Paper 238-28

METADATA APPLICATION ON CLINICAL TRIAL DATA IN DRUG DEVELOPMENT

Julia Zhang, GlaxoSmithKline, Collegeville, PA
David Chen, EDP Contract Services, Bala Cynwyd, PA
Tor-Lai Wong, PSCI, Wilmington, DE

ABSTRACT

As we move forward to an electronic environment, information proliferates rapidly. How do we efficiently manage this voluminous quantity of information and use this data appropriately? With metadata knowledge, we learned that we need to develop an effective and efficient metadata strategy to prevent us from suffering the chaotic situation associated with proliferation of mass amounts of information. Using metadata appropriately in the drug development environment can vastly increase work efficiency and reduce the probability of mistakes. In order to take advantage of the metadata methodology, we developed an application using Visual Basic® for Applications (VBA) calling upon the Windows® Application Programming Interface (API) in Microsoft® Excel in conjunction with the SAS® system to efficiently apply metadata on clinical trial data.

INTRODUCTION

Metadata is descriptive information about an object or resource, whether it is physical or electronic. The simplest definition of metadata is "structured data about data". It allows a piece of information to be discovered by potential users, assessed for its usefulness, used or analyzed in an appropriate context, and protected and deleted as appropriate. While metadata itself is relatively new, the underlying concepts behind metadata have been in use for as long as collections of information have been organized. For example, library card catalogs represent a well-established type of metadata that has served as a collection of management and resource discovery tools for decades.

As we step into the E-era, information sources seem to be growing, and data information spreads much faster with each passing day. Proper information management is a necessity to ensure that important research is not lost or duplicated unnecessarily. Willing or not, you will be caught up in the proliferation of information. Applying the metadata concept to your information will allow you, and those you share information with, to search across multiple files for the exact piece of information you need easily. Without the proper information management practices and tools, the quantity of information, diversity of information sources, and multiplicity of information types found on your desktop can lead to a needle-in-a-haystack scenario!

As information moves from an individual arena to one where it is shared, applying metadata principles becomes more and more important. The more critical the information is to the business, the more metadata is needed to ensure its retrievability.

METADATA APPLIED ON CLINICAL TRIAL DATA

Metadata can be found almost everywhere. Since metadata describes a piece of information which could be it's content, format, intended use, origin, derivation, etc., we are able to facilitate information flow, improve retrievals, help discovery and provide context by applying metadata principles to information. During the data analysis process in a drug development environment, we are faced with the sometimes-overwhelming task of efficiently and effectively managing different kinds of data. The examples of those data are clinical trial patient data which are collected on Case Report Forms (CRF), definition data for manipulating patient information, SAS data sets for statistical analysis, and the table of contents for statistical outputs. We realize that as long as there is data, metadata will be needed. We apply the metadata principles all the time in the process of drug development.

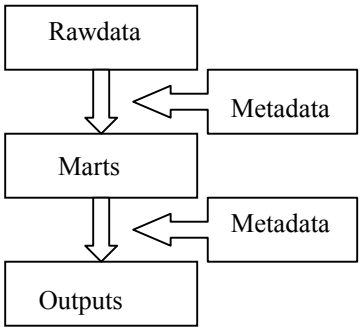

Figure 1. Workflow of statistical data analysis using metadata.

Figure 1 illustrates workflow for the statistical data analysis process in the drug development environment. Rawdata here is defined as the patient data collected on CRFs from the investigative sites – for example, ACE (adverse clinical events), ECG (Electrocardiogram), LAB (laboratory), and VITAL (Vital Signs) data. After raw data has been stored into the clinical trial database, the second step is data manipulation. Data marts (derived SAS data sets) are generated based on specific rules and criteria that are defined by metadata. We call the repository for this metadata the Data Definition Table (DDT). The final step in the workflow is to produce SAS outputs using the data marts. In order to create SAS outputs efficiently and to reduce spelling mistakes for titles and footnotes to a minimum, another metadata entity called the titles driver file is created. Other metadata e.g., clinical concern flagging criteria for LAB or VITAL data, may also be needed during the statistical reporting process or New Drug Application (NDA) submission. These metadata are defined centrally as the DDT and titles driver file. The three primary examples of our metadata are defined as follows:

DATA DEFINITION TABLE (DDT)

DDT is a collection of data manipulation guidelines, which defines the structure, content, and derivation of the derived data marts. It contains drug name, protocol number, table (we refer to data sets as tables in the DDT) name, table label, variable name, variable label, variable type, variable length, variable format, variable derivation, and variable description. You can also add other fields based on your needs. The Clinical Data Interchange Standards Consortium (CDISC) has proposed some pharmaceutical industry standards for the structure and content of data domains used to store clinical trial study data. Our DDT model is referenced from the model defined by CDISC. By following the guidelines in the DDT, programmer analysts can derive the SAS data marts using rules for mapping derivations described. Utilizing the DDT a quality control (QC) programmer can efficiently quality assure (QA) the derived SAS data marts.

TITLES DRIVER FILE

We use appropriate statistical methods to perform data analysis and processing, then produce the tables and listings to interpret the data. To efficiently generate the outputs, a titles driver file containing information of the output titles and footnotes is created. The titles file is the link from the physical output to the

study's reporting analysis plan. This driver file keeps all the titles and footnotes in a centralized location and makes any update, modification, and QC easier and faster.

TABLE OF CONTENT (TOC)

Generally, we provide a Table of Contents (TOC) to our clinical customers as a review aid. The TOC is a supplemental document for the clinical group to facilitate review of the statistical output. It contains a list of item numbers and titles for the output delivered. It can also be used to provide comments and feedback if desired. It can easily be produced from the titles driver file using the VBA metadata application.

METADATA APPLICATION

Metadata is structured data about data, and it is centralized data management. It is a tool to control and maintain data entities. Metadata consistently and effectively describes data and reduces the probability of the introduction of errors in the data framework by defining the content and structure of the target data. The uses of metadata in the pharmaceutical industry are growing, and the industry is seeking to develop supporting standards through independent organizations such as CDISC. CDISC's mission is the development of industry standards to support the acquisition, exchange, submission, and archiving of clinical trial data. Inspired by the data model from CDISC and based on the need for an effective means to control the content and structure of data marts, manage the data contained within the DDTs (as well as other study reporting related resources), and maintain the consistency of these elements across studies during the data analysis and processing phases, the idea for a tool to integrate and facilitate these related components was conceived.

Among the requirements identified for efficiently using the metadata were:

- User friendliness
- Simple, intuitive interface
- Seamless link between SAS and this application
- Ability to map metadata from source to target platform
- Implement with standard computing resources

The Metadata Application is implemented as a VBA add-in module (.xla) for use with Microsoft Excel. Other implemented features not already available in Excel are provided through the Windows API (Application Programming Interface) and the SAS System. The Metadata Application tool provides three primary services- data manipulation and management via Excel, interface and interaction with the user via the Windows API routines, and conversion of data found in the DDT into SAS data sets on our target platform via the SAS System. The VBA add-in allows for the integration of these services and centralizes them.

Microsoft Excel is a feature rich application for manipulating tabular data. From integrated spell checking, to sorting, filtering, finding, replacing, importing and exporting data capabilities- all of these built-in features contribute to the ease of use, power, and flexibility of the overall metadata application. The general popularity of Excel in the computing community and the fact that most people have had some exposure to Excel means that they can manage and maintain these metadata more effectively and quickly. As such, Microsoft Excel was chosen to create and maintain this metadata and was determined to be the logical starting point from which to build upon.

