



2018
EUROPE
INTERCHANGE
BERLIN
23-27 APRIL



CDISC Standards in the age of Artificial Intelligence

FH-Prof. Dr. Jozef Aerts
University of Applied Sciences FH Joanneum
Institute of e-Health
Graz, Austria



A few facts ...

- CDISC Standards documents come as PDF / HTML
 - And thus are not machine-readable ...
 - Leading to many different interpretations ...
 - We even don't define what "**must**", "**should**", "**may**" mean
 - Other SDOs do so as the first thing in a specification
 - "Rules" are not machine-readable and are often confusing...

The following Qualifiers would not generally be used in VS:

- Protocols are not machine-readable

SDTM is difficult to learn ...

- It takes years to become an expert

November 29, 2017

Hi,

I would like to ask if somebody knows if in LB domain, is lab category (LBCAT) for **Alcohol Breathalyzer Results?**

Thank you,

Konrad

**Alexa, can VSDY
be 0 or negative?**

Although we do have standards Clinical Research remains highly **inefficient**

- FDA/PMDA can still not compare results between studies
- Study design and mapping to SDTM **reuse** remains limited
- Essentially, AI should be able to do 80% of the mapping
- There is still no "**Alexa for SDTM**"
 - Nor for other standards within CDISC

What is needed?

- Machine-readable / interpretable protocols
- Machine-readable standard specifications
- Clear (and machine-readable / executable) rules
 - And no "*would generally not be used ...*"
- Further:
 - Better coding systems, **interoperable** with healthcare

Without these, AI and ML
Will be hard to implement

Ok Jozef ...

