Dataset-XML

The better transport format
for electronic submission data

A bit of history

- 19xx: FDA wants electronic submissions
- Looks for a transport format
- As most reviewers use SAS, a SAS transport format would be welcome
- However, SAS Transport is not open
- SAS and FDA agree to publish the "Transport 5" specification
- The "TS-140" document is published

- But wouldn't it have been better to simply use CSV?

The TS-140 specification

TS-140

RECORD LAYOUT OF A SAS® VERSION 5 OR 6 DATA SET IN SAS® TRANSPORT (XPORT) FORMAT

INTRODUCTION

All transport data set records are 80 bytes in length. If there is not sufficient data to reach 80 bytes, then a record is padded with ASCII blanks to 80 bytes. All character data are stored in ASCII, regardless of the operating system. All integers are stored using IBM-style integer format, and all floating-point numbers are stored using the IBM-style double (truncated if the variable's length is less than 8). [An exception to this is noted later.]

See the section "NUMERIC DATA FIELDS" for information on constructing IBM-style doubles.

RECORD LAYOUT

1. The first header record consists of the following character string, in ASCII:

```
HEADER RECORD*******LIBRARY HEADER RECORD!!!!!!!0000000000000000000000000
00000
```

2. The first real header record uses the following layout:

```
aaaaaaaabbbbbbbbcccccccddddddddeeeeeeee ffffffffffffffff
```

XPT and TS-140: the problematic part

TS-140

RECORD LAYOUT OF A SAS® VERSION 5 OR 6 DATA SET IN SAS® TRANSPORT (XPORT) FORMAT

INTRODUCTION

All transport data set records are 80 bytes in length. If there is not sufficient data to reach 80 bytes, then a record is padded with ASCII blanks to 80 bytes. All character data are stored in ASCII, regardless of the operating system. All integers are stored using IBM-style integer format, and all floating-point numbers are stored using the IBM-style double (truncated if the variable's length is less than 8). [An exception to this is noted later.]

See the section "NUMERIC DATA FIELDS" for information on constructing IBM-style doubles.

- But modern computers do NOT use "IBM-style" integers and doubles any more
- "IBM-style" was only meant for IBM mainframes and VAX computers

XPT and IBM mainframes
Do you still have one at home?

https://en.wikipedia.org/wiki/IBM_mainframe

Some more history: Dataset-XML

- Around 2005, CDISC and FDA performed a pilot to use ODM for submission purposes
- The pilot was interrupted and discontinued, as FDA decided that future submissions would be done using HL7-v3 messages
- CDISC: "HL7-v3 messages won't work!"
- FDA outsourced the development of the HL7-v3 messages to an external party
- After a number of years and xxx,xxx US$, it was reported by the external party that: "HL7-v3 messages don't work!"

Some more history: Dataset-XML

- 2014: CDISC publishes the Dataset-XML standard
- XML based standard building on define.xml
- To transport ANY tabular data
  - Submission and non-submission data

Dataset-XML builds on define.xml

Define.xml:

```
<ItemGroupDef IsReferenceData="No" Name="EX" OID="EX" Purpose="Tabulation"
    Repeating="Yes" def:ArchiveLocationID="Location.EX" def:Class="Interventions"
    def:Structure="One record per constant dosing interval per subject">
    <Description>
        <TranslatedText xml:lang="en">Exposure</TranslatedText>
    </Description>
    <ItemRef ItemOID="EX.STUDYID" Mandatory="Yes" OrderNumber="1" Role="IDENTIFIER" RoleCodeListOID="ROLES" KeySequence="1"/>
    <ItemRef ItemOID="EX.DOMAIN" Mandatory="Yes" OrderNumber="2" Role="IDENTIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.USUBJID" Mandatory="Yes" OrderNumber="3" Role="IDENTIFIER" RoleCodeListOID="ROLES" KeySequence="2"/>
    <ItemRef ItemOID="EX.EXSEQ" Mandatory="Yes" OrderNumber="4" Role="IDENTIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXTRT" Mandatory="Yes" OrderNumber="5" Role="TOPIC" RoleCodeListOID="ROLES" KeySequence="3"/>
    <ItemRef ItemOID="EX.EXDOSE" Mandatory="No" OrderNumber="6" Role="RECORD QUALIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXDOSU" Mandatory="No" OrderNumber="7" Role="VARIABLE QUALIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXDOSFRM" Mandatory="No" OrderNumber="8" Role="RECORD QUALIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXDOSFRQ" Mandatory="No" OrderNumber="9" Role="VARIABLE QUALIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXROUTE" Mandatory="No" OrderNumber="10" Role="VARIABLE QUALIFIER" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.VISITNUM" Mandatory="No" OrderNumber="11" Role="TIMING" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.VISIT" Mandatory="No" OrderNumber="12" Role="TIMING" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.VISITDY" Mandatory="No" OrderNumber="13" Role="TIMING" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXSTDTC" Mandatory="No" OrderNumber="14" Role="TIMING" RoleCodeListOID="ROLES" KeySequence="4"/>
    <ItemRef ItemOID="EX.EXENDTC" Mandatory="No" OrderNumber="15" Role="TIMING" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXSTDY" Mandatory="No" OrderNumber="16" Role="TIMING" RoleCodeListOID="ROLES"/>
    <ItemRef ItemOID="EX.EXENDY" Mandatory="No" OrderNumber="17" Role="TIMING" RoleCodeListOID="ROLES"/>
    <def:leaf ID="Location.EX" xlink:href="EX.xml">
        <def:title>ex.xml</def:title>
    </def:leaf>
</ItemGroupDef>
```

Dataset-XML builds on define.xml

Dataset-XML:

```
<ClinicalData StudyOID="LZZT" MetaDataVersionOID="CDISC.SDTM.3.1.0">
    <ItemGroupData ItemGroupOID="EX" data:ItemGroupDataSeq="1">
        <ItemData ItemOID="EX.STUDYID" Value="CDISCPILOT01"/>
        <ItemData ItemOID="EX.DOMAIN" Value="EX"/>
        <ItemData ItemOID="EX.USUBJID" Value="01-701-1015"/>
        <ItemData ItemOID="EX.EXSEQ" Value="1"/>
        <ItemData ItemOID="EX.EXTRT" Value="PLACEBO"/>
        <ItemData ItemOID="EX.EXDOSE" Value="0"/>
        <ItemData ItemOID="EX.EXDOSU" Value="mg"/>
        <ItemData ItemOID="EX.EXDOSFRM" Value="PATCH"/>
        <ItemData ItemOID="EX.EXDOSFRQ" Value="QD"/>
        <ItemData ItemOID="EX.EXROUTE" Value="TRANSDERMAL"/>
        <ItemData ItemOID="EX.VISITNUM" Value="3"/>
        <ItemData ItemOID="EX.VISIT" Value="BASELINE"/>
        <ItemData ItemOID="EX.VISITDY" Value="1"/>
        <ItemData ItemOID="EX.EXSTDTC" Value="2014-01-02"/>
        <ItemData ItemOID="EX.EXENDTC" Value="2014-01-16"/>
        <ItemData ItemOID="EX.EXSTDY" Value="1"/>
        <ItemData ItemOID="EX.EXENDY" Value="15"/>
    </ItemGroupData>
```

