Clinical Data Model and FDA/CDISC Submissions

Mark Edwards/Gajanan Bhat, Boston Scientific Corporation, Natick, MA

ABSTRACT

Development of data model in clinical trial management environment is an ongoing process but of utmost importance for a good clinical information management. Companies are trying to come up with efficient global standards (e.g., ICH) in data modeling to work with. Global standards coupled with CDISC guidelines are also important and imminent in view of the new electronic submission guidelines of the Food and Drug Administration (FDA) for New Drug Application.

CDISC (Clinical Data Interchange Standards Consortium) partnered with the FDA to standardize on electronic submissions to improve the review/approval process of an NDA. CDISC is an open, multidisciplinary, non-profit organization committed to the development of worldwide industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trials data and metadata for medical and biopharmaceutical product development. Their mission is to lead the development of global, vendor-neutral, platform-independent standards to improve data quality and accelerate product development to our industry.

The most important factor to consider when designing the analysis and reporting data structure are the standard CDISC SDM v2.0 (Submission Data Model) data structures, (v3.0 will be out later this year), coupled with the ODM v1.1 (Operations Data Model), metadata, naming conventions, documentation, and data normalization. The SDM v3.0 enhancement will include 3 models, events, labs and efficacy data for certain therapeutic areas.

The main objective of this paper is to provide a data model following CDISC SDM for FDA submissions. This paper focuses on data model for medical device trials using Oracle Clinical and Data Warehousing database systems. Also discusses specifically the development of CDISC compatible model from OC systems, OC to SAS (CDISC) data mapping and the data flow processes. This data model defines the CDISC data structure for the Adverse Events domain **(See Figure 1, ae.xpt Model)** that includes SAS Names, and Datatype.

INTRODUCTION

Many companies in recent years have been striving to produce global and project standards. These standards are the basis for the data sets from which analyses and reports are produced. The aim is to standardize the naming and definition for SAS variables that occur in any clinical study. Food and Drug administration has brought new electronic submission standards, recently and requires the sponsors to comply with the electronic submission standards for application (NDA) as soon as possible and Device approval (PMA) soon to follow.

Today Ninety-three companies are using CDISC as a guideline to simplify their submissions based on CDISC ‘s 12 standard safety domains, (i.e., Adverse Events, etc.), to improve their data modeling, metadata, naming conventions and documentation.

The device approach is still selective in using electronic submission standards but is expected to move forward with all clinical studies in the near future.

FDA Guidelines for Electronic Submission

In order to expedite the drug approval process not compromising the approval standards and also with the advantage of the current information technology, FDA requires sponsors to provide the regulatory NDA submission in electronic

format. This provides many advantages such as global standards in submission formats, increased efficiency gain in sponsors' R&D efforts, faster and more accurate reviews and approval. However, this requires many deviations from the current submission practices, and calls for new standards in formats, data models, and metadata. One of the main sections of the NDA submission is Case Report Tabulation (CRT). This section includes Patient Profiles and SAS databases in transport files that include data definition files. The data definition file is mainly a metadata file describing data structure, data model, definitions of contents, sources, etc. This requires a detailed data model to be developed in clinical data management and analysis projects

Clinical Data Environment

Clinical Trial Management describes the general process that the clinical data pass through from data collection until data is analyzed for the statistical reports. Project standards augment the global standards already defined to include extra data sets and variables that are specific to a particular clinical project. The standard data sets and variables for the project are then used to define the analysis data sets for the actual study. Once the raw data structure is known and the analysis data structure has been specified, the data must be run through a transformation process to convert it from native database standards to the CDISC SDM standard model. This includes renaming of SAS variable names since some clinical trial data management systems (Oracle Clinical) only allow six characters for SAS names while CDISC uses up to eight characters for SAS names.

Therefore a mapping routine **(See Figure 2)** is required using the SAS PROC Format feature to rename the sponsors SAS Names to the CDISC SAS Names. This process includes a metadata table (Oracle table) that SAS will read and process to map the clinical name SAS names to the corresponding CDISC SAS names to create the SDM.