***You are
complaining
again,
But what did
YOU do?***

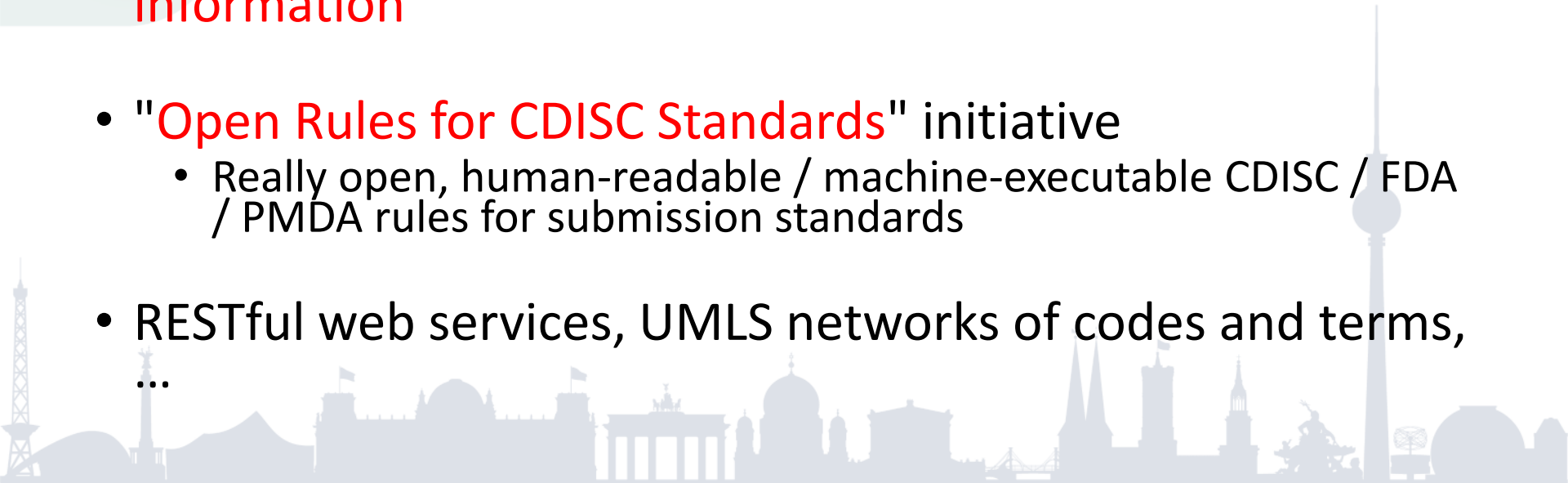
**CDISC
needs
you !!!**



What Jozef is doing ... - a few projects

- SDTM-IG in XML
- SDRG in XML (Phuse project)
- Annotating clinical research protocols with coded information
- "Open Rules for CDISC Standards" initiative
 - Really open, human-readable / machine-executable CDISC / FDA / PMDA rules for submission standards
- RESTful web services, UMLS networks of codes and terms,

...



The SDTM-IG in XML

- Although the SDTM-IGs are highly structured, they are still not machine readable
 - Move to HTML is not helpful ...
- Bachelor students project 2017
- Machine-readable IG for 46 SDTM-IG domains
- XSLT stylesheet reconstructing the "human view"

The SDTM-IG in XML: Results

```
1  <?xml version="1.0" encoding="UTF-8"?>
2
3  <!-- Domain Pharmacokinetics Parameters (PP) -->
4  <SDTMClass Name="Findings" Version="3.2">
5    <Domain ShortName="LB" Label="Laboratory Test Results">
6      <DomainDescription>
7        <TranslatedText xml:lang="en">Laboratory test findings including, but is not l
8        include microbiology or
9        pharmacokinetic data, which are stored in separate domains.</TranslatedText>
10     </DomainDescription>
11     <Specification>
12       <Structure>One record per lab test per time point per visit per subject, Tabul
13     </Specification>
14     <!--Start der Tabelle -->
15     <VariableList>
16       <Variable Name="STUDYID">
17         <VariableLabel>Study Identifier</VariableLabel>
18         <SASXPTDataType>Char</SASXPTDataType>
19         <RecommendedXMLDataType>string</RecommendedXMLDataType>
20         <Role>Identifier</Role>
21         <ControlledTerminology/>
22         <NCICodeList/>
23         <Core>Required</Core>
24         <CDISCNotes>Unique identifier for a study</CDISCNotes>
25         <Rules/>
26       </Variable>
27       <Variable Name="DOMAIN">
28         <VariableLabel>Domain Abbreviation</VariableLabel>
29         <SASXPTDataType>Char</SASXPTDataType>
```

The SDTM-IG in XML: Results

```
<Variable Name="LBSEQ">  
  <VariableLabel>Sequence Number</VariableLabel>  
  <SASXPTDataType>Num</SASXPTDataType>  
  <RecommendedXMLDataType>positiveInteger</RecommendedXMLDataType>  
  <Role>Identifier</Role>  
  <ControlledTerminology/>  
  <NCICodeList/>  
  <Core>Required</Core>  
  <CDISCNotes>Sequence Number given to ensure uniqueness of subject records within a  
    domain. May be any valid number.</CDISCNotes>  
  <Rules/>  
</Variable>
```

Variable definitions

```
<AssumptionSet>  
  <Assumption>For lab tests where the specimen is collected over time, i.e., 24-hour urine collection,  
    the start date/time of the collection goes into LBSTC and the end date/time of collection goes into LBENDTC.  
    See Section 4: 4.1.4.8, Date and Time Reported in a Domain Based on Findings.</Assumption>  
</AssumptionSet>
```

Assumptions

The SDTM-IG in XML: Human View (through stylesheet)

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 sponsor's TM 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

The SDTM-IG in XML - Future

- This is all still **extremely simple**
- "Rules" have only been added partially
- Assumptions are still "human text"
 - Part of it has been structured (discouraged variables)
 - But could already be interpreted by machines
- This is the way the SDTM team **SHOULD** publish the IG
 - And not as damned HTML or PDF
- It is a very first step only to come to an **"Alexa for SDTM"**

Poster in the poster session!