Advantages of Dataset-XML

- Modern technology
- Works 1:1 with define.xml
  - Easy validation against the define.xml
  - P.S.: the "define.xml" is the YOUR TRUTH about your study's metadata
  - Validation tool used by FDA (and probably you too) does NOT validate the define.xml correctly - it validates against the own idea of one company what the define.xml should be
- Allows audit trail on submission data
- Allows embedding of source data points (e.g. FHIR data point)

Disadvantages of Dataset-XML

- File size
  - Usually about 2-3x file size of XPT files
    (except for SUPPQUAL files)
  - But XML can easily be zipped - and tools can read zipped XML
    (zipped XML does not need to be unzipped - XPT is also binary …)
- Technology not known by medical reviewers
- "Not-invented-here" at the FDA
  - But used by the rest of the world
  - Pharma/FDA is the only industry using XPT

Dataset-XML and file sizes

Dataset	XPT file size	XML file size	Zipped XML file size
DM	0.1 MB	0.3 MB	0.02 MB
VS	23 MB	32 MB	0.8 MB
LB	33 MB	66 MB	2.0 MB
QS	33 MB	110 MB	2.8 MB
SUPPLB	55 MB	40 MB	2.0 MB

REMARKS:

- Also XPT files can be zipped
  - Also XPT is very inefficient
- XML can also be transformed into JSON, RDF Turtle, …

Does file size matter?

- It does NOT matter when information is immediately stored into a database or data warehouse
- It DOES matter when using memory sticks, file servers with slow intranet lines
- Which of both is the FDA doing?
- Does Amazon use XPT files?

Dataset-XML and audit trails

- As Dataset-XML is a subset of ODM, audit trails can easily be added

```
<ItemData ItemOID="AE.AESEV" Value="MILD">
    <AuditRecord EditPoint="Monitoring" UsedImputationMethod="Yes">
        <UserRef UserOID="ZBIuser000"/>
        <LocationRef LocationOID="XML4PharmaLocation"/>
        <DateTimeStamp>2013-12-21T11:59:59.9+01:00</DateTimeStamp>
        <ReasonForChange>Originally classified as moderate,
        then corrected to mild as subject had worked the whole day in the forest</ReasonForChange>
    </AuditRecord>
    <Signature>
        <UserRef UserOID="XML4Pharmauser000"/>
        <LocationRef LocationOID="XML4PharmaLocation"/>
        <SignatureRef SignatureOID="XML4PharmaSignature000"/>
        <DateTimeStamp>2013-12-31T11:59:59.9+01:00</DateTimeStamp>
    </Signature>
</ItemData>
```

Dataset-XML and Electronic Health Records

- As Dataset-XML is based on ODM, and ODM is extensible, EHR data points can easily be embedded. For example: FHIR

Structure

Name	Flags	Card.	Type	Description & Constraints
Observation	I		DomainResource	Measurements and simple assertions *+ If code is the same as a component code then associated with the code SHALL NOT be present* *+ dataAbsentReason SHALL only be present if Obs present* Elements defined in Ancestors: id, meta, implicitR contained, extension, modifierExtension
identifier	Σ	0..*	Identifier	Business Identifier for observation
basedOn	Σ	0..*	Reference(CarePlan | DeviceRequest | ImmunizationRecommendation | MedicationRequest | NutritionOrder | ProcedureRequest | ReferralRequest)	Fulfills plan, proposal or order
status	?! Σ	1..1	code	registered | preliminary | final | amended + ObservationStatus (Required)
category		0..*	CodeableConcept	Classification of type of observation Observation Category Codes (Preferred)

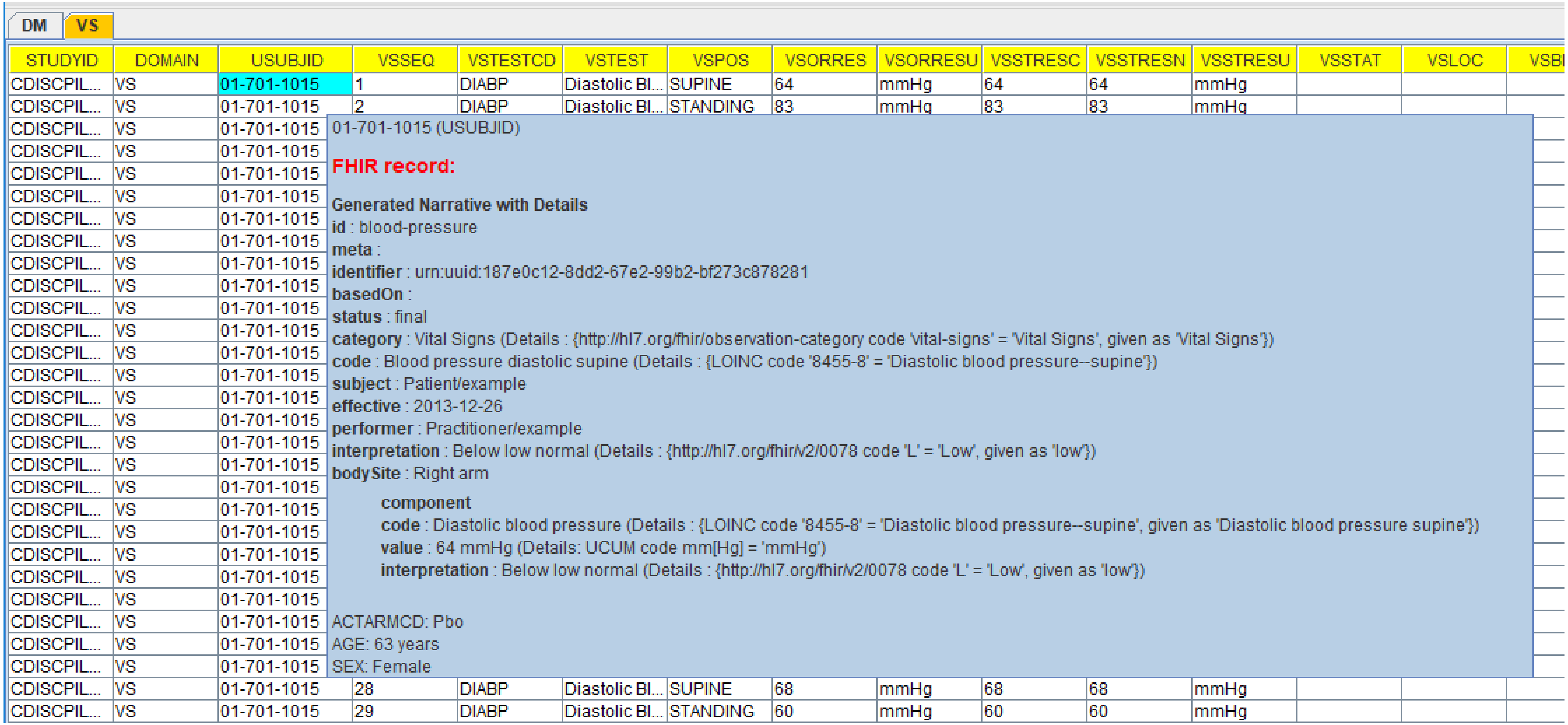
FHIR source record in SDTM record


```xml
<ItemGroupData data:ItemGroupDataSeq="1" ItemGroupOID="VS">
   <ItemData ItemOID="VS.STUDYID" Value="CDISCPILOT01"/>
   <ItemData ItemOID="VS.DOMAIN" Value="VS"/>
   <ItemData ItemOID="VS.USUBJID" Value="01-701-1015"/>
   <ItemData ItemOID="VS.VSSEQ" Value="1"/>
   <ItemData ItemOID="VS.VSTESTCD" Value="DIABP"/>
   <ItemData ItemOID="VS.VSTEST" Value="Diastolic Blood Pressure"/>
   <ItemData ItemOID="VS.VSPOS" Value="SUPINE"/>
   <ItemData ItemOID="VS.VSORRES" Value="64"/>
   <ItemData ItemOID="VS.VSORRESU" Value="mmHg"/>
   <ItemData ItemOID="VS.VSSTRESC" Value="64"/>
   <ItemData ItemOID="VS.VSSTRESN" Value="64"/>
   <ItemData ItemOID="VS.VSSTRESU" Value="mmHg"/>
   <ItemData ItemOID="VS.VISITNUM" Value="1"/>
   <ItemData ItemOID="VS.VISIT" Value="SCREENING 1"/>
   <ItemData ItemOID="VS.VISITDY" Value="-7"/>
   <ItemData ItemOID="VS.VSDTC" Value="2013-12-26"/>
   <ItemData ItemOID="VS.VSDY" Value="-7"/>
   <ItemData ItemOID="VS.VSTPT" Value="AFTER LYING DOWN FOR 5 MINUTES"/>
   <ItemData ItemOID="VS.VSTPTNUM" Value="815"/>
   <ItemData ItemOID="VS.VSELTM" Value="PT5M"/>
   <ItemData ItemOID="VS.VSTPTREF" Value="PATIENT SUPINE"/>
   <Observation xmlns="http://hl7.org/fhir">
      <id value="blood-pressure"/>
      <meta>
         <profile value="http://hl7.org/fhir/StructureDefinition/vitalsigns"/>
      </meta>
      <text> <status value="generated"/> <div xmlns="http://www.w3.org/1999/xhtml"><p> <b> Generated Narrative with Details</b> </p> <p> <b> id</b> : blood-pres
         given as 'Vital Signs'})</span> </p> <p> <b> code</b> : Blood pressure diastolic supine<span> (Details : {LOINC code '8455-8' = 'Diastolic blood pressu
                  pressure supine'})</span> </p> <p> <b> value</b> : 64 mmHg<span>  (Details: UCUM code mm[Hg] = 'mmHg')</span> </p> <p> <b> interpretation</b>
                     <system value="urn:ietf:rfc:3986"/>
                     <value value="urn:uuid:187e0c12-8dd2-67e2-99b2-bf273c878281"/>
                  </identifier>
      <!-- demonstrating the use of the baseOn element with a fictive identifier   -->
      <basedOn>
```