Normalization

For reasons of both data integrity and performance, it is the goal of a production database, typically, to represent each fact or data item in only one place. This is usually consistent with the typical RDBMS snowflake schemas. The data redundancy not only causes potential errors in data maintenance; it also requires added storage. Normalization is the technical name for the process that reveals and then eliminates data redundancy. The normalization in RDBMS is a key factor of a good database design. The normalized database in the relational database system gives advantages such as elimination of redundancy, ease of use, update, and modification. Each table in a normalized database has a primary key, which is a field or fields that uniquely identifies each record in the table. The tables may also have one or more fields that serve as foreign keys, which are fields with the same name as a primary key field in another table. Normalization is done by matching the tables with first, second, and then, third normal forms.

CDISC recommends that three, (Labs, Vital Signs and ECG), of the Standard Safety domains be in a Normalized format for statistical analysis as well as a De-Normalized, Vital Signs and ECG only format for online review using CDISC ODM (Operational Data Model) XML.

SAS provides the means to both read and write XML data using the XML Libname engine for processing the ODM XML to SAS data sets, and the SAS Output Delivery System (ODS) for processing SAS files to ODM XML that is compliant with the CDISC ODM standards. In addition to operating on SAS data sets and the SAS XPORT transport format, SAS provides seamless interfaces to relational databases and other data sources on various hardware platforms **(See Figure 3).**

SAS also provides SXLE XML Map technology to facilitate input processing that conforms to the CDISC XML and W3C standards. Conforming to XML standards eliminates the need for special

browsers to view data and facilitates development of more effective software applications.

Clinical Data Complexity

Analysis data structures in clinical projects do not completely conform to normalization in real life for many reasons. Main reasons are attributed to the way programmers and statisticians use the data at that stage to create the reports and the nature and the source of data. Normalization, as discussed in the previous section, makes the task of updating and modification in the original data tables easier as it eliminates duplication of information. However, the main place of update and modification of data is the original CDBS and not the analysis database. The other important reason is that statistical programming needs not the completely normalized structure in the analysis data structure. This will pose an extra overhead of combining data sets to create meaningful and workable data sets for the production of Tables and Listings. Thus, it leads to normalization and de-normalization at the same time. Also, it is better to keep some basic information such as visit number, visit date, etc. in every analysis data set.

CLINICAL DATA MODEL

The Role of Metadata

Before beginning data modeling from the data warehousing perspective, an extremely thorough understanding of the user's requirements is necessary. This includes a thorough knowledge and assessment of metadata. Aside from the obvious concerns of data cleanliness, integrity, and integration, the understanding of data value cardinality and distribution pattern are of utmost importance in determining various keys and indexing of final data models.

Data Model Overview

One of the most important decisions affecting the modeling efficiency is the schema with which you choose to model your data. This also depends on the level of granularity (level of details) that is required for the project. The user requirements may vary depending on the type of user. In the clinical data analysis projects, from the data warehousing perspective, the summarized or granular structure is preferable, where as less normalized data structure is preferable from the statistical analysis perspective to reduce the overhead involved in creating reports.

Two subject-based data models are of widely use. They are Snowflake Schema, most widely used for the operational data and the Star Schema that is always used for Data Warehousing models. The differences between the two models are the Snowflake Schema is a third normal form model or a normalized data model while the Star Schema is a de-normalized data model.

In a Star schema, the relational tables are used to model the subject of interest known as a Fact table. This schema utilizes explicit modeling of relationships. Because it holds large volumes of numeric factual data about the subject, the Fact table (mostly Keyvar data) is typically the largest table. The Fact tables (i.e., Demography, Vital Signs, Adverse Events, LAB, MEDS, etc) are surrounded by many "Dimension" or coding tables (i.e., Date, Time, Study, Country, etc) that are used as lookups or dictionary data that supports the subject base data of information.