Figure 2 is an example of an Excel spreadsheet that contains the data definition table (DDT). This DDT contains all the information necessary to define the content and structure of the derived SAS data marts for a single study. While the specification of the Food and Drug Administration (FDA) compliant DDT actually requires that distinct data domains reside on separate worksheets within the same workbook, a single worksheet is used to hold metadata specifics from more than 50 different data marts. The main purpose of using one worksheet to aggregate all the mart information is that it is easier to maintain and update information across data domains consistently. Another reason for aggregating all the domains is so that this metadata can be converted into single SAS data set in the reporting environment. This scheme allows for the content and structure of each data mart during data mart generation to be driven by a single metadata SAS data set. Additionally, this circumvents the extra processing necessary to collect and combine the metadata for the data domains if they were to reside on distinct worksheets- prior to creating the data mart definition data set on the reporting platform.

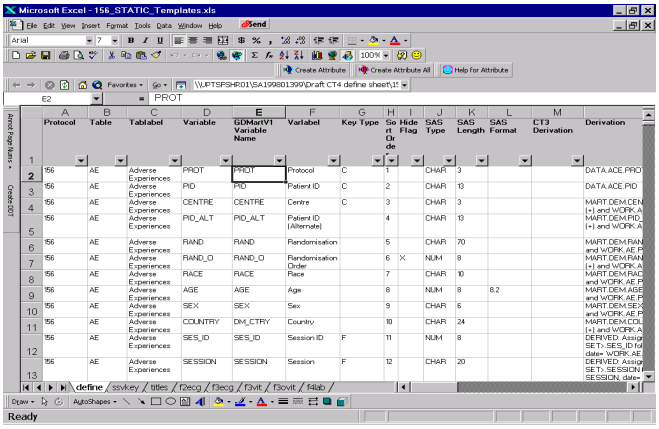

Figure 2. An example of Excel spread sheet for DDT. It also contains different data sheets for other metadata information.

Taking into account the requirements, the goal was to develop an application, making use of Excel and extending it's capabilities by adding customized commands and specialized features. The routines available in the Windows API provide features/services not otherwise available to the Metadata Application i.e., in Excel. The API routines handle most of the specialized interaction with the user through the VBA add-in module's interface. Among these services are: interface navigation, file management, creating starter template metadata worksheets, creating the TOC deliverable on the desktop from the titles driver file worksheet, and creation of the FDA compliant DDT workbook for a NDA submission effort. In addition, the API handles one of the primary services to the metadata application, that of uploading and synchronizing the titles driver file on the desktop with the corresponding file used in output generation on the reporting platform through File Transfer Protocol (FTP).

The final service/feature provided by the Metadata Application is the synchronization of the DDT on the desktop and the data mart generation DDT on the target reporting platform. This service is provided by the SAS System (Base SAS, SAS CONNECT). Specifically, SAS on the desktop is invoked in the background to interface with Excel, utilizing DDE (Dynamic Data Exchange), and acts as an OLE (Object Linking and Embedding) automation server for the VBA front-end i.e., OLE client. This service is the mechanism through which the DDT metadata information is uploaded onto the target reporting platform. From there, the uploaded DDT, in the form of a SAS data set, is used in the construction and definition of the attributes of the derived SAS data sets used in reporting.

These services are bundled into the Metadata Application- the glue being the VBA add-in, which provides a customized front-end interface for ease of navigation. Invocation of the front-end can only take place after the VBA add-in has been loaded into Excel, either directly or by placing the add-in within Excel's XLSTART folder on your desktop. The add-in inserts a drop down

menu command under the TOOLS menu item within Excel for easy access.

Figure 3 displays a general layout of the metadata application. In the upper panel, you can select your project or clintrial protocol, study number and its indication. The Active Workbook box indicates the name of the selected workbook.