FHIR source record in SDTM record (detail)


```
<ItemData ItemOID="VS.VSTPTREF" Value="PATIENT SUPINE"/>
<Observation xmlns="http://hl7.org/fhir">
   <id value="blood-pressure"/>
   <meta>
      <profile value="http://hl7.org/fhir/StructureDefinition/vitalsigns"/>
   </meta>
   <text> <status value="generated"/> <div xmlns="http://www.w3.org/1999/xhtml"><p> <b> Generated Narrative with Details</b> </p> <p> <b> id</b> : blood-pressure</p>
      given as 'Vital Signs'})</span> </p> <p> <b> code</b> : Blood pressure diastolic supine<span> (Details : {LOINC code '8455-8' = 'Diastolic blood pressure--supi
               pressure supine'})</span> </p> <p> <b> value</b> : 64 mmHg<span>  (Details: UCUM code mm[Hg] = 'mmHg')</span> </p> <p> <b> interpretation</b> : Below
                  <system value="urn:ietf:rfc:3986"/>
                  <value value="urn:uuid:187e0c12-8dd2-67e2-99b2-bf273c878281"/>
               </identifier>
   <!-- demonstrating the use of the basedOn element with a fictive identifier   -->
   <basedOn>
      <identifier>
         <system value="https://acme.org/identifiers"/>
         <value value="1234"/>
      </identifier>
   </basedOn>
   <status value="final"/>
   <category>
      <coding>
         <system value="http://hl7.org/fhir/observation-category"/>
         <code value="vital-signs"/>
         <display value="Vital Signs"/>
      </coding>
   </category>
   <code>
      <coding>
         <system value="http://loinc.org"/>
         <code value="8455-8"/>
         <display value="Diastolic blood pressure--supine"/>
      </coding>
      <text value="Diastolic blood pressure--supine"/>
   </code>
```

And can easily be visualized to the reviewer



STUDYID	DOMAIN	USUBJID	VSSEQ	VSTESTCD	VSTEST	VSPOS	VSORRES	VSORRESU	VSSTRESC	VSSTRESN	VSSTRESU	VSSTAT	VSLOC	VSB
CDISCPIL...	VS	01-701-1015	1	DIABP	Diastolic Bl...	SUPINE	64	mmHg	64	64	mmHg			
CDISCPIL...	VS	01-701-1015	2	DIABP	Diastolic Bl...	STANDING	83	mmHg	83	83	mmHg			
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015												
CDISCPIL...	VS	01-701-1015	28	DIABP	Diastolic Bl...	SUPINE	68	mmHg	68	68	mmHg			
CDISCPIL...	VS	01-701-1015	29	DIABP	Diastolic Bl...	STANDING	60	mmHg	60	60	mmHg			

01-701-1015 (USUBJID)

FHIR record:

Generated Narrative with Details
id : blood-pressure
meta :
identifier : urn:uuid:187e0c12-8dd2-67e2-99b2-bf273c878281
basedOn :
status : final
category : Vital Signs (Details : {http://hl7.org/fhir/observation-category code 'vital-signs' = 'Vital Signs', given as 'Vital Signs'})
code : Blood pressure diastolic supine (Details : {LOINC code '8455-8' = 'Diastolic blood pressure--supine'})
subject : Patient/example
effective : 2013-12-26
performer : Practitioner/example
interpretation : Below low normal (Details : {http://hl7.org/fhir/v2/0078 code 'L' = 'Low', given as 'low'})
bodySite : Right arm

> **component**
> **code** : Diastolic blood pressure (Details : {LOINC code '8455-8' = 'Diastolic blood pressure--supine', given as 'Diastolic blood pressure supine'})
> **value** : 64 mmHg (Details: UCUM code mm[Hg] = 'mmHg')
> **interpretation** : Below low normal (Details : {http://hl7.org/fhir/v2/0078 code 'L' = 'Low', given as 'low'})

ACTARMCD: Pbo
AGE: 63 years
SEX: Female

Visualization by the "Smart Dataset-XML Viewer"

Tools for working with Dataset-XML

- See: https://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources

Name	Description	Provided By	Links
	• Freely available		source forge
Smart Dataset-XML Viewer	• Similar to the SAS Viewer, but with additional functionality • Supports working with Define-XML + Dataset-XML files • Supports SDTM, SEND, and ADaM data • Basic validation • Open source	Univ. Appl. Sciences FH Joanneum Graz - eHealth	• The application and tutorial is available under the Smart SDS-XML View project on source forge • Youtube video on the Smart Dataset-XML Viewer
EZ Convert	• Converts Dataset-XML files into SAS datasets • Supports Define-XML Version 1 or Version 2 • Open Source	@Sally Cassells	• EZConvert Demonstration video • Beta version of EZConvert