Table 1. Clinical Database Attributes

Column Name	Description
Data set name	Name of the data set
Description/ Label	Description of the data set regarding its purpose, contents, and key information/Label given to the data set
Data set Program	Name of the program
SAS Engine	Version of SAS Engine used

	to create the data set
Observations	Number of observations/records in the data set
Variables	Number of variables/columns in the data set
Index	Names of the indexes defined in the data set
Compression	Yes/No
Protection	Yes/No
Primary key	Unique identifier variable(s) used to distinguish the records
SAS Names	CDISC SAS Names for each of the variables.
Variables	List of the all the variables
Sort by	Name of the variable(s) that the data set is sorted by
Type	Data type: Numeric/Character/Date/use defined
Length	Length of the variables (8 for numeric)
Format	The formats associated with the variables
Informat	The informat associated with the variables to read in
Labels	Labels associated with the variables.

Data Model Attributes/Roles

There are several considerations when selecting primary keys. They should uniquely identify individual patients and records and should be as short as possible. The primary keys must be identical across each of the patient level data files i.e. the variables have the same length and type. Most used variables, as primary keys in the clinical databases are Protocol and Patient ID.

Defining the Data Structure

Consider the following Vital Signs Data set example in Table 2.

Table 2. Vertical vs. Horizontal Data Structure

Horizonal Data Structure

Visit	Systolic	Diastolic	Weight
1	128	84	161
2	125	82	158
3	121	81	159

Vertical Data Structure

Visit	Parameter	Value
1	Systolic	128
1	Diastolic	84
1	Weight	161
2	Systolic	125
2	Diastolic	82
2	Weight	158
3	Systolic	121
3	Diastolic	81
3	Weight	159

Invariably the question arises, “Why not just use de-normalized or rather more robust data sets with all the data in one observation instead of Fact and Dimension tables (data sets)?” This situation most often creates far too much redundant data for project Database in the cases where in considerably large storage is required. The models using Star Schema and such represent a viable option to allow reasonably good query performance and elimination of some redundant data.

The analysis data structure must be optimized to facilitate reporting and analysis process. The most important factor to consider when designing the analysis and reporting data structure is data normalization. The efficiency and ease of extracting the data from the database and use them for reporting are very important. Depending on the nature of data, de-normalized data structure is better suited for some data sets such as Vital Signs where as normalized data

structure is better suited for some data sets such as labs. For example, Single Vital Signs listings and tables include all parameters where as there will be separate listings/tables for each lab or efficacy parameters.

CONCLUSION

This paper describes in brief, the transformation from a native database system OC (Oracle Clinical as an example) to CDISC compatible data model following SDM v2.0 Adverse Events domain was used for illustration purposes. The process of OC to SAS mapping was explained. This paper provides an alternative approach to transform or comply with CDISC SDM for data residing on OC system without changing the sponsor's database and objects standards. Also, the paper gives a brief account of the FDA/CDISC guidelines in terms of data model for new electronic submissions.

CONTACT INFORMATION

Mark Edwards
Boston Scientific Corporation
1 Boston Scientific Place
Natick, MA 01760
Phone: 508.650.8620
E-mail: edwardsm@bsci.com

Gajanan Bhat
Boston Scientific Corporation
1 Boston Scientific Place
Natick, MA 01760
Phone: 508.650.5254
E-mail: Gajanan.Bhat@bsci.com

CDISC – Adverse Events

AE.xpt, Adverse Events, Version 2.0, October 22, 2002, One Record per subject per Adverse Event, Date: Monday, January 06, 2003

