phuse•CONNECT
CONNECT SHARE ADVANCE

EU 2018

4th–7th November
KAP Europa
Frankfurt, Germany

The Clinical Data Science Conference

From Machine-readable CDISC Standard Specifications to the e-Protocol

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

FH | JOANNEUM
University of Applied Sciences

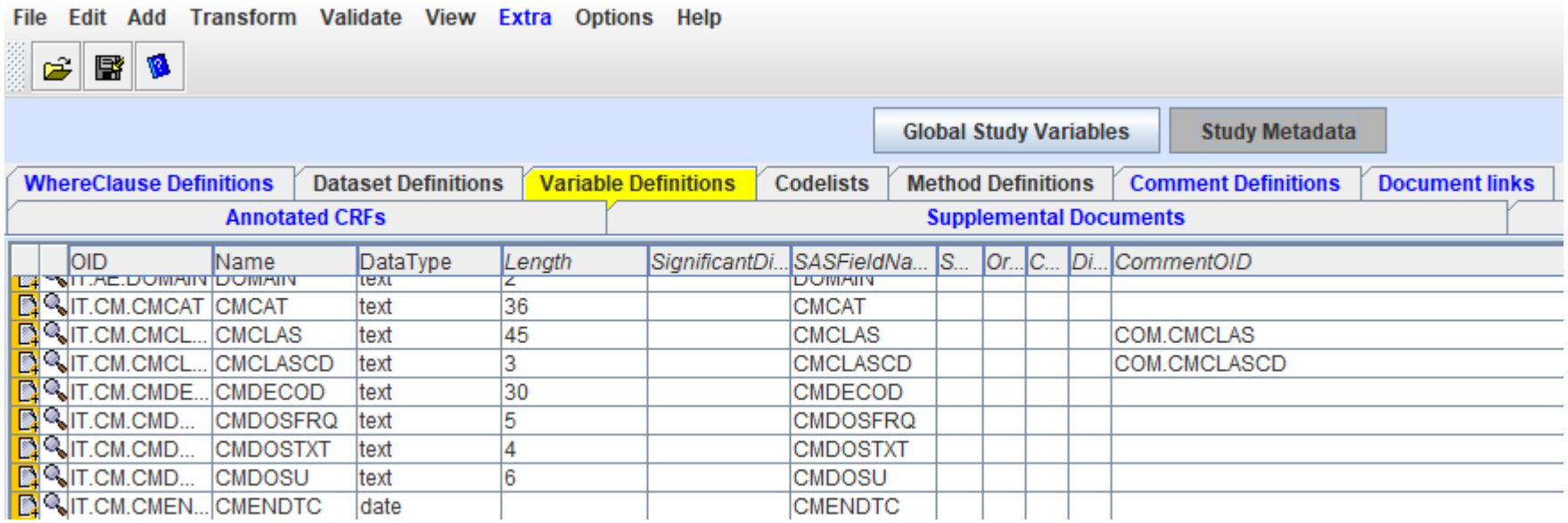
cdisc

The problems

- 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…
    - Validation tools (even those used by the FDA) are based on overinterpretation, misinterpretation and own-invented-interpretation of the standard
    - Often leading to many "false positives"
- Protocols are not machine-readable

The problems

- Only a few CDISC standards partially do have a machine-readable specification
- Example: Define-XML through XML-Schema and Schematron
- Allow to automate tool development and validation



OID	Name	DataType	Length	SignificantDi...	SASFieldNa...	S...	Or...	C...	Di...	CommentOID
IT.AE.DOMAIN	DOMAIN	text	2		DOMAIN					
IT.CM.CMCAT	CMCAT	text	36		CMCAT					
IT.CM.CMCL...	CMCLAS	text	45		CMCLAS					COM.CMCLAS
IT.CM.CMCL...	CMCLASCD	text	3		CMCLASCD					COM.CMCLASCD
IT.CM.CMDE...	CMDECOD	text	30		CMDECOD					
IT.CM.CMD...	CMDOSFRQ	text	5		CMDOSFRQ					
IT.CM.CMD...	CMDOSTXT	text	4		CMDOSTXT					
IT.CM.CMD...	CMDOSU	text	6		CMDOSU					
IT.CM.CMEN...	CMENDTC	date			CMENDTC					

The problems

- Our CDISC Controlled Terminology is completely disconnected from CT used in healthcare-IT
  - How the hell can we retrieve information from EHRs when we use completely different CT?
- We have even invented our own notation for units
  - Not used in healthcare-IT nor anywhere else in the world.
  - Not suitable for unit conversion calculations

Ok Jozef …

You are complaining again,
But what did <u>YOU</u> do?