SAS Clinical Standards Toolkit	• Dataset-XML support (writing/reading /validation) will be part of the next release of SAS® Clinical Standards Toolkit. Updated information will be published at the SAS web site.
	• Support for Dataset-XML is available as a pre-production package that contains SAS macros, XML schema files, sample data, and sample programs to support the following functionality:
	  • Creating Dataset-XML files from SAS data sets
	  • Creating SAS data sets from Dataset-XML files
	  • Validating Dataset-XML files against an XML schema
	  • Comparing original SAS data sets with SAS data sets created from Dataset-XML files
	• These macros are standalone and do not require SAS® Clinical Standards Toolkit.
		SAS Institute Inc.	• SAS Clinical Standards Toolkit
			• SAS Macros to support Dataset-XML v1.0.0

OpenCDISC v1.5	• OpenCDISC v1.5 works with Dataset-XML files and Define-XML v2.0	OpenCDISC	• OpenCDISC.org
R4CDISC	• R4CDISC package includes functions for reading Dataset-XML and Define-XML files.	Ippei Akiya	• CRAN project page with downloads • Reference manual

Tools for working with Dataset-XML

Name	Description	Provided By	Links
XPT2DatasetXML	• Transforms XPT datasets into Dataset-XML datasets • Freely available	XML4Pharma	• Available under the Smart SDS-XML View project on source forge
Smart Dataset-XML Viewer	• Similar to the SAS Viewer, but with additional functionality • Supports working with Define-XML + Dataset-XML files • Supports SDTM, SEND, and ADaM data • Basic validation • Open source	Univ. Appl. Sciences FH Joanneum Graz - eHealth	• The application and tutorial is available under the Smart SDS-XML View project on source forge • Youtube video on the Smart Dataset-XML Viewer
EZ Convert	• Converts Dataset-XML files into SAS datasets • Supports Define-XML Version 1 or Version 2 • Open Source	@Sally Cassells	• EZConvert Demonstration video • Beta version of EZConvert

SAS Clinical Standards Toolkit	• Dataset-XML support (writing/reading/validation) will be part of the next release of SAS® Clinical Standards Toolkit. Updated information will be published at the SAS web site.
	• Support for Dataset-XML is available as a pre-production package that contains SAS macros, XML schema files, sample data, and sample programs to support the following functionality:
	  • Creating Dataset-XML files from SAS data sets
	  • Creating SAS data sets from Dataset-XML files
	  • Validating Dataset-XML files against an XML schema
	  • Comparing original SAS data sets with SAS data sets created from Dataset-XML files
	• These macros are standalone and do not require SAS® Clinical Standards Toolkit.
		SAS Institute Inc.	• SAS Clinical Standards Toolkit
			• SAS Macros to support Dataset-XML v1.0.0

OpenCDISC v1.5	• OpenCDISC v1.5 works with Dataset-XML files and Define-XML v2.0	OpenCDISC	• OpenCDISC.org
R4CDISC	• R4CDISC package includes functions for reading Dataset-XML and Define-XML files.	Ippei Akiya	• CRAN project page with downloads • Reference manual

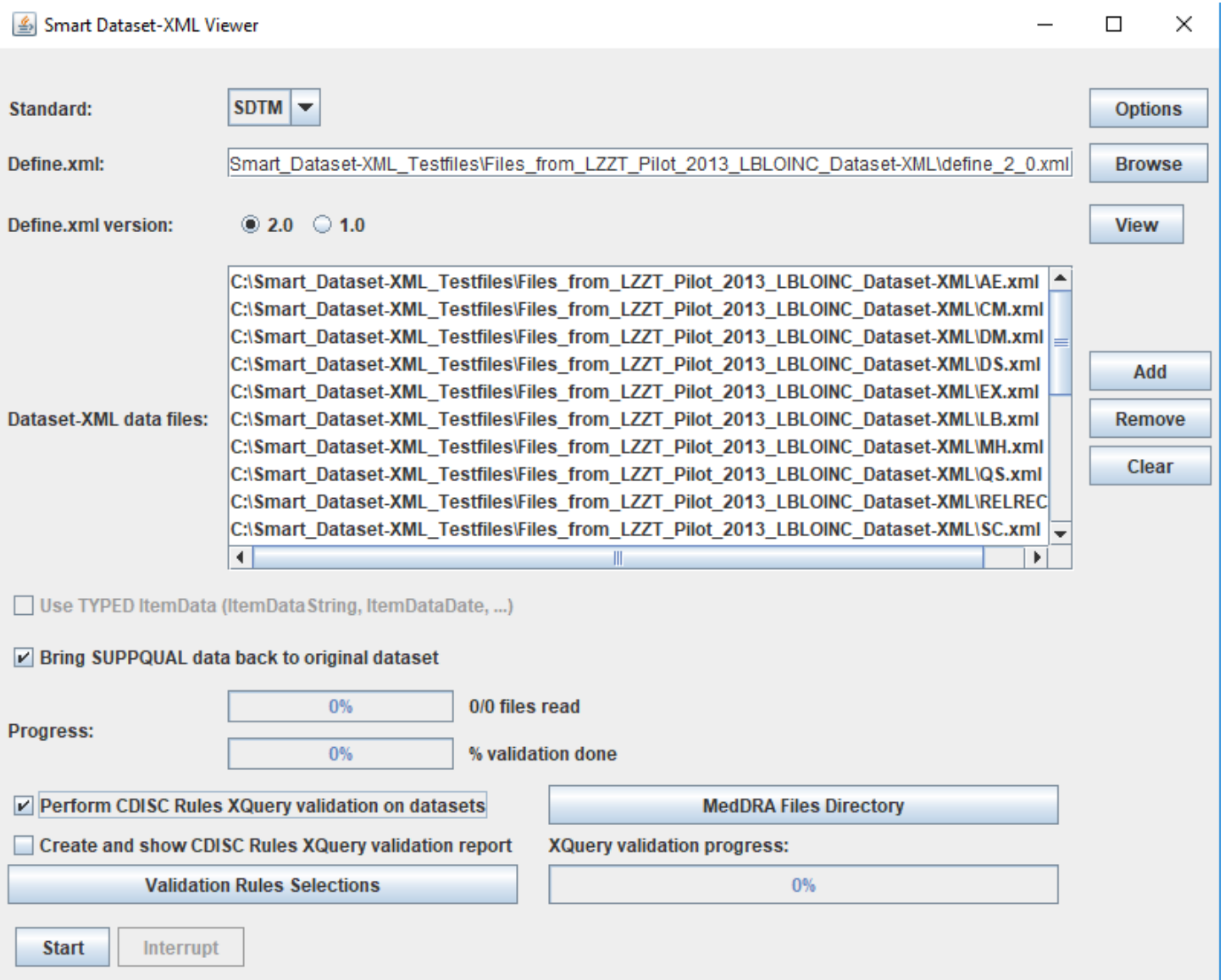
Smart Dataset-XML Viewer

- Viewer software for inspecting SDTM/SEND/ADaM submissions
- Similar to "SASViewer" or "SAS Universal Viewer"
  - But much smarter for SDTM, SEND and ADaM files
- Soon to extended (to also convert XPT files) and to be renamed to "Smart Submission Dataset Viewer"
- Essentially, reviewers should NOT use such viewers, but load the data into databases, and query these databases - they don't however

Smart Dataset-XML Viewer

- Can use modern technologies such as RESTful Web Services
- Can connect to scientific information systems such as these from
  - The National Library of Medicine
  - LOINC
  - SNOMED-CT
  - UMLS (Unified Medical Language System)
- The FDA systems do apparently use none of these

Smart Dataset-XML Viewer - Demo time!