CDISC Variable Name	SAS Variable Name	CDISC Variable Label	CDISC Type	Decodes/Format	CDISC Origin	CDISC Notes	CDISC Core Variable
AEACTSDY	SAS Calculation	Actual Study Day of Start of Event	Number		Derived	Algorithm for calculations must be relative to the sponsor defined DMREFDT (and DMREFTM, if appropriate) variable in Demographics. This formula should be consistent across the submission.	Y
AEACTEDY	SAS Calculation	Actual Study Day of End of Event	Number		Derived	Algorithm for calculations must be relative to the sponsor defined DMREFDT (and DMREFTM, if appropriate) variable in Demographics. This formula should be consistent across the submission.	Y
AESEQ	REPEATSN Given	Sequence Number	Number		Adverse Events CRF Page	Can be used as an optional identifier to ensure uniqueness within data set.	N
AETERM	ADVEVT	Reported Term	Char		Adverse Events CRF Page	Verbatim text.	Y
	S	Modified Reported Term	Char		Adverse Events CRF Page	If AETERM is modified for coding, AEMODIFY should be provided. Procedure defined by sponsor.	N
AEDECOD	TMS	Dictionary Term	Char		Adverse Events CRF Page	1. Dictionary-derived text description of AETERM or AEMODIFY. The sponsor should specify the dictionary name and version in the Sponsor Comments column. 2. If AETERM or AEMODIFY are not coded, then value for AETERM should appear here.	Y
AEBODSYS	TMS	Body System/Organ Class	Char		Adverse Events CRF Page		Y
AESTDT	STRTDT	Start Date of Event	Number	ISO 8601 YYYY-MM-DD	Adverse Events CRF Page		Y

Figure 1

Figure 1

CDISC – Adverse Events

AE.xpt, Adverse Events, Version 2.0, October 22, 2002, One Record per subject per Adverse Event, Date: Monday, January 06, 2003

CDISC Variable Name	SAS Variable Name	CDISC Variable Label	CDISC Type	Decodes/Format	CDISC Origin	CDISC Notes	CDISC Core Variable
AEENDT	STOPDT	End Date of Event	Number	ISO 8601 YYYY-MM-DD	Adverse Events CRF Page		Y
AESTTM	Exclude	Start Time of Event, 24 hr Clock	Number	ISO 8601 HH:MM	Adverse Events CRF Page		N
AEENTM	Exclude	End Time of Event, 24 hr Clock	Number	ISO 8601 HH:MM	Adverse Events CRF Page		N
AEDUR	Exclude	Duration of Event (in Days)	Number		Derived		N
AEDURU	Exclude	Units of Time for this Event	Char		Sponsor Defined		N
AESER	SAEYN	Seriousness Criteria	Char	Y,N	Adverse Events CRF Page		Y
AESEV	SEVER	Severity/Intensity of Event	Char		Adverse Events CRF Page		Y
SAEDTH	SAE	Results in Death	Char	Y,N	Adverse Events CRF Page		Y
SAELIFE	SAE	Is Life-Threatening	Char	Y,N	Adverse Events CRF Page		Y
SAEDISAB	SAE	Permanent/Serious/Disability/Incapacity	Char	Y,N	Adverse Events CRF Page		Y
SAEHOSP	SAE	Requires/Prolongs Hospitalization	Char	Y,N	Adverse Events CRF Page		Y
SAECAN	Exclude	Involves Cancer, NA	Char	Y,N	Adverse Events CRF Page		Not considered to be core since this SAE category is not required by E2B.
SAEOD	Exclude	Occurred with Overdose	Char	Y,N	Adverse Events CRF Page		Not considered to be core since this SAE category is not required by E2B.
SAECONG	SAE	Congenital Anomaly/Birth Defect	Char	Y,N	Adverse Events CRF Page		Y
SAEOTH	SAE	Other Medically Important	Char	Y,N	Adverse Events CRF Page	Additional seriousness criteria can be defined by sponsor if needed.	Y

Figure 1

CDISC – Adverse Events

AE.xpt, Adverse Events, Version 2.0, October 22, 2002, One Record per subject per Adverse Event, Date: Monday, January 06, 2003