Annotated Protocols

- Protocols are still written using office software
 - "Templates" help to structure, but "that's it"
- Humans need to interpret the protocol and transform it to:
 - A study design
 - Submission data sets

[HOME](#) / [CDISC BLOG](#)

CDISC Blog

Can CDISC variables withstand a game of telephone?

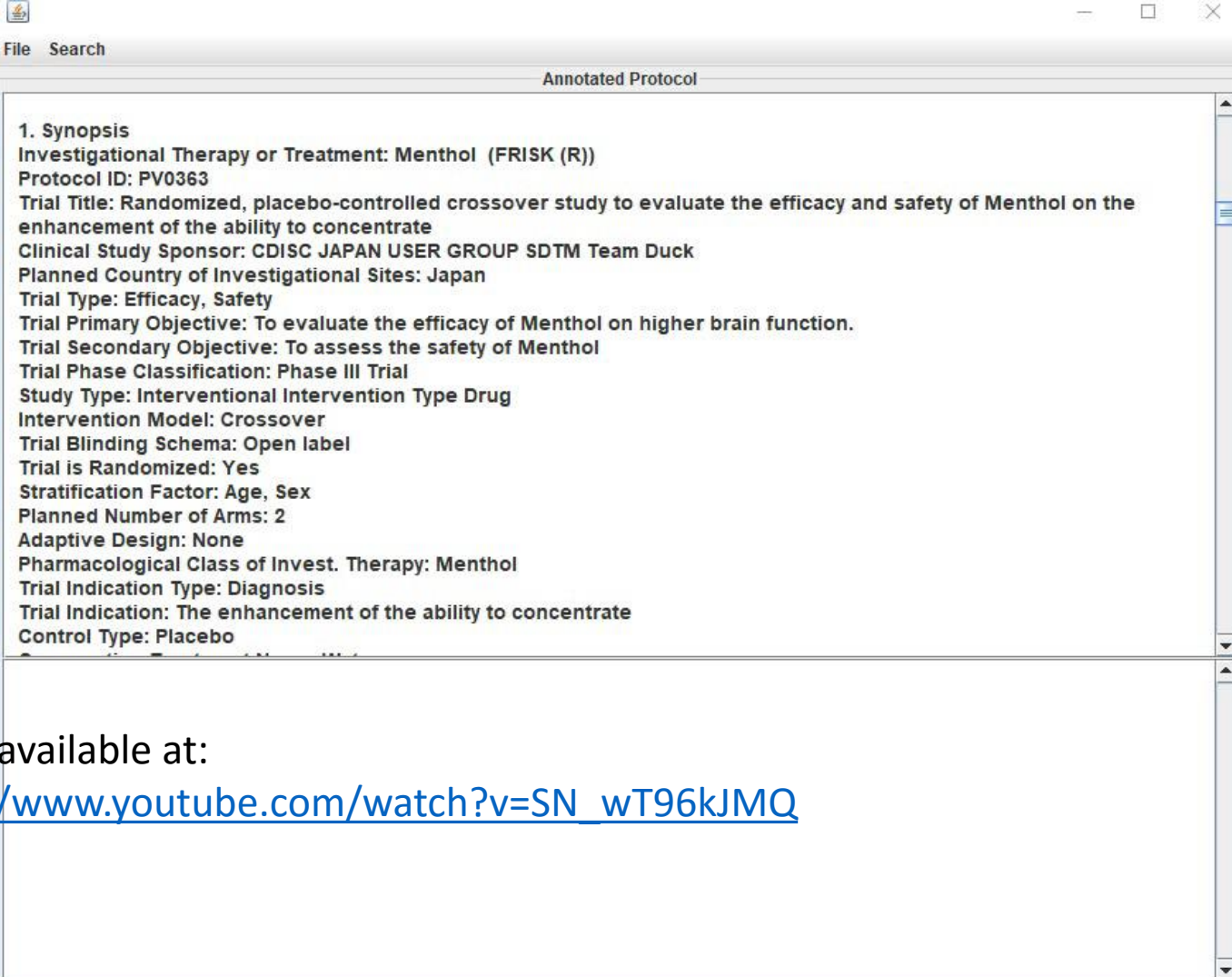
1 Feb 2018

by Anthony Chow, Sr. Manager, Technical Development, CDISC

Annotated Protocols

- A format and software tool was developed to annotate "narrative" protocols with codes and terms
 - SDTM Trial Design Parameters
 - => Automated generation of TS data sets
 - CDISC Controlled Terminology
 - LOINC, SNOMED-CT, ATC, ICD-10, UMLS, ...
 - Making it possible to use eSource and EHRs
- The "tool" uses RESTful web services for suggesting suitable codes and terms for protocol text snippets

Annotated Protocols - Movie



The screenshot shows a software window titled "Annotated Protocol" with a menu bar containing "File" and "Search". The main content area displays the following text:

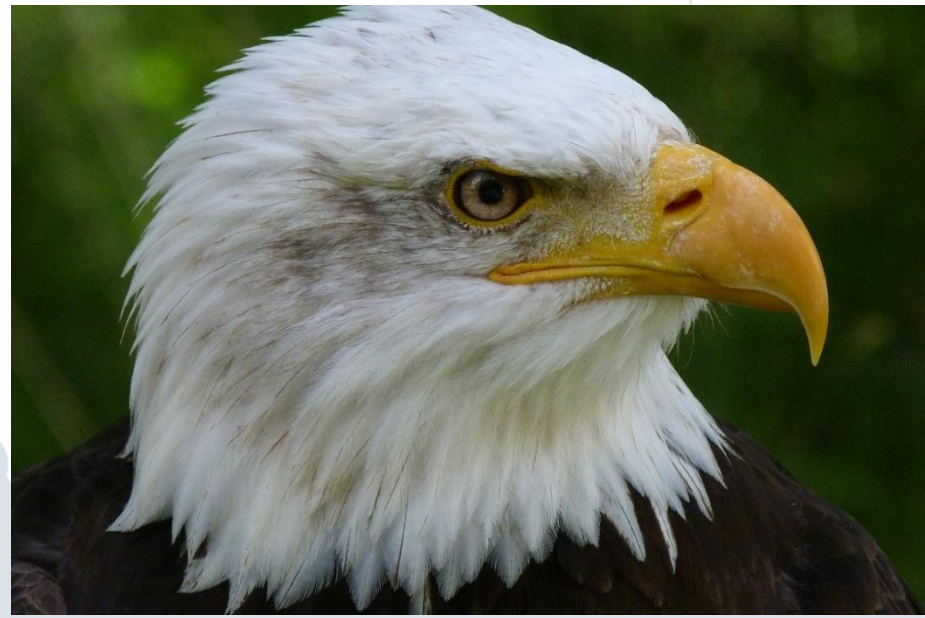
1. Synopsis
Investigational Therapy or Treatment: Menthol (FRISK (R))
Protocol ID: PV0363
Trial Title: Randomized, placebo-controlled crossover study to evaluate the efficacy and safety of Menthol on the enhancement of the ability to concentrate
Clinical Study Sponsor: CDISC JAPAN USER GROUP SDTM Team Duck
Planned Country of Investigational Sites: Japan
Trial Type: Efficacy, Safety
Trial Primary Objective: To evaluate the efficacy of Menthol on higher brain function.
Trial Secondary Objective: To assess the safety of Menthol
Trial Phase Classification: Phase III Trial
Study Type: Interventional Intervention Type Drug
Intervention Model: Crossover
Trial Blinding Schema: Open label
Trial is Randomized: Yes
Stratification Factor: Age, Sex
Planned Number of Arms: 2
Adaptive Design: None
Pharmacological Class of Invest. Therapy: Menthol
Trial Indication Type: Diagnosis
Trial Indication: The enhancement of the ability to concentrate
Control Type: Placebo

