It's Easy If You Know How: Importing, Processing, and Exporting CDISC XML with SAS®

Michael C. Palmer, Zurich Biostatistics, Inc., Morristown, NJ

ABSTRACT

The Clinical Data Interchange Standards Consortium (CDISC) has published an XML-based clinical data standard that will make it easier for a pharmaceutical company or other study sponsor to develop a single, low maintenance, vendor-neutral gateway for clinical trials data, including central lab results and FDA-compliant data archiving. The tutorial will cover the basics of data-centric XML, the structural and semantic requirements of CDISC's Operational Data Model (ODM) standard, and techniques for importing, processing, and exporting clinical data in SAS using the ODM. The tutorial is needed because SAS users accustomed to flat or relational files have found it challenging to import, process, and export CDISC XML. Trouble has come from two characteristics of XML. First, it is a text stream, and, second, it is hierarchical in structure. SAS, on the other hand, is oriented towards data in fields and flat or relational file structures. CDISC is a consortium of some 60 leading pharmaceutical companies and vendors, with a formal liaison to the FDA, pledged to the development of non-proprietary, open standards to streamline the collection, analysis, and review of clinical trials data intended for regulatory submissions.

INTRODUCTION

The Clinical Data Interchange Standards Consortium (CDISC) began in 1998 as a special interest group within the Drug Information Association (DIA) dedicated to the development of vendor-neutral, platform-independent data standards for clinical trials. CDISC incorporated as an independent entity in June 2000 and shortly thereafter, in November 2000, released initial versions of two complementary data models. The Operational Data Model (ODM) is an XML-based clinical data model designed for data transfer and archiving. The Submissions Domain Model (SDM) is a SAS dataset-based model designed for regulatory needs. Discussions have taken place in CDISC about migrating the SDM to XML.

CDISC is a consortium of US and European companies active in the pharma industry and the ODM and SDM are the beneficiaries of this broad-based, largely volunteer effort to develop an industry consensus for clinical data models. FDA has supported CDISC and officially acknowledged the role of the SDM in electronic submissions (eSubs). FDA has taken a keen interest in the development of the ODM and, as of February 2002, is discussing the role that it will have in submissions.

OPERATIONAL DATA MODEL

At the topmost level, an ODM instance contains metadata and data. Conceptually, ODM metadata is an annotated case report form (CRF) formatted as XML. Like an annotated CRF, ODM metadata describes the study events, such as visits, and types of data collected, such as adverse events, physical exams, or efficacy. The metadata, like an annotated CRF, also describes the fields on the CRF with their attributes such as numeric or character, length, codelists, and so on. The data part of an ODM instance contains the actual clinical data, formatted as XML, and laid out by subject, study event, type of data, and field, in other words, laid out precisely the way clinical data are arrayed in a paper CRF.

Just as a CRF is designed to transport data, from site to sponsor and to archive data in a file cabinet, the XML version of a CRF, that is the ODM, is designed to transport data and to archive data. The analogy between CRF and THE ODM extends even further. The hierarchical, subject-study event structure of a CRF makes it a clumsy database for queries. Likewise, the hierarchical subject-study event structure of an ODM instance makes it a clumsy database, even though it is a good data transport and archiving format.

SUBMISSIONS DOMAIN MODEL

In contrast to the hierarchical, subject-study event structure of the ODM, the SDM is organized around 12 clinical domains: demography, adverse events, concomitant medications, ECG, drug exposure, chemistry labs, hematology labs, urinalysis labs, medical history, physical examinations, vital signs and subject disposition. At the present time, the SDM does not include efficacy data. SDM data exist, at the present time, in SAS version 5 transport datasets. CDISC has discussed migrating the SDM to XML and, anticipating that, the SDM will be discussed in this tutorial.

The SDM contains metadata, but not in machine-readable format. SDM metadata exist in PDF files. If the SDM is migrated to XML, the metadata will become machine-readable.

A CLINICAL DATA GATEWAY

Sponsors of clinical trials always have a gateway for clinical data. That is, they have a way to import study data from sites, partners, CROs, central labs, and other data sources and they have a way to export clinical data to FDA, partners, subsidiaries, and data archives. It is not unusual for these gateways to require manual revision, tinkering, and tweaking for every new source of data, whether it is a new central lab, CRO, or other data source. Not uncommonly, tinkering and tweaking are required every time personnel change at a data source because different people interpret even written specifications differently. In other words, clinical data gateways are often not vendor-neutral nor low maintenance.

