Paper FC08

An FDA-Requested XML Replacement for SAS Version 5 Transport Files in U.S. Regulatory Submissions

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

ABSTRACT

At the FDA's request, a CDISC-sponsored group has developed an XML-based replacement for the SAS version 5 transport files that are currently used to send case report tabulations clinical trials data to the FDA. Limitations of version 5 transport files such as 8 character variable names, 40 character labels for variables, a 200 character limit for character data, an idiosyncratic representation for dates and times, and sparse built-in metadata stored with the data have become noticeable and troublesome as such limitations have disappeared from other software used in clinical data management. CDISC has developed the more flexible XML-based Operational Data Model (ODM). ODM provides XML analogs to "forms" and "fields" and it has built-in support for extensive metadata. ODM does not specify names for clinical domains (i.e., adverse events, medical history) or data fields. CDISC's Submissions Data Standards (SDS) specifies a structure and metadata for clinical domains and data fields. ODM, by design, provides an excellent XML format for clinical trials data with the SDS-specified structure and metadata. This ODM/SDS combination is the basis for the XML replacement for version 5 transport files in regulatory submissions.

INTRODUCTION

At the FDA's request, a CDISC-sponsored group is developing an XML-based replacement for the SAS version 5 transport files that are currently used to send case report tabulations from clinical trials to the FDA. The author is a co-author of this specification for the XML replacement for version 5 transport files.

The FDA's 1999 guidance, "Providing Regulatory Submissions in Electronic Format," recommends the use of SAS version 5 transport files to send case report tabulations data to the FDA. To enable that use, the SAS Institute put version 5 transport file specification in the public domain and making version 5 transport files non-proprietary. SAS version 5 was released in the mid-1980s. Since then, database structures and data transport have advanced and clinical database management systems (CDMS) used for clinical trials have advanced, too. But, version 5 transport files have stayed the same. This mismatch between the CDMS world and the transport of data to the FDA has becoming an increasing problem as the SAS version 5 transport format grows more and more obsolete.

Recently, the FDA recognized the need to look into a successor data transport format and encouraged a CDISC-sponsored project to develop an XML-based successor to the version 5 transport format for clinical trials data.