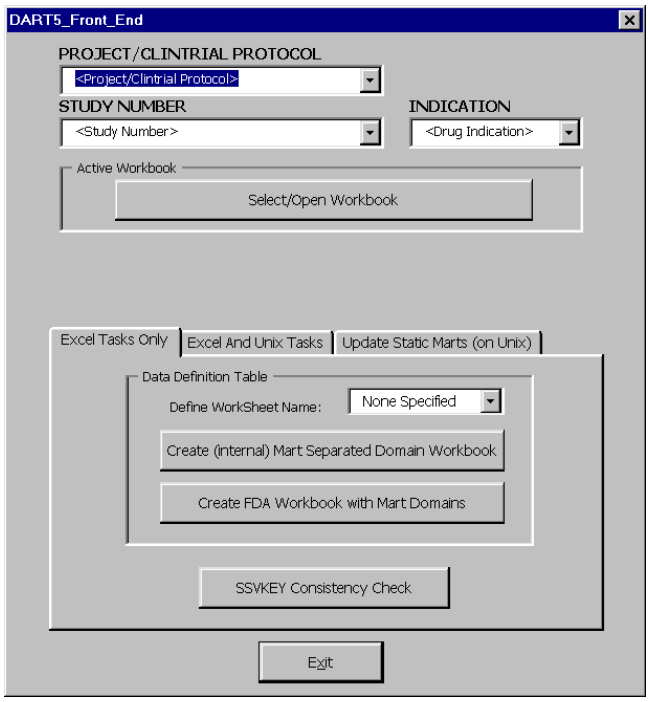

Figure 3. The display of our developed metadata application.

There are three major functional areas within this application:

Excel Tasks Only tab performs the following functions:

1. For documentation purposes, you can create another Excel workbook with one mart per worksheet. This makes reading easier.
2. Create FDA version of Excel workbook. The FDA has its own standards or preferences; for example, workbook contains one table/data domain per sheet, and contains only fields required by the regulatory agency, such as variable name, variable type, variable length, and description. This application can read the original Excel DDT worksheet, which has more information than needed and create a version of Excel workbook that only contains the fields FDA requires in a preferred format.

Excel and UNIX Tasks box performs the following functions using the titles file:

1. FTP titles driver file to UNIX using the titles worksheet.
2. Create titles driver file or other study-specific marts such as clinical concern flagging criteria data sets using the standard metadata templates.

Update Marts on UNIX box performs the following functions:

1. Upload all data domains DDT metadata Excel worksheet transformed into SAS data set onto UNIX.
2. Upload other metadata in SAS data set form onto UNIX system.

Our target reporting environment is a UNIX system; as such, our front-end refers to UNIX, but is extensible to other computing systems.

The Metadata Application is the aggregate collection of the VBA add-in module, Microsoft Excel, the Windows API routines, and the SAS System. The VBA front-end is the glue, while the API routines and the SAS System extend the functionality of the tool, it is Excel that is the primary reason for the ease and flexibility possible in the management and maintenance of the metadata. This metadata application is convenient and reduces a lot of the redundant work such as in maintaining the DDT and creating data marts of like structures for similar studies.

CONCLUSION

The benefits of using metadata are countless and the impact is enormous. This developed metadata application provides us with a valuable tool to use metadata more efficiently and maintain metadata more easily. It increases work efficiency and reduces the probability of errors. Using the metadata application, we are able to streamline the whole process of constructing SAS data marts as defined in metadata in a consistent manner. It also allows us to produce the FDA version of data documentation, generates other specific metadata, and eliminates the need for duplication of effort. This application simplifies complicated work, which would otherwise require more steps and more people to complete, thereby saving us time and resources. It makes quality control (QC) and data validation easier. Using metadata is a definite trend in the E-time.

REFERENCES

(1) CDISC, Submissions Data Model (SDM) Version 2.0, http://www.cdisc.org/models/sds/v2.0.

(2) CDISC, Analysis Dataset Model, http://www.cdisc.org/models/adam/ADaM_Guidelines_V1.pdf

(3) CDISC, Standards, http://www.cdisc.org/standards/index.

TRADEMARKS

CONTACT INFORMATION

Julia Zhang
GlaxoSmithKline
1250 S. Collegeville Road, Collegeville, PA 19426-0989
(610) 917-6914 (W)
julia_z_zhang@gsk.com

David Chen
EDP Contract Services
Bala Cynwyd, PA 19004
(610) 917-6062 (W)
david_chen-1@gsk.com

Tor-Lai Wong
PSCI Inc.
Wilmington, DE 19810
(610) 917-5096 (W)
tor-lai_2_wong@gsk.com