The CDISC standards have the potential to change this situation.

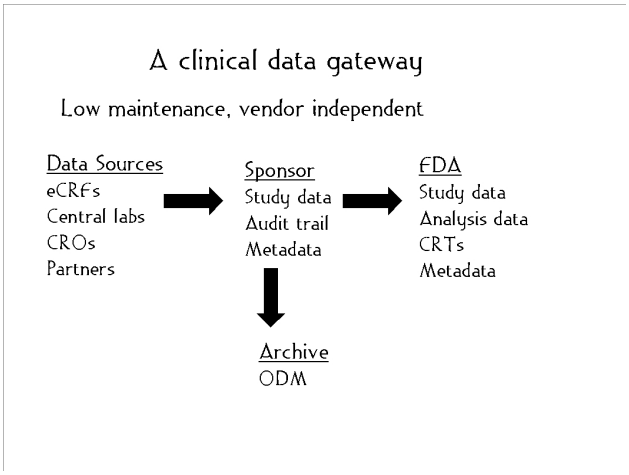

Figure 1. CDISC standards can make clinical data gateways low maintenance and vendor independent.

LOW MAINTENANCE

For exporters of clinical data, such as clinical labs and CROs, the ODM is a consistent, well-documented, machine-readable target to export to. Well-documented consistency has at least two very desirable consequences. First, it makes the writing of generic, importer-independent tools feasible, and, second, it enables a very high level of automation so that manual tinkering and tweaking are not necessary for individual instances of the ODM, and, hopefully, not even for different clients or projects.

On the pharma side, that is, the ODM import side, consistency and documentation have the same desirable consequences: generic tools and automation. In addition, machine-readable metadata is as important as consistency and documentation. Each ODM instance includes ***all*** the information, as machine-readable metadata, needed to import the clinical data into a clinical database. This has been demonstrated by numerous vendors at CDISC-sponsored Connectathons in July 2001 and October 2001.

The SDM enjoys some but, as of this writing, not all of the technical sophistication of the ODM. The SDM is a consistent, well-documented machine-readable target for export and source for import. But, SDM version 2.0 does not have machine-readable metadata. The SDM metadata facilitates generic import and export tools but the lack of machine-readable metadata limits the degree of automation that is possible. It is anticipated that the SDM will migrate to XML and will then include machine-readable metadata.

VENDOR-NEUTRAL

The ODM is an XML vocabulary. XML is just plain text so The ODM is just plain text that can be read in any text reader, such as Notepad on Windows PCs. There is no need to buy proprietary software to read an ODM instance. CDISC freely publishes the precise features of the ODM vocabulary, or DTD in XML jargon, for all on its web site. Anyone can use it.

CDISC does allow proprietary vendor extensions to the ODM DTD but with significant restrictions. An ODM instance extended by a vendor must still be valid ODM when the vendor extension is stripped out. In other words, the extended ODM will still make sense to ODM-savvy applications that know nothing of the extension. In addition, the DTD for a vendor extension must be made freely available to CDISC and, through CDISC, to any ODM user.

Like, THE ODM, the SDM standard is freely available from CDISC. FDA acknowledged SDM version 2.0 in December 2001, in effect recognizing it's vendor neutrality. The specification for the electronic format for SDM, SAS version 5 datasets, is freely available from the SAS Institute, having been put into the public domain for use with the SDM and eSubs. The SDM metadata exists in PDF, a proprietary format owned by Adobe Systems, Inc., but the Adobe Acrobat® Reader for PDF is freely available from Adobe. If the SDM migrates to XML, it will enjoy the same vendor-neutrality as the ODM does.

GATEWAY SUMMARY

The one word summary of the clinical data gateway is **consistency**. The ODM and SDM provide a consistent way to import and export clinical data no matter where they come from or where they're going in the pharma world. Consistency, coupled with documentation, can translate to low maintenance, automated implementations. The ODM electronic CRF is a well-documented means for transporting and archiving clinical data. The SDM is a well-documented means for submitting non-efficacy clinical data to FDA.

ODM

The ODM relies on the CRF paradigm and a few words about that paradigm are a useful introduction to the ODM data model. In the ODM paradigm, a CRF belongs to a clinical study protocol. The protocol, and the CRF collecting data for it, consist of one or more study events. Study events can be scheduled visits or other clinical events like concomitant medications, for instance. For a given study event, data are collected on forms. Forms consist of groups of individual data items. These item groups are often called data panels. The items are data fields in an item group. So, a given item is nested in an item group, form, and study event.