LIMITATIONS OF VERSION 5 TRANSPORT FILES

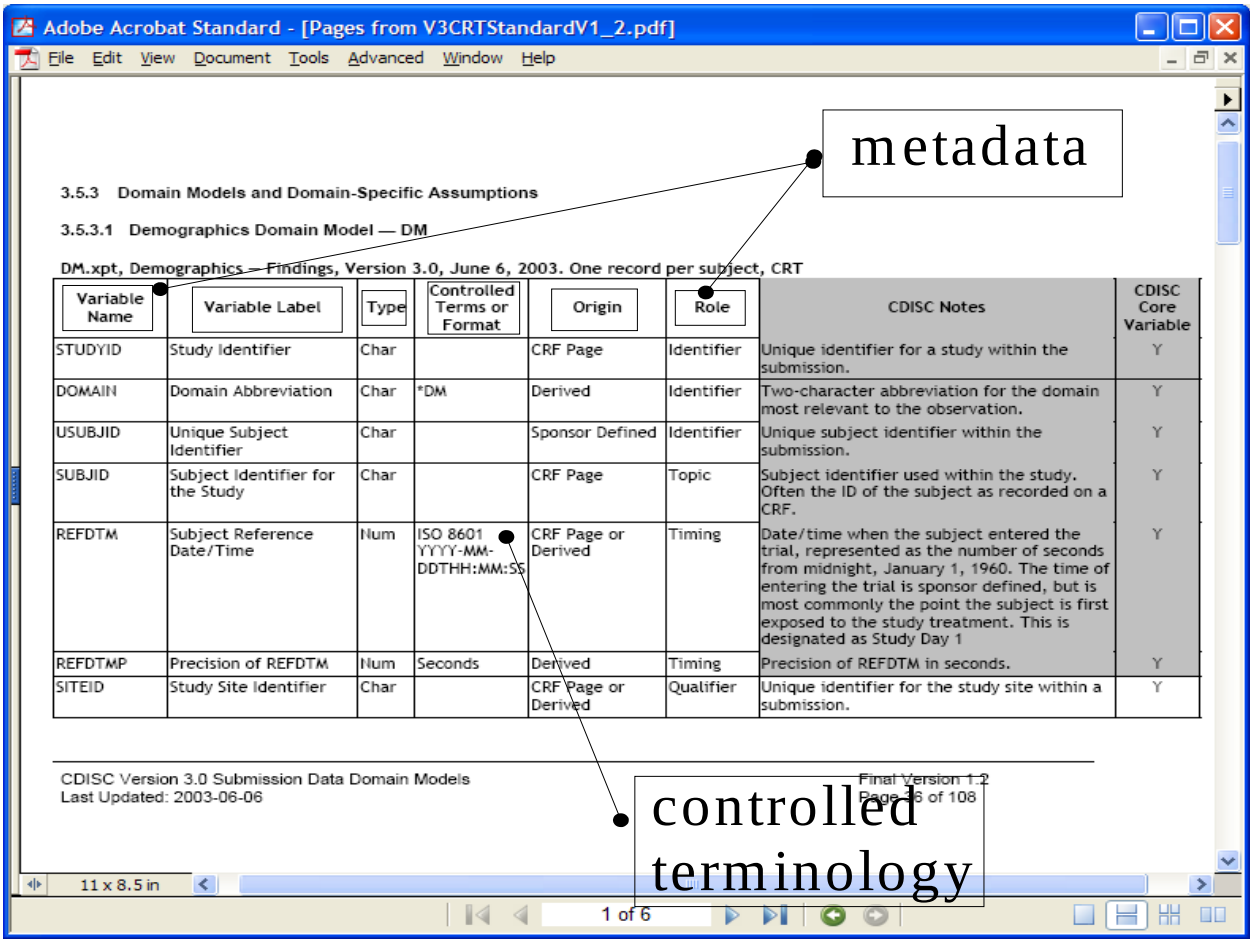
Limitations of version 5 transport files such as 8 character variable names, a 40 character limit for variable labels, a 200 character limit for character data, an idiosyncratic representation for dates and times based on elapsed time since a fixed reference date, no standard encoding for non-English characters, and sparse metadata have become noticeable and troublesome as such limitations have disappeared from other software used in clinical data management.

THE CDISC REPLACEMENT

The XML replacement for SAS version 5 transport files will be CDISC's Operational Data Model (ODM) version 1-2-0 configured for CDISC's Submission Data Standards.

Case report tabulations submitted to the FDA as SAS version 5 transport files use SDS clinical domains and metadata and the ODM replacement for SAS v5 transport files will also use SDS.