The role of define.xml in submissions

- Define.xml is PRIMARILY meant to be used as a machine-readable specification of the submission metadata
- Most reviewers however only use the human-readable VIEW
- The define.xml is YOUR TRUTH of what is in the submission and not that of CDISC, Pinnacle21 or anyone else
- So better take care the define.xml is of high quality

High Quality define.xml

- Made long time before you do the submission
- Generated BEFORE the datasets are generated
  - And not generated at the last moment using crap software
- Possibly generated already at or before study start
  - Although you cannot know everything in advance
- Possibly used as a specification
  - For the CRO or service provider
  - With the mapping instructions between operational data and SDTM/SEND (ADaM may be different)

Using define.xml as a mapping specification

```
<MethodDef Name="Computation method for AEREL" OID="IMP.MyStudy:AE.32.AE.AEREL"
           Type="Computation">
   <Description>
      <TranslatedText xml:lang="en">SDTM-ETL mapping for AEREL</TranslatedText>
   </Description>
   <FormalExpression Context="SDTM-ETL"># Mapping using ODM element ItemData with ItemOID IT.AEREL
# Generalized for all StudyEvents
# Generalized for all StudyEvents
# Using decoded values from ODM CodeList CL.AEREL
$AE.AEREL = xpath(/StudyEventData/FormData[@FormOID='FORM.AE']/ItemGroupData[@ItemGroupOID='IG.AE']/ItemData
$TEMP = "";
if ($AE.AEREL == '0') {
    $TEMP = 'NONE';
} elsif ($AE.AEREL == '1') {
    $TEMP = 'UNLIKELY';
} elsif ($AE.AEREL == '2') {
    $TEMP = 'POSSIBLE';
} elsif ($AE.AEREL == '3') {
    $TEMP = 'PROBABLE';
} else {
    $TEMP = '';
}
$AE.AEREL = $TEMP;</FormalExpression>
</MethodDef>
```

- Define.xml used by the "SDTM-ETL" mapping software

Stylesheets for define.xml

- Are the sponsor's responsibility
- FDA should essentially use their own stylesheet
  (but they have no idea how to do that)
- The stylesheet helps the reviewer to find things easier,
  in a better and user-friendly way …
- Check for stylesheets made available by Phuse, CDISC, …
  - And if you don't like it, find/hire an XSLT specialist

Use of Dataset-XML: Validation: Open Rules for CDISC Standards (ORCS)

- Initiative by a few CDISC volunteers
- Goal is to have all validation rules in a format that is both:
  - Human-readable (at least to people who have minimal programming skills)
  - Machine-readable (machines should be able to execute it)
- To be used as a "reference implementation" of the rules
  - Anyone can develop ist own validation software, but the results minimally need to be identical to that of the reference implementation
  - Usual methodology in software language development (e.g. Java)

Open Rules for CDISC Standards (ORCS)

- Problem: there is no universal "rule description language"
  - Even not a standardized "pseudo code language"
- The SDTM team published rules with pseudo code
  - Has been ignored by the FDA validation software

--LNKGRP	--LNKGRP present in another domain	--LNKGRP present in a domain
--LNKID	--LNKID present in another domain	--LNKID present in a domain
--RFTDTC	--RFTDTC = null	--TPTREF = null
--SCAT	--SCAT ^= --CAT	--SCAT ^= null

- For XML documents, there is however XQuery
  - W3C standard

W3C®INK

W3C XML Query (XQuery)

Open Rules for CDISC Standards - XQuery

- We can use XQuery for defining rules that are as well human-readable as machine-readable
- But XQuery only works on XML …
- If we use Dataset-XML, we can thus have open, vendor-neutral rules definitions using XQuery
- Which can still act as a "reference implementation"

ORCS Rule example - FDAC017-FDAC018

```
(: Rule FDAC017-FDAC018: SDTM Required variable not found - Variables described in SDTM as Required
 must be included in the dataset :)
(: The following Query relies on that the define.xml is complete and
 that Mandatory='Yes' is set for each required variable :)
xquery version "3.0";
declare namespace def = "http://www.cdisc.org/ns/def/v2.0";
declare namespace odm="http://www.cdisc.org/ns/odm/v1.3";
declare namespace data="http://www.cdisc.org/ns/Dataset-XML/v1.0";
declare namespace xlink="http://www.w3.org/1999/xlink";
declare namespace request="http://exist-db.org/xquery/request";
(: "declare variable ... external" allows to pass $base and $define from an external programm :)
(: declare variable $base external; :)
(: declare variable $define external; :)
let $base := '/db/fda_submissions/cdisc01/'
let $define := 'define2-0-0-example-sdtm.xml'
(: iterate over all datasets mentioned in the define.xml :)
for $itemgroup in doc(concat($base,$define))//odm:ItemGroupDef
    (: get all the ItemRef-OIDs which have 'Mandatory="Yes" :)
    let $mandatory := $itemgroup/odm:ItemRef[@Mandatory='Yes']/@ItemOID
    (: get the dataset itself :)
    let $datasetfilename := $itemgroup/def:leaf/@xlink:href
    let $dataset := doc(concat($base,$datasetfilename))
    let $datasetname := $itemgroup/@Name
    (: iterate over all the records :)
    for $record in $dataset//odm:ItemGroupData
        let $recnum := $record/@data:ItemGroupDataSeq
        (: iterate over the 'mandatory' OIDs  :)
        for $m in $mandatory
            (: and give an error when there is no such ItemData/@ItemOID  :)
            let $varname := doc(concat($base,$define))//odm:ItemDef[@OID=$m]/@Name
            where not($record/odm:ItemData[@ItemOID=$m])
            return <error rule="FDAC017" datasetname="{$datasetname}" variable="{data($varname)}"
            rulelastupdate="2015-08-31" recordnumber="{$recnum}">
            No data found for required variable {data($varname)} in record number {data($recnum)}
            in dataset {data($datasetname)}</error>
```

Open Rules for CDISC Standards - Principles

- Basis is the define.xml (which is YOUR truth about the submission)
- Information from the SDTM-IG can be queried using RESTful web services from SHARE and other CDISC services
  - E.g. whether a variable is "required", "expected" or "permissible"
- Descriptive error messages are provided
  - Including the "record number"
- ANYONE can implement these open rules in their own software independent of programming language (Java, C#, SAS, Python, …)

