customization necessary from one study to the next results in cost and time savings
- Accelerated response to metadata update drivers like new versions of standards, new regulations and/or changes in study-level requirements facilitate regulatory compliance.
- Reusability of standard algorithms result in enhanced efficiencies for validation
- Dynamic and automated processes driven by metadata limit the need of change control
- Full traceability for audit tracking provides controlled standards management.

In addition to this use case, there are several other benefits that can be realized with a metadata driven approach using SM in combination with a source agnostic data integration platform. E.g. integrate data from a variety of sources such as EDC, CTMS, Labs, ePRO, Imaging etc., and leverage the centrally managed and governed metadata to support potential visualizations and reports to support functions such as trial operations reporting, aggregated EDC reporting, interactive safety review, and risk based monitoring. In context of the use case described in this paper, metadata creation and governance in a centralized repository and a source agnostic data aggregation platform are the key tenets to support analysis ready dataset generation across studies.

CONCLUSION

A metadata-driven data flow that dynamically generates target datasets from raw source datasets based on structural and transformational metadata managed in a repository is a more advantageous approach compared to traditional data flow processes. The use of a centralized repository to manage and govern metadata and standards in a flexible source agnostic data integration platform allows a dynamic, automated data flow – managing all changes in the centralized repository and providing views or visualizations for downstream consumption of data (AS IS, standardized or transformed), while maintaining versions of standards and full traceability from source to target. This brings significant savings in time and costs for the CDLC, bringing in greater agility to adapt to changes quickly.

As mentioned in conclusion above, the benefits can be realized with Akana's Semantics Manager which provides a flexible mechanism for creation and governance of data standards, clinical artifacts and transformational metadata, combined with a metadata driven implementation approach for data integration and dissemination through subject matter and technology expertise and services provided by Intelent.

REFERENCES

1. Smiley, J, Moving to a Standards-Based, Agile Clinical Development Lifecycle, white paper, http://resource.akana.com/resource/whitepaper/moving-standards-based-agile-clinical-development-lifecycle

CONTACT INFORMATION

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

Judith Goud
Akana, Inc.
12100 Wilshire Blvd, Suite 1800
Los Angeles, CA 90025
(866) SOA-9876
Judith.Goud@akana.com
www.akana.com

Priya Shetty
Intelent, Inc.
500 Alexander Park, #102
Princeton, NJ 08540
(609) 832-0000
Priya.Shetty@intelent.com
www.intelent.com