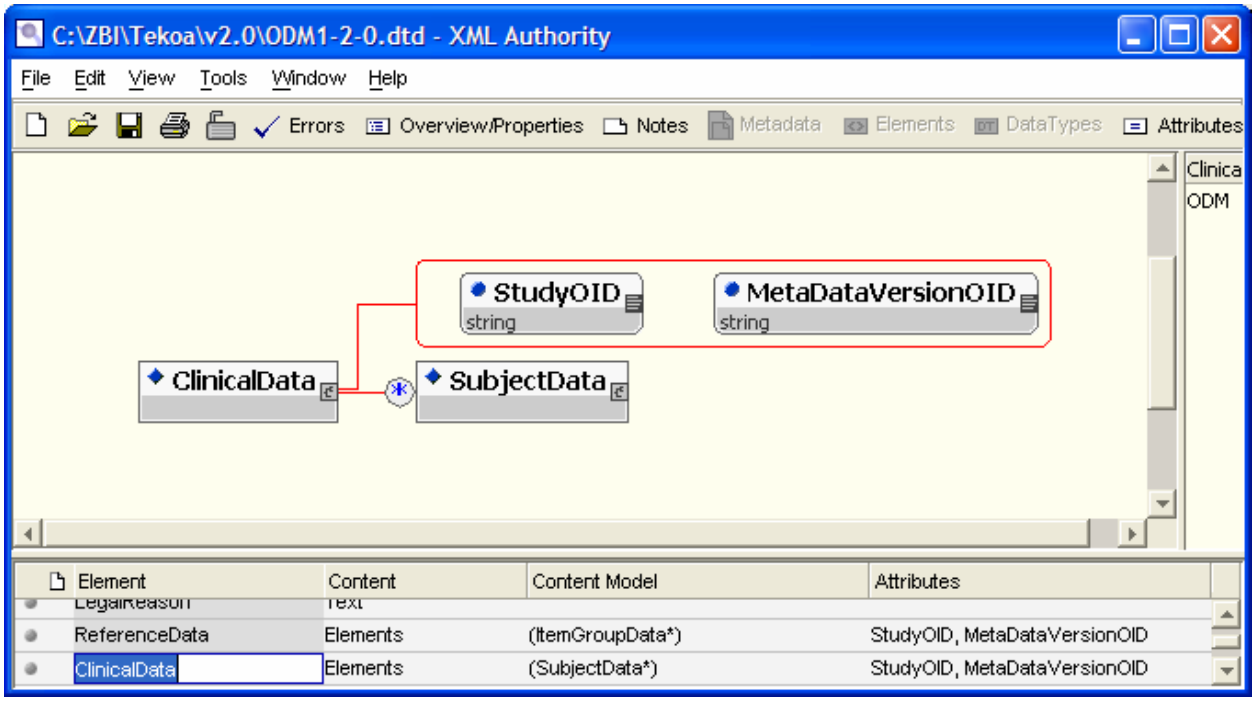
Figure 1. A page from Submissions Data Standards version 3.0. CDISC's Submission Data Standards specification will define the clinical domains and fields included in case report tabulation data submitted to the FDA in the new XML format.



SDS VERSION 3.X: CLINICAL DOMAINS IN A VERTICAL FORMAT

SDS version 3.x uses a vertical format to represent case report tabulations. Each record in an SDS v3 dataset has one field with the role of "topic" and that field is the principal piece of information relevant to the clinical domain of the dataset. For instance, in the spec for demography data in Figure 1, the topic field is "SUBJID" the subject identifier for the study, because demography data is whole-subject information such as birthday, race, and gender. All the other fields on a record have one of three other roles. They can be "timing" data, "identifier" data, or "qualifier" data. Timing fields have to do with some aspect of when the topic information took place in the context of the study. Identifier fields, as the name indicates, say something about the identity of the topic information. Qualifier fields hold data about the topic field that are neither timing nor identifier data. For the demography domain, REFDTM, the datetime when a subject entered the study is a timing field. STUDYID identifies the study and is an identifier field. Birthday, race, and gender relate information about a subject and they are qualifiers in SDS version 3.

Figure 2. A top-level view of the <ClinicalData> section of the ODM XML schema. ODM semantics put clinical trials data in the <ClinicalData> section and associate that data to a study with the StudyOID attribute and to metadata with the MetaDataVersionOID attribute. Clinical data are organized by study subject.