CDISC Variable Name	SAS Variable Name	CDISC Variable Label	CDISC Type	Decodes/Format	CDISC Origin	CDISC Notes	CDISC Core Variable
AEACTTRT	ACTION	Action Taken with Study Treatment	Char	E2B Codes: 1=Drug Withdrawn 2=Dose Reduced 3=Dose Increased 4=Dose Not Changed 5=Unknown 6=Not applicable	Adverse Events CRF Page		Y
AECONTRT	Exclude	Concomitant/Additional Treatment Given	Char		Adverse Events CRF Page		N
AEACTOTH	ACTASP	Other Action Taken, Specify	Char		Adverse Events CRF Page	Algorithm is sponsor defined.	N
AEOUT	OUTCM	Outcome of Event	Char	E2B Codes: 1=Recovered/ Resolved 2=Recovering/Resolving 3=Not Recovered/ Not Resolved 4=Recovered/ Resolved with Sequel 5=Fatal 6=Unknown	Adverse Events CRF Page		Y
AEREL	RELDEV	Causality (Relationship to Treatment)	Char		Adverse Events CRF Page		Y
AECOM	COMMTS	Comment	Char		Adverse Events CRF Page		N
AEONGO	ONGO	Ongoing Adverse Event	Char	Y,N			Y
AETRTEM	Exclude	Treatment Emergent	Char	Y,N			N