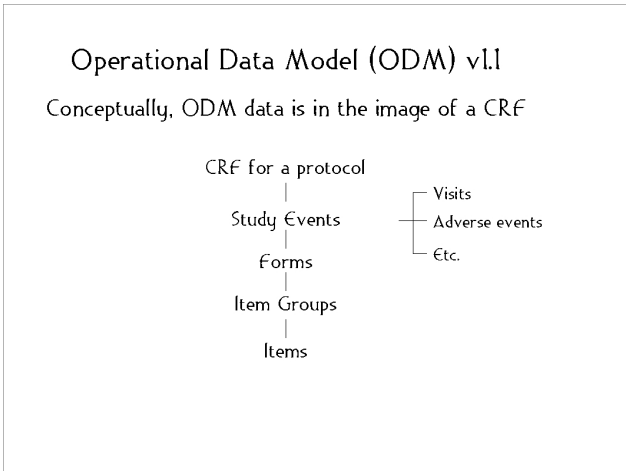

Figure 2. the ODM provides an XML-formatted case report form.

In an additional parallel to CRFs, study event and its nested data are organized by study subject. An ODM instance for a study is like an XML file cabinet. Each subject has one drawer in this file cabinet and in that drawer there's one file folder for each visit. The individual CRF pages are in the folders.

For a CRF, data can change, and the audit trail has to be maintained. The ODM provides for an audit trail parallel to the original data. Signatures are required for the original data and for changes to the data and the ODM provides for 21 CFR Part 11-compliant electronic signatures.

CRFs collect subject data but to enter that data into a database, one needs an annotated CRF that describes field attributes such as data types, lengths, data coding, mapping to datasets, and other database specifications. In other words, one needs data about the data, that is, metadata. This metadata, in XML, is an integral and essential part of the ODM. The metadata in the ODM specifies completely how to import clinical data from the ODM's XML format to a clinical database.

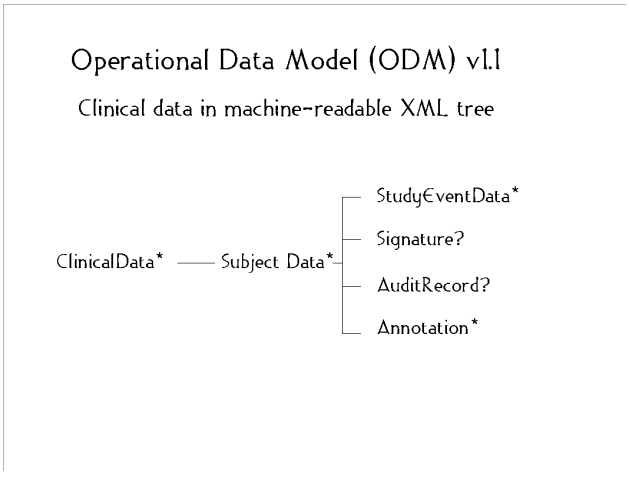

Figure 3. Clinical data exist in the ODM with electronic signatures, audit trails, and comments.

Subject data are not the only kind of data collected in a clinical study. Clinical lab normals, for instance, are part of a clinical study database but are not study subject data. The ODM includes a reference data section for non-subject data.

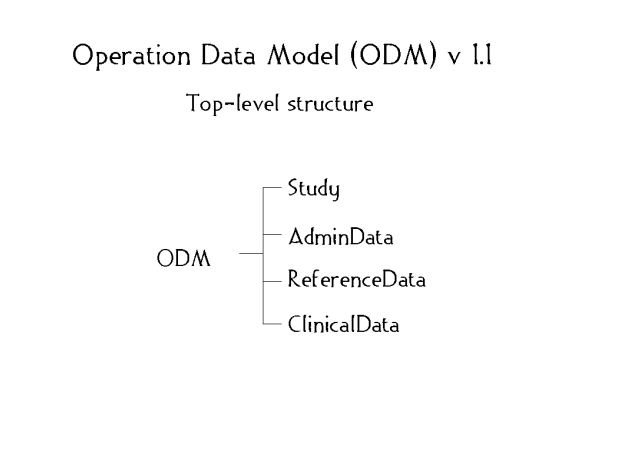

Figure 4. The ODM is a single data source for all clinical study information.