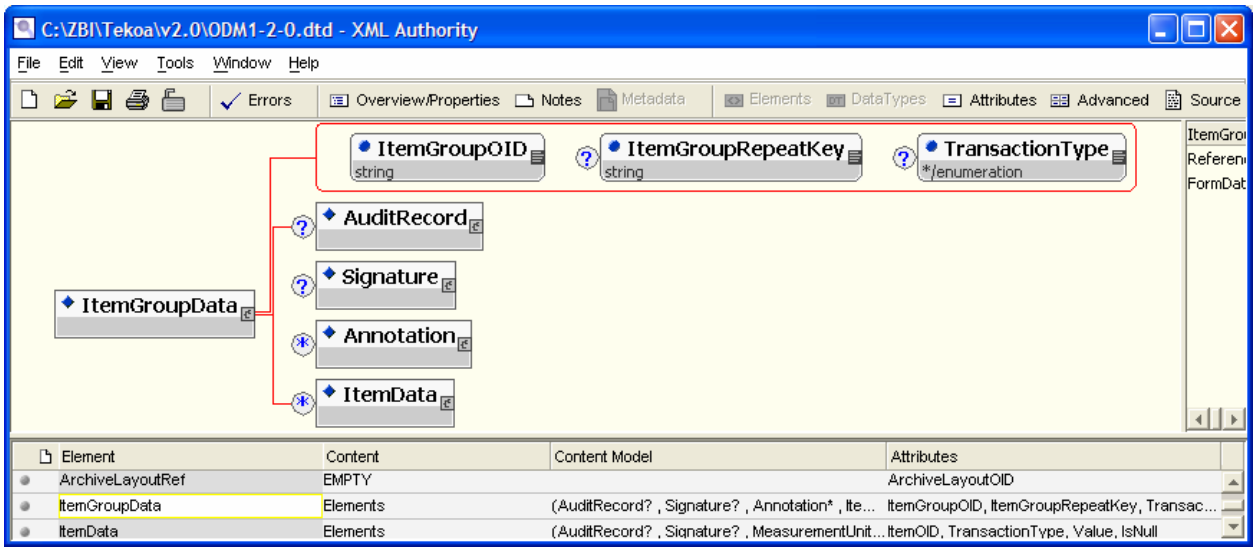


ODM

ODM is an XML schema for clinical trials data. It provides XML analogs to "forms" and "fields" and it has built-in support for extensive metadata. ODM does not specify names for clinical domains (i.e., adverse events, medical history) or data fields. CDISC's Submissions Data Model (SDS) specifies those domains and fields and the SDS structure and names are part of the ODM replacement for version 5 transport files. This ODM/SDS combination is the basis for the XML replacement for version 5 transport files in regulatory submissions.

ODM offers a number of advantages over version 5 transport format. ODM allows long variable names, long variable labels, ISO 8601 datetimes, non-English characters, and it does not limit the length of a character variable value to 200 characters. Also, ODM includes built-in, machine-readable metadata. The initial use of ODM to replace SAS version 5 transport files will maintain compatibility with SAS version 5 transport files so it will continue to have short variable names and labels.

Figure 3. The <ItemGroupData> section of the ODM schema. Each clinical domain in SDS will map to an <ItemGroupData> section in ODM. Data fields in a domain will map to <ItemData> elements.



SDS AND ODM TOGETHER REPLACE SAS VERSION 5 TRANSPORT FILES

The replacement data transport format uses CDISC's Operational Data Model (ODM) with some extensions and CDISC's Submission Data Model (SDS). Currently, SDS provides the logical structure for case report tabulations submitted to the FDA and it will continue to do this with the XML transport format.

SDS defines domains for clinical data, i.e., adverse events and physical exams, and data fields for each domain. In version 5 transport format, a clinical domain is a dataset and a data field is a variable in a dataset. ODM provides an analogous XML structure. In ODM, a clinical domain is an "item group" and a data field is an "item." The ODM item group/item structure is exactly analogous to the familiar dataset/variable structure. Figure 3 shows the <ItemGroupData> section of the ODM schema. The ItemGroupOID attribute references an <ItemGroupDef> element in the ODM metadata. The actual data values are in the <ItemData> tag.

Clinical domain metadata are defined in the SDS standard. These include the dataset name, a description of the dataset, dataset structure, purpose, key fields, and a reference to a file containing the data. Each of these metadata elements is mapped to an XML structure in ODM. For each variable in a domain, there are seven metadata fields: variable name, variable label, type, format, origin, role, and comments. These metadata elements are also mapped to XML structures in ODM. The result is machine-readable metadata in XML that will replace the non-machine-readable define.pdf file currently used for metadata.

Standardized codes and decodes for data values are an important part of clinical data. In the SAS world, the proprietary SAS format representation is used to store these code lists as datasets in version 5 format. In the non-proprietary ODM world, machine-readable code lists are stored with the other metadata and this practice will continue with the XML replacement for version 5 transport files.