Movie available at:

https://www.youtube.com/watch?v=SN_wT96kJMQ

Annotated Protocols

- Such annotated protocols are an "easy prey" for ML systems
 - Automated Study Design generation (in a consistent way)
 - Ideally in combination with MDRs
 - LOINC / SNOMED-CT coding
=> BCs
- Limitations
 - "Schedule of Events"
 - => should be replaced by "workflows"



The Schedule of ... disaster

So when we move to an electronic form of the SoA, I would suggest that the machine can maybe store the design in more novel ways. We can present it in this tabular form back to users should they need that presentation. The one comment from the Copenhagen event was that investigators like this form. I can understand that, it provides an overview and is readily consumable by the human.

But it's hopeless for a machine. I think we could look at a more timeline type of definition that can be used for multiple purposes including much more rapid deployment of studies into collection systems, while also supporting such presentation in protocol document and elsewhere.

Topic	Variable	Year							
		0	1	2	3	4	5/7	6/8	
		Pre-V	V1a	V1b	V2	V3	V4	V5/V7	V6/V8
	Eligibility Form (Inclusion & Exclusion Criteria)	•							
	Consent Form and Study Brochure	•							
	Family Binder	•	•	•	•	•	•	•	•
Kidney	Isotonic-based GFR	X		X	X	X	X	X	X
	Cystatin C	X		X	X	X	X	X	X
	Serum Creatinine	X		X	X	X	X	X	X
	Central Renal Panel*	X		X	X	X	X	X	X
	Central Uric Acid*	X		X	X	X	X	X	X
	Central Urine Creatinine and Protein	X		X	X	X	X	X	X
	Central Urine Albumin	X		X	X	X	X	X	X
	Local Complete Blood Count*	X		X	X	X	X	X	X
	Local Pregnancy Tests*	X		X	X	X	X	X	X
	Local Renal Panel*	X		X	X	X	X	X	X
	Local Urine Creatinine and Urine Protein*	X		X	X	X	X	X	X
Cardiovascular	Classical Blood Pressure (centrally collected)	•	•	•	•	•	•	•	•
	Classical Blood Pressure (locally measured)	•	•	•	•	•	•	•	•
	Lipid Profile			•	•	•	•	•	•
	Ambulatory Blood Pressure Monitoring			•	•	•	•	•	•
	Echocardiography			•	•	•	•	•	•
	Carotid Intima-Media Thickness*			•	•	•	•	•	•
	Pulse Wave Velocity*			•	•	•	•	•	•
Neurocognitive	Cardiac Magnetic Resonance Imaging (CMRI)		•	•	•	•	•	•	•
	Pediatric Quality of Life		•	•	•	•	•	•	•
Growth	Cognitive and Development Assessments		•	•	•	•	•	•	•
	Behavioral Assessments		•	•	•	•	•	•	•
	Height/Length and Weight	•	•	•	•	•	•	•	•
	Head Circumference*	•	•	•	•	•	•	•	•
	Mid-Arm Circumference*	•	•	•	•	•	•	•	•
	Waist and Hip Circumference*	•	•	•	•	•	•	•	•
	Tanner Stage	•	•	•	•	•	•	•	•
	Frost Pregnancy Questionnaire (FPQ)	•	•	•	•	•	•	•	•
	Intact Parathyroid Hormone (iPTH)	•	•	•	•	•	•	•	•
	High Sensitivity CKP (hsCKP)	•	•	•	•	•	•	•	•
	Vitamin D	•	•	•	•	•	•	•	•
	Pubertal Growth Factor-23 (PGF-23)	•	•	•	•	•	•	•	•
6 Minute Walk Test (6MWT)			•	•	•	•	•	•	
Grip Strength			•	•	•	•	•	•	