ORCS: using RESTful web services

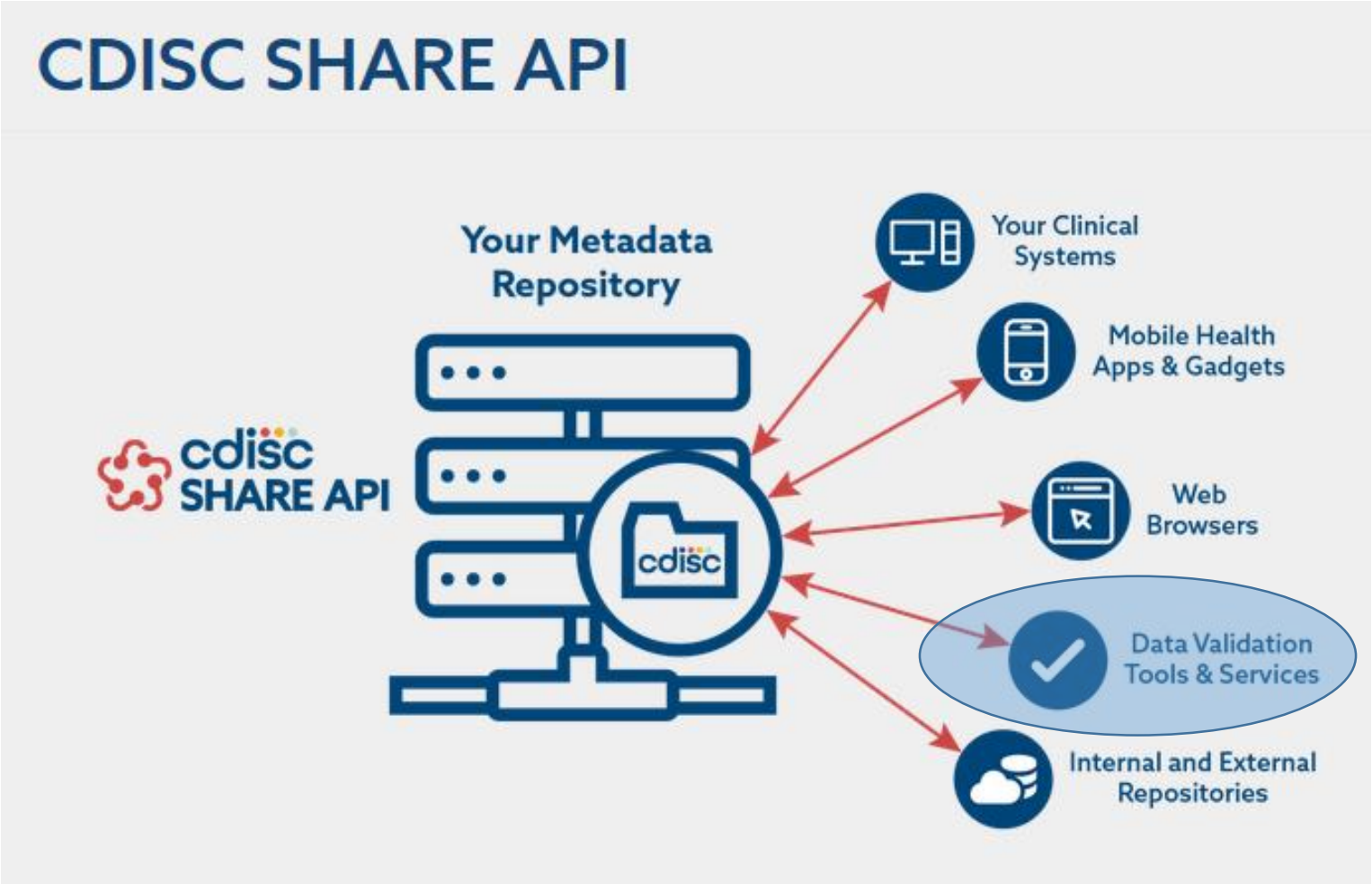
- CDISC RESTful web services API at
  - http://xml4pharmaserver.com/WebServices/CDISCSDSVariables_webservices.html
- Will also be available through the SHARE v2 API


```
▼<XML4PharmaServerWebServiceResponse ServerDateTime="2017-03-07T20:05:55">
  ▼<WebServiceRequest>
     http://xml4pharmaserver.com:8080/CDISCCTService/rest/SDTMVariableInfoForDomainAndVersion/3.2/LB/LBSTDTC
   </WebServiceRequest>
  ▼<Response>
    ▼<Variable Name="LBSTDTC" SDTMIGVersion="3.2" Domain="LB">
       <Observation_Class>All classes</Observation_Class>
       <Variable_Label_From_Model>Start Date/Time of Observation</Variable_Label_From_Model>
       <Type>Char</Type>
       <Role>Timing</Role>
     ▼<CDISC_Notes>
        Start date/time of an observation represented in ISO 8601 character format.
       </CDISC_Notes>
       <Core>Perm</Core>
     </Variable>
   </Response>
 </XML4PharmaServerWebServiceResponse>
```

ORCS: using RESTful web services - Example

- Rules engine needs to check whether variable is "required", "expected" or "permissible"
  - This depends on the version of the SDTM-IG
  - Define.xml may "upgrade" this
    - E.g. may state that permissible variable LBLOINC is "required" in the current submission / dataset
- Rule definition asks SHARE whether the variable is "required", "expected" or "permissible"
  - Using the RESTful web service API, given the variable name and SDTM-IG version
- And checks whether the define.xml did not "upgrade" this
- And then checks whether all records comply with the requirement

ORCS: using RESTful web services
Future: working with the SHARE API



Open Rules for CDISC Standards: Advantages

- Really open
- Freely available
- Software language independent
- Vendor neutral
- Human-readable as well as machine-readable
- Clear and exact error messages
- Error messages come as XML for further processing
  - But can also be transformed to … Excel …
- Soon an official CDISC project => will hopefully later go into SHARE

Open Rules for CDISC Standards: Disdvantages

- Currently, use of Dataset-XML necessary
  - But the FDA does not use Dataset-XML
- Slower than Pinnacle21 validation
  - Each rule must first be compiled "on the fly"
  - Rules must be executed sequentially
    - Though some people have already tried parallelization
- XQuery pretty unknown among SAS programmers

Open Rules for CDISC Standards
Call for volunteers

- We especially need volunteers for allowing us to implement the "ADaM Validation Checks v.1.3"
- No XQuery nor XML knowledge required
- Good knowledge of ADaM required
- Willing to provide test examples

A short overview of other "Jozef projects"

- Annotated Protocol in XML
- Machine-readable SDTM-IG
- Connecting CDISC-CT to healthcare controlled terminology

Annotated Protocol in XML

- Currently, protocols come as Word, or PDF documents
- Must be interpreted by humans, in order e.g. to:
- Define which forms with what content
- Which tests need to be performed

Annotated Protocol in XML: Example: CDISC Diabetes TAUG

3.1 Laboratory Tests

The audience for this laboratory section is not targeted for medical professionals, but is meant to find a balance between general and detailed.