CURRENT STATUS

Currently, the FDA is developing guidances that will replace the 1999 eSubmission guidances and these guidances will reference the CDISC SDS and ODM standards for case report tabulations. It is anticipated that the guidances will continue to reference SAS version 5 transport format as well. The FDA's parent department, the Department of Health and Human Services, has designated the Health Level 7 (HL7) organization as the standards development organization for healthcare information standards in the United States. In consequence of this, CDISC and HL7 are closely cooperating on the development of SDS and ODM. SDS v3.0 was approved as an HL7 informative document in spring 2003. FDA conducted a pilot project with SDS v3.0 in summer 2003 and the SDS team at CDISC has been revising SDS v3.0 since then. As of this writing, the revised spec is to be balloted as an informative document at HL7 in May 2003. The current status of ODM is that a configuration of ODM for SDS v3 metadata was balloted as an informative document at HL7 in January 2004 and passed. This ODM configuration is known as "define.xml" because it will replace the define.pdf document for case report tabulation metadata descriptions. define.pdf is discussed in the 1999 guidances. ODM configured for SDS v3 data as well as metadata will be balloted at HL7 after SDS v3's

approval. That approval is anticipated by June 2004. The bottom line is that define.xml for SDS v3 metadata is an approved HL7 document today and it is expected that it will be referenced by the FDA in its eCTD guidance by summer 2004 . ODM for case report tabulation data should follow by the end of 2004.

TOOLS TO SUPPORT THE NEW STANDARD

The SAS Institute is developing support for define.xml and for the ODM replacement for SAS version 5 transport files. This support is often referred to as PROC CDISC but, as of February 2004, PROC CDISC has not been seen outside of the SAS Institute so it's impossible to report on its capabilities. Zurich Biostatistics, Inc. (ZBI) has a set of XML tools known as the Tekoa toolkit for importing and exporting define.xml and ODM (and other XML schemas) to and from SAS. Tekoa was used to create the define.xml in the HL7 ballot and it has been used at CDISC to test ODM functionality. It is distributed for free by the author.

SUMMARY

SAS version 5 transport files have had a venerable second life since SAS v5 transporting case report tabulation data from companies to the FDA. But, their limitations such as short variable names, short labels, short text fields, lack of built-in metadata, idiosyncratic datetimes, and lack of standardization for non-English characters have finally caught up with them. An XML-based replaced using CDISC's Operational Data Model (version 1-2-0) and CDISC's Submission Data Standards (version 3.x) is being developed. In the ODM replacement, SDS clinical domains are mapped to <ItemGroupData> sections in ODM and fields in a domain are mapped to <ItemData> elements. A standard way to do this will be balloted as an informative document at HL7 in 2004. The SAS software will support the new transport format.

CONCLUSION

Many pharma and biotech companies have considerable investments in infrastructure designed to create and manage version 5 transport files and their accompanying define.pdf files. The replacement of both those formats with a single XML-based format will have significant impact on the existing infrastructure.

REFERENCES

CDISC "Standards" http://www.cdisc.org/standards/index.html (February 23, 2004)

Regulated Clinical Research Information Management (RCRIM) Technical Committee at HL7 http://www.hl7.org/ (February 23, 2004)

SAS Institute "FDA and SAS Technology " http://www.sas.com/govedu/fda/index.html (February 23, 2004)

World Wide Web Consortium "Date and Time Formats" World Wide Web Consortium (W3C) http://www.w3.org/TR/NOTE-datetime.html (February 23, 2004)

Zurich Biostatistics, Inc. "Presentations" http://www.zbi.net/NewFiles/leadership.html (February 23, 2004)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. The DATA-step toolkit for working with XML in SAS (Tekoa toolkit) is available for free from the author. Contact the author at:

Michael Palmer
Zurich Biostatistics, Inc.
45 Park Place, So., PMB 178
Morristown, New Jersey 07960
Work Phone: 973-277-9034
Email: mcpalmer@zbi.net
Web: www.zbi.net