<https://www.a3informatics.com/phuse-single-day-event-copenhagen/>

Presentation versus data

- In CDISC we are constantly mixing up "presentation" with "data" (and "information")
- We use technologies meant for "presentation" (e.g. **tables**) and think they are the "data"
- Everything that we have is "flattened" to "Wiener Schnitzel" => Quality loss
- We must learn to **return to first principles** and generate machine-readable standards



Open Rules for CDISC Standards

- Current validation rules & software:
- Have been "hijacked" by regulatory authorities and a for-profit company
- Are over-interpretations of the IGs
- Are often completely incorrectly implemented in software
 - Extremely many "false positives"

Open Rules for CDISC Standards

- New initiative to publish CDISC (and FDA/PMDA?) rules in **machine-executable** as well as human-readable format
- Can be used in **any** modern software
- Are owned by the CDISC community
- Can be written in the machine-readable IGs itself
- New formal CDISC project

Other things we are working on

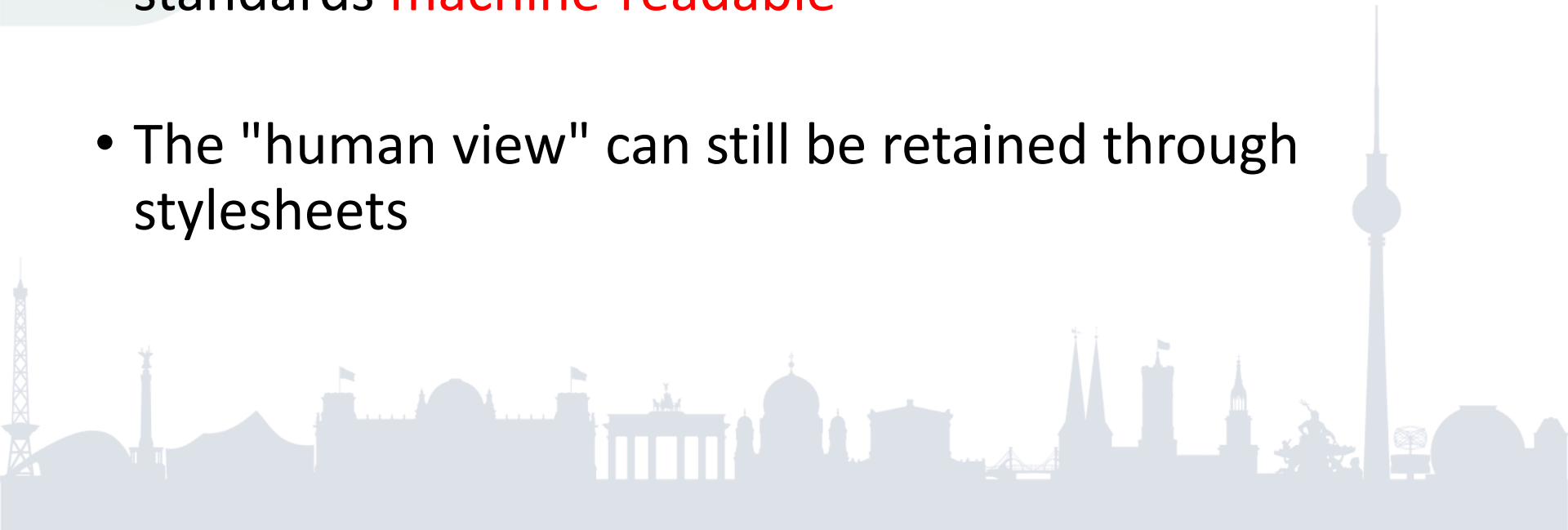
- ML (Word2Vec) on ClinicalTrials.gov entries
- Workflows in e-Protocol
- Protocol elements in ODMv2

Looking for
Funding!



Conclusions

- Using Standards creates huge potential to enormously increase efficiency through ML and AI
- In order to use this potential, we need to make our standards **machine-readable**
- The "human view" can still be retained through stylesheets



The "Alexa for SDTM"

*Well Jozef,
VSDY can be negative,
but it cannot be zero,
because ...*