What Jozef is doing … - a few projects

- SDTM-IG in XML
- SDRG in XML (Phuse project)
- UCUM Units conversion and validation
- 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
- UMLS Controlled Terminology Explorer
- SHARE API 2.0 Implementations
- RESTful web services for CDISC standards and CT

The SDTM-IG in XML

- Although the SDTM-IGs are highly structured, they are still not machine readable
  - Move to HTML is <u>not</u> 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

```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>
```

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 LBDTC 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 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

Specification of the LB domain
Human-readable VIEW

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"

UCUM units validation and conversion

- Unified Code for Units of Measure (UCUM) is THE notation used for units in healthcare-IT
- CDISC still refuses to allow usage of UCUM notation in SDTM
  - Has "invented" its own terminology
- UCUM essentially allows conversion between ANY unit (for the same property)
- A RESTful web service was developed for conversions and validations
- Has been donated to and is now run at the National Library of Medicine server

CDISC SEND example:

UCUM notation for the measurement in the US: `[us_gal]{waterconsumption}/([ft_i]2.{chicken}.[oz_av]{food}.h)`

UCUM notation for the measurement in Europe: `l{waterconsumption}/(m2.{chicken}.g.{food}.d)`

UCUM units validation and conversion

https://ucum.nlm.r 120% ucum conve

NIH U.S. National Library of Medicine

LHNCBC

Unified Code for Units of Measure (UCUM)

UCUM Resources from the Lister Hill National Center for Biomedical Communications

UCUM Web Service

This is a set of web services (APIs) for programs to use when working with units from the Unified Code for Units of Measure (UCUM) system. These are the same APIs as those that are running at xml4pharmaserver.com, and are based on that website's web service code which has been donated to the U.S. National Library of Medicine by FH-Prof. Jozef Aerts and Mr. Milos Ilic MSc, Institute of eHealth, University of Applied Sciences FH Joanneum in Graz Austria.

Currently, three web services are available:

- UCUM unit Conversion web service
- UCUM unit validation web service
- UCUM unit to base units conversion

UCUM Service version: 2.1.2 (changes)

https://ucum.nlm.nih.gov/ucum-service.html

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
  - CRFs
  - Lab instructions
  - Submission data sets …

  - CRFs
  - Trial Design datasets
  - Clinical Trial Registry entries
  - And the results to SDTM and ADaM

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

File Search

Annotated Protocol

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

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"

https://www.a3informatics.com/biomedical-concepts/

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"

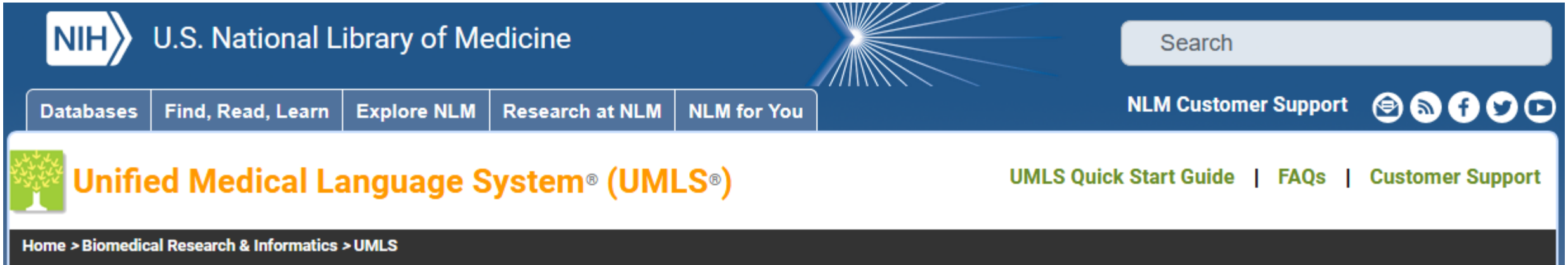
Open Rules for CDISC Standards

- New initiative to publish CDISC (and FDA/PMDA?) rules in machine-executable as well as human-readable format
  - Envisaged to become THE reference implementation
- Can be used in any modern software
  - By any vendor or organization
- Are owned by the CDISC community
- Can be written in the machine-readable IGs itself
- New formal CDISC project

More information coming soon …

UMLS Controlled Terminology Explorer