3.1.1 Glucose Homeostasis and Diabetes Related Markers

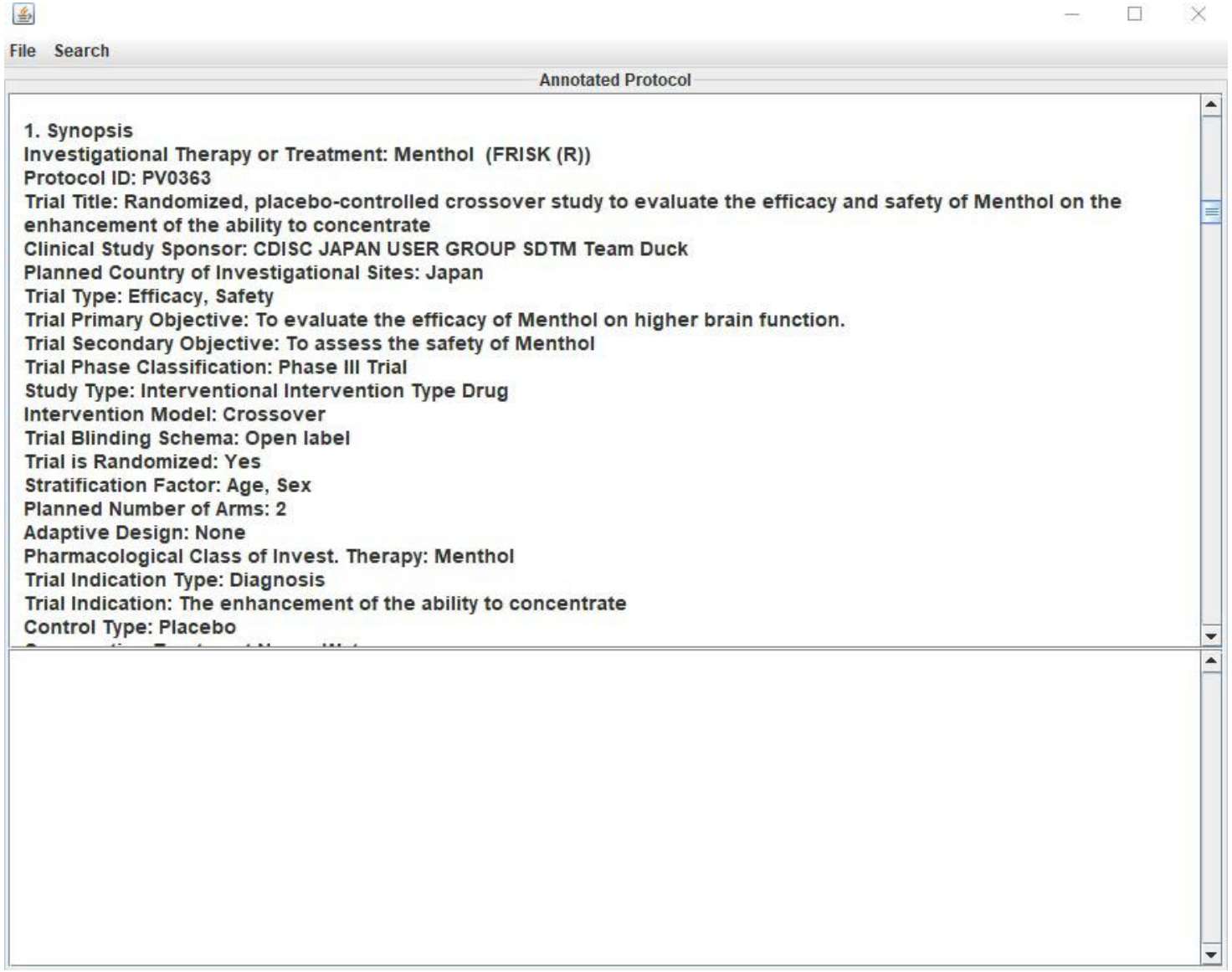
Diabetes is generally diagnosed by blood tests; pre-diabetes and early T2DM may have few or no markers. Blood glucose concentrations are affected by many factors, but particularly by meals, so random blood glucose may not be a reliable basis for diagnosis, unless markedly elevated (e.g. >200 mg/dL or 11.1 mmol/L) and accompanied by typical symptoms of hyperglycemia. Fasting blood glucose and measurements obtained during an oral glucose tolerance test (see Section 3.2.2) are more reliable, but they measure glucose concentrations only in the short term and require fasting or glucose loading. Standardized glycosylated hemoglobin A1c assays reliably estimate average glucose concentrations over a longer term, have less variability during stress and illness, and are sometimes more specific for identifying individuals with diabetes or at increased risk for diabetes.

The test names in the following tables should not be relied upon for current controlled terminology. Refer to the NCI EVS page (http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc) for current CDISC terminology.

Common Test Abbreviation	Test Name	Description	Specimen(s)
A1c HbA1c	Glycosylated Hemoglobin, Glycated Hemoglobin, Hemoglobin A1c, Glycosylated Hemoglobin A1c	Glycosylated hemoglobin is formed in a non-enzymatic glycation pathway by hemoglobin's exposure to plasma glucose. As the average amount of plasma glucose increases, the fraction of glycosylated hemoglobin increases in a predictable way. This serves as a marker for average blood glucose concentrations over the previous two to three months prior to the measurement.	Blood
	Glucose	Glucose is a carbohydrate and is the most important simple sugar in human metabolism. The body naturally tightly regulates the glucose concentrations as a part of metabolic homeostasis. Glucose is transported from the intestines or liver to body cells via the bloodstream and is made available for cell absorption via the hormone insulin. Glucose concentrations are usually lowest in the morning before the first meal of the day or an extended time (e.g. 8 hours) since the last meal. This is called "fasting glucose". A consistently high glucose concentration is referred to as hyperglycemia. Low glucose concentrations are referred to as hypoglycemia. Diabetes mellitus is characterized by consistent hyperglycemia from any of several causes. T1DM is characterized by a state of insulin deficiency, while T2DM is characterized by insulin resistance. It is the most prominent disease related to failure of blood glucose regulation.	Serum, Plasma, Blood, Urine

- Not a single LOINC or SNOMED code mentioned …
- So, how can we find the data in the HIS or EHR ?

Annotated Protocol in XML
Short demo movie: Trial Summary data



Machine-readable SDTM-IG

- Students undergraduate project 2016-2017

- Generated an XML Structure for the SDTM-IG 3.2 content
  - Tables
  - Assumptions
  - Other metadata such as define.xml datatypes

Generated a stylesheet to display the content of the machine-readable SDTM-IG to humans in a browser

Machine-readable SDTM-IG: Results

```xml
<?xml version="1.0" encoding="UTF-8"?>

<!-- Domain Pharmacokinetics Parameters (PP)  -->
<SDTMClass Name="Findings" Version="3.2">
    <Domain ShortName="LB" Label="Laboratory Test Results">
        <DomainDescription>
            <TranslatedText xml:lang="en">Laboratory test findings including, but is not l
            include microbiology or
pharmacokinetic data, which are stored in separate domains.</TranslatedText>
        </DomainDescription>
        <Specification>
            <Structure>One record per lab test per time point per visit per subject, Tabul
<!--Start der Tabelle -->
        <VariableList>
                <Variable Name="STUDYID">
                    <VariableLabel>Study Identifier</VariableLabel>
                    <SASXPTDataType>Char</SASXPTDataType>
                    <RecommendedXMLDataType>string</RecommendedXMLDataType>
                    <Role>Identifier</Role>
                    <ControlledTerminology/>
                    <NCICodeList/>
                    <Core>Required</Core>
                    <CDISCNotes>Unique identifier for a study</CDISCNotes>
                    <Rules/>
                </Variable>
                <Variable Name="DOMAIN">
                    <VariableLabel>Domain Abbreviation</VariableLabel>
                    <SASXPTDataType>Char</SASXPTDataType>
```

Machine-readable SDTM-IG: Results

- And the human-readable content:
  - 100% identical in text content
  - >95% identical in presentation (HTML instead of PDF)

Class: Findings

Laboratory Test Results (LB)

LB - Description/Overview for the Laboratory Test Results Domain Model

Laboratory test findings including, but is not limited to hematology, clinical chemistry and urinalysis data. This domain does not include microbiology or pharmacokinetic data, which are stored in separate domains.