The ODM also includes a section for administrative data. The CDISC documentation details what should be in each section of an ODM instance.

SDM

The ODM is an abstract CRF that models a clinical study as subjects undergoing study events and accommodates whatever data are collected. The SDM takes a less abstract, less operational point of view and explicitly models non-efficacy clinical data in clinical studies. The SDM categorizes non-efficacy clinical data into 12 clinical domains: demography, adverse events, concomitant medications, ECG, drug exposure, chemistry labs, hematology labs, urinalysis labs, medical history, physical exam, vital signs, and subject disposition.

Each domain has the same, consistent record structure: key variables, selection variables, review variables, and support variables. Key variables uniquely identify records in a dataset. Selection variables are used to subset data for reporting. Review variables are source or derived outcomes used in clinical and statistical analysis. Support variables are anything else of relevance. Every record will have key and selection variables and generally one or more variables with other roles. A single variable can have multiple roles, key and selection, for instance. In addition, CDISC or a sponsor may extend this list of roles.

Unlike the ODM which has machine-readable metadata, SDM metadata exists in PDF files and so is not readily machine-readable. If the SDM migrates to XML, it's metadata will also and will be machine-readable.

Dataset metadata describes basic SAS datasets attributes such as dataset name, description, and file location and it also includes SDM-specific attributes. The SDM-specific attributes are key variables, purpose, and dataset structure with respect to the number of records per subject and study event. Dataset purpose can be "CRT" for case report tabulation or "Analysis" if the dataset is intended for statistical or clinical analysis.

Metadata for variables includes the basic SAS variable attributes like variable name, label or description, data type, and format as well as SDM-specific information. SDM-specific information includes variable origin and role. Metadata for a variable can also include a comment. Origin indicates where the values for a variable came from: source documents such as CRFs, derived variables including the derivation algorithm or a hyperlink in the PDF to the algorithm, or from another dataset. Role can be key, selection, review, or support. The SDM does include a provision for CDISC to extend this list of roles as the SDM and other CDISC standards mature, or for sponsors to extend it if they need to.

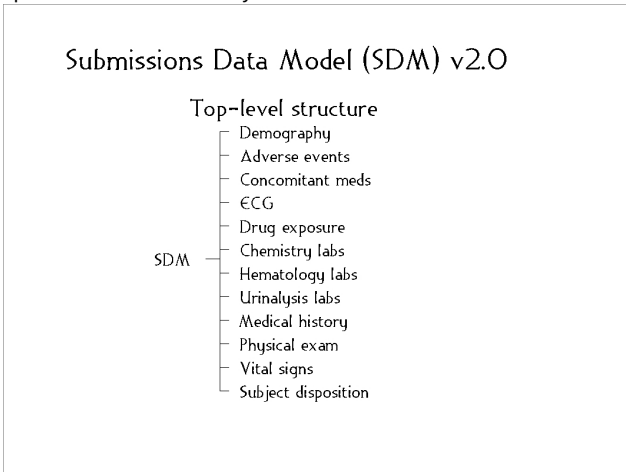

Figure 5 The SDM models non-efficacy data in 12 clinical domains.

XML IN SAS

Two features of XML form a barrier to accessibility in SAS. First, XML is hierarchical and relationships between data fields in XML are expressed in terms of ancestors, sibs, and descendents rather than through common values of key fields as in SAS and other flat and relational file structures. Second, XML is always just plain text. Ancestors, sibs, and descendents are all explicitly named with text tags that identify their beginning and end in the hierarchy. Since XML is always just plain text, it's clumsy to work with in SAS.

TWO STAGES

One way to overcome this barrier is by flattening a hierarchical, text-based XML file into a long, skinny SAS dataset. This dataset has to preserve all of the native XML hierarchy information in numeric indices. Hierarchy information and data are both available in the long, skinny dataset and can be easily processed in SAS to create, manage, and update the traditional short, fat SAS datasets with their horizontal structure. Short, fat datasets have lots of non-key data fields on a record. Long, skinny datasets have one non-key data field on a record. The use of the long, skinny dataset as a staging area between XML and traditional short, fat SAS datasets for both data importing and exporting provides a natural environment for someone familiar with SAS to work with XML.