Figure 1

CDISC SAS Mapping

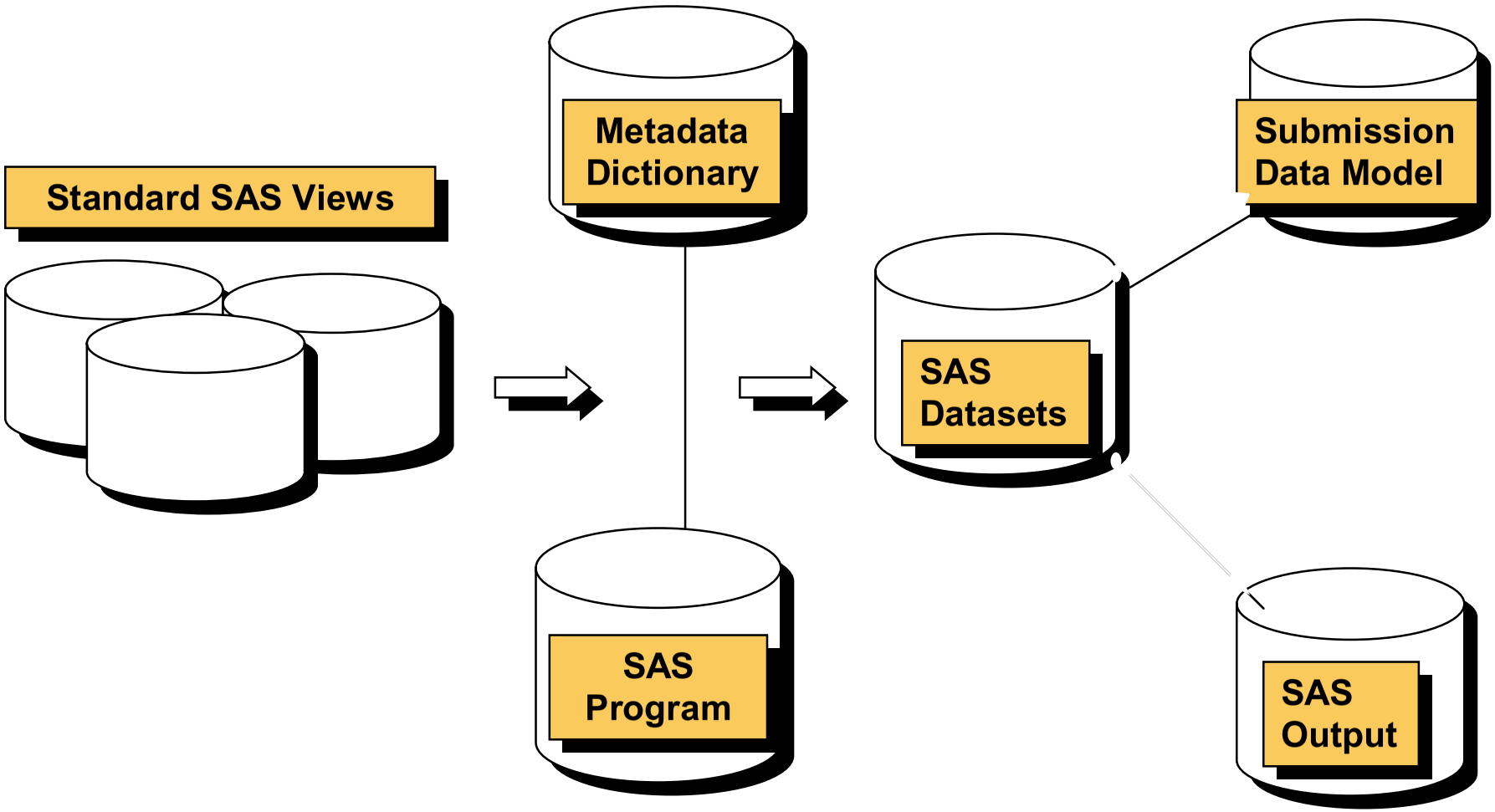

Figure 2

ODM XML and SAS

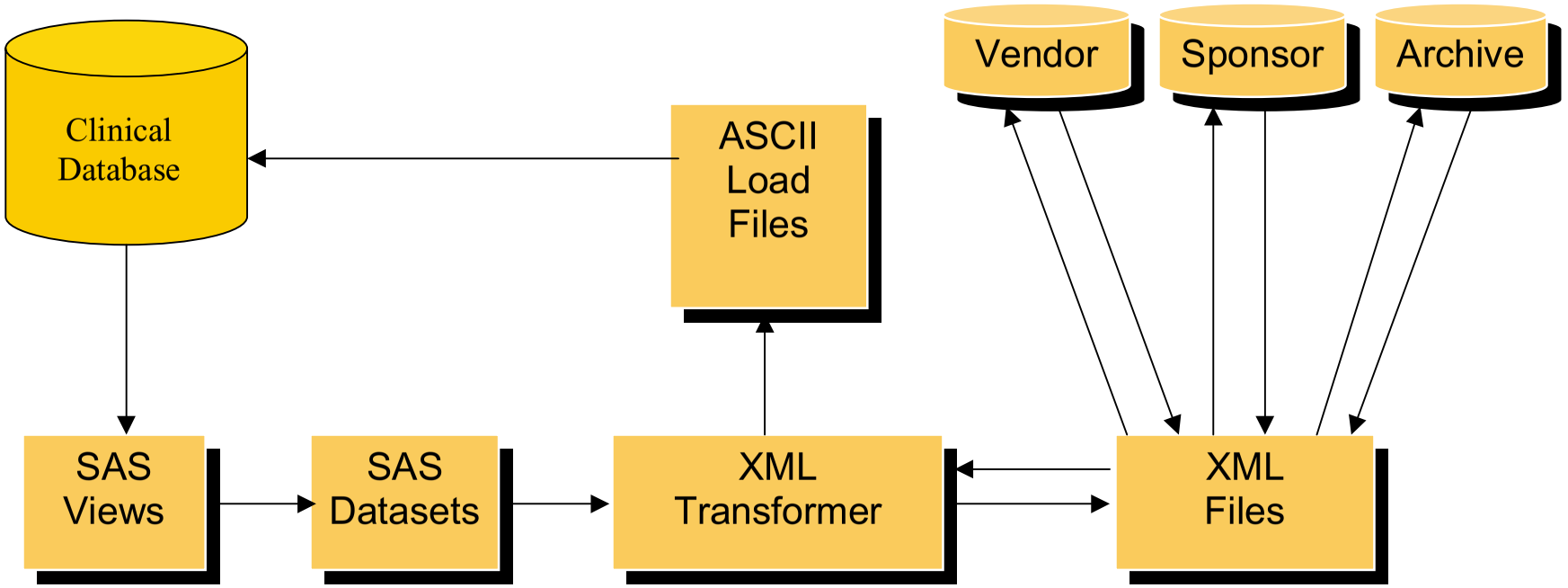

Figure 3