LB - Specification for the Laboratory Test Results Domain Model

lb.xpt, Laboratory Test Results - Findings, Version 3.2. One record per lab test per time point per visit per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, CodeList or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study	Required
DOMAIN	Domain Abbreviation	Char	LB	Identifier	Two-character abbreviation for the domain	Required
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Required
LBSEQ	Sequence Number	Num		Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Required
LBGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Permissible
LBREFID	Specimen ID	Char		Identifier	Internal or external specimen identifier. Example: Specimen ID.	Permissible
LBSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsorâ€™s operational database. Example: Line number on the Lab page.	Permissible
LBTESTCD	Lab Test or Examination Short Name	Char	(LBTESTCD)	Topic	Short name of the measurement, test, or examination described in LBTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. The value in LBTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g."1TEST"). LBTESTCD cannot contain characters other than letters, numbers, or underscores. Examples: ALT, LDH.	Required
LBTEST	Lab Test or Examination Name	Char	(LBTEST)	Synonym Qualifier	Verbatim name of the test or examination used to obtain the measurement or finding. Note any test normally performed by a clinical laboratory is considered a lab test. The value in LBTEST cannot be longer than 40 characters. Examples: Alanine Aminotransferase, Lactate Dehydrogenase.	Required
LBCAT	Category for Lab Test	Char	*	Grouping Qualifier	Used to define a category of related records across subjects. Examples: such as HEMATOLOGY, URINALYSIS, CHEMISTRY.	Expected
LBSCAT	Subcategory for Lab Test	Char	*	Grouping Qualifier	A further categorization of a test category such as DIFFERENTIAL, COAGULATON, LIVER FUNCTION, ELECTROLYTES.	Permissible
LBORRES	Result or Finding in Original Units	Char		Result Qualifier	Result of the measurement or finding as originally received or collected.	Expected

Machine-readable SDTM-IG
Why doesn't CDISC do this?

- SDTM-IG developers need an infrastructure to put the content in
- We cannot expect the SDTM-IG developers to write/edit XML
- SDTM-IG developers are used to work in Word
  - But latest SDTM-IG (v.3.3) was developed in Wiki/Jira environment
- Can we use Wiki/JIRA to generate the SDTM-IG in XML?
- Will SHARE deliver everything so that we don't need an IG?

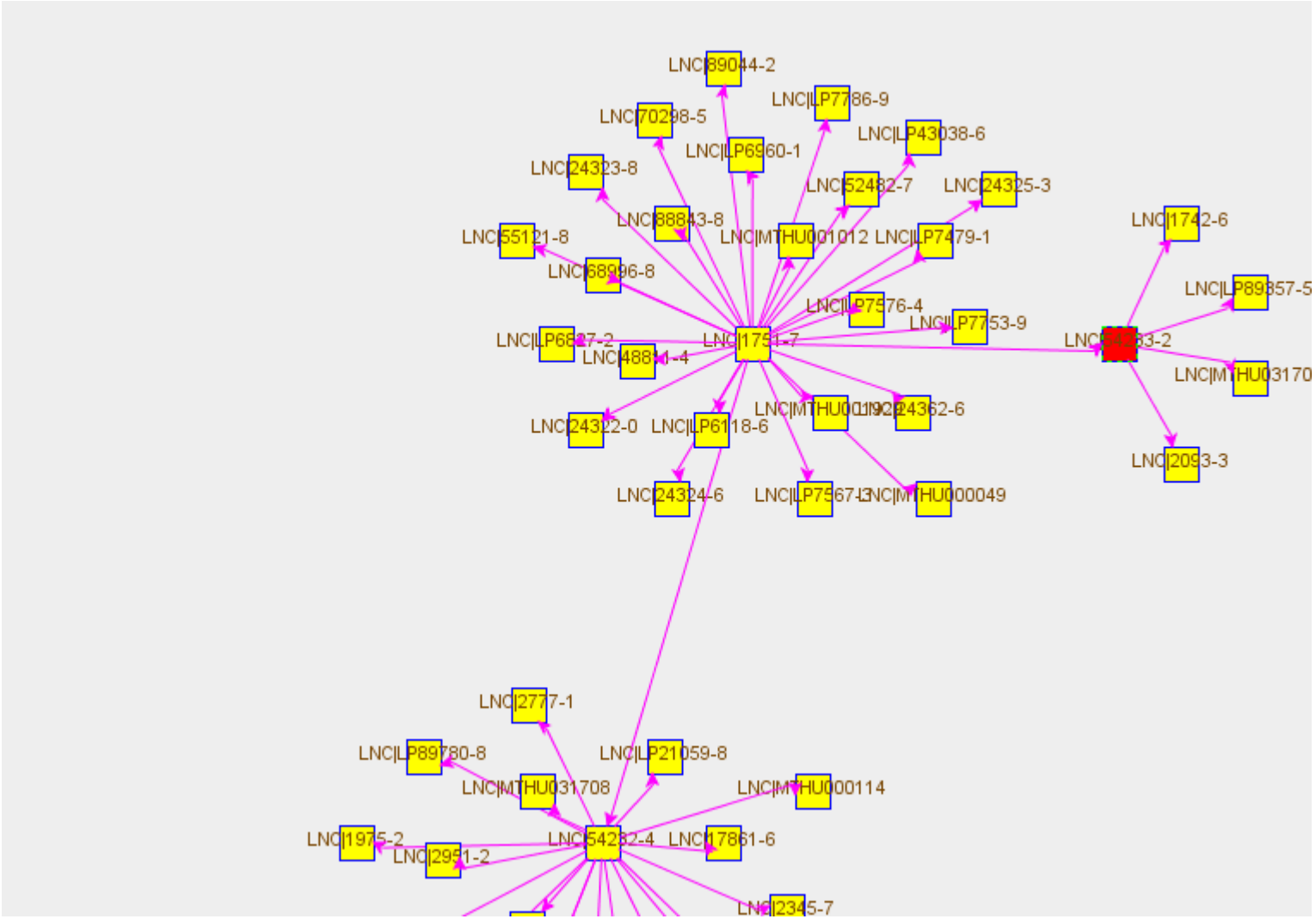
Connecting CDISC-CT to healthcare controlled terminology

- CDISC-CT is completely disconnected from (controlled) terminology used in healthcare (LOINC, SNOMED-CT, ICD-10, …)
- This makes it difficult to use information from electronic health records (EHRs) in clinical research
- Ideally, CDISC should give up some of its coding systems and use those from healthcare
- For the moment, we need … mapping

Mapping between CDISC-CT and Healthcare-CT

- Mapping between most used Laboratory LOINC Codes LBTESTCD / LBSPEC / LBMETHOD in development
- Can we automate things?
- Fortunately, we have UMLS (Unified Medical Language System)
  - Tries to connect terms between different medical coding systems (including NCI-CDISC)
  - An open RESTful Web Service is available
    - So we can use that in our tools

Connecting CDISC-CT to Healthcare CT



Connections between CDISC-CT "ALB" and LOINC codes and panels as used in healtcare IT

Movie available

And Jozef …

- If you do all this volunteer and research work, what are you living from?

- XML4Pharma provides
  - CDISC consultancy
  - Software for working with CDISC standards (ODM, Define.xml, SDTM)
  - Not for free, but not expensive either
  - Always with intelligent Graphical User Interfaces and many Wizards
  - 1000 times better than the crap software that is often offered for free by other companies