INDEXING

The indexing is done in two dimensions for each item of data in an XML file. The first dimension depends on the number of ancestor XML elements a data point has and the second dimension depends on the number of sibling XML elements each ancestor and the data element being indexed have. In the terminology of SAS programming, the index for a data point is an array with dimension equal to the number of ancestors the data point has plus one. The value of each variable in this array is the number of siblings the corresponding ancestor has in the XML. A record in a stage 1 dataset, therefore, has one content field and several corresponding index fields. This very simple, consistent record layout is always used, for all XML-formatted data.

The indices preserve all of the XML hierarchy information in a numeric, text-free, flat file layout where it is highly accessible for conventional SAS programming. This gives SAS programmers the capabilities to transform one XML instance into a different XML instance in SAS without XSLT, perl, Java, Xpath, SAX, the DOM, or other exotic technologies.

It is highly desirable that a given element type in a particular context of element types always have the same index regardless of the amount and types of data preceding it in a particular XML instance. This can be done by creating an XML document that shows a data model, a canonical document. Canonical documents ("candoc") are the means to derive the index for each element in an instance from a data model, the candoc itself. Since the indexing is invariant to the data in a particular instance, stage 2 programs are generic and highly reusable.

FREE TOOL

Zurich Biostatistics' Tekoa Technology[sm] toolkit for working with XML in SAS uses this two stage approach with a simple plug-in interface. The toolkit with an ODM add-on is distributed for free by Zurich Biostatistics.

Figure 6 Two-stage import and export of XML into SAS provides a familiar DATA-step interface to XML.

Tekoa Technology breaks data transfer between XML and SAS into two stages. Stage 1 is generic. It never requires programming and always produces a long, skinny SAS dataset from XML, or for export, an XML file from a long, skinny dataset. Stage 2 is where long, skinny stage 1 datasets are mapped into short, fat stage 2 datasets. The stage 2 datasets are typically what's used for statistical analysis and reports. In Tekoa, both stage 1 and stage 2 are true invertible processes, they work in both directions, import from XML to SAS and export from SAS to XML.

In stage one, a generic dataset is created that preserves the hierarchical XML structure information in a way that is accessible for conventional SAS DATA step programming. In stage two, the structure information is used to map the content to clinical domain datasets.

EXPERIMENTAL SAS INSTITUTE METHOD

The SAS Institute (SI) has an experimental XMLMAP macro that has been used to import CDISC XML into SAS datasets. For export from SAS datasets to XML, SI has custom-programmed routines to export CDISC XML from SAS datasets. As of this writing, the Institute has not made these routines available to me for evaluation. From what I've seen demoed, XMLMAP uses XPATH notation and an underlying event-driven approach to the import of data from XML to SAS. SI has not published its CDISC export technology.

SUMMARY

The ODM and SDM provide pharma industry consortium-developed, vendor-neutral, FDA-aware standards for clinical trials data to be transferred and archived. These standards provide consistent, stable targets for organizations to use in the development of innovative, low maintenance clinical systems. ODM is an XML format for clinical data and machine-readable metadata. SDM uses SAS version 5 transport format for data and PDF files for metadata but may very well be migrated to XML by CDISC.

The import, processing, and export of XML with SAS presents new challenges because XML-formatted data has a hierarchical structure and exists as plain text. A natural programming environment for programmers familiar with SAS to work with XML can be created in the DATA-step by using a two-stage process to import and export XML. In stage 1, data exist in long, skinny SAS datasets indexed according to position in the XML. In stage 2, data exist in the customary short, fat SAS datasets for analysis and reporting. Using the indices, generic, reusable programming can move data back and forth between stage 1, stage 2, and XML. Zurich Biostatistics offers a free SAS plug-in to enable this two-stage processing of XML.

CONCLUSION

The era of vendor-neutral, low maintenance clinical data gateways is here. They should help to shorten the time to market for new drugs, biologics, and medical devices and thus add to the health and well-being of society as well as to the profitability of their developers and manufacturers.

REFERENCES

1. CDISC ODM and SDM documentation at the CDISC web site: http://www.cdisc.org
2. Tekoa Technology documentation at the Zurich Biostatistics, Inc. web site: http://www.zbi.net
3. SAS Institute XML information at http://www.sas.com/rnd/base/topics/templateFAQ/xmlfaq.htm

ACKNOWLEDGMENTS

CONTACT INFORMATION

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

Michael Palmer
Zurich Biostatistics, Inc.
45 Park Place South
PMB 178
Morristown, NJ 07960
Phone: 973-727-0025
Email: mcpalmer@zbi.net
Web: www.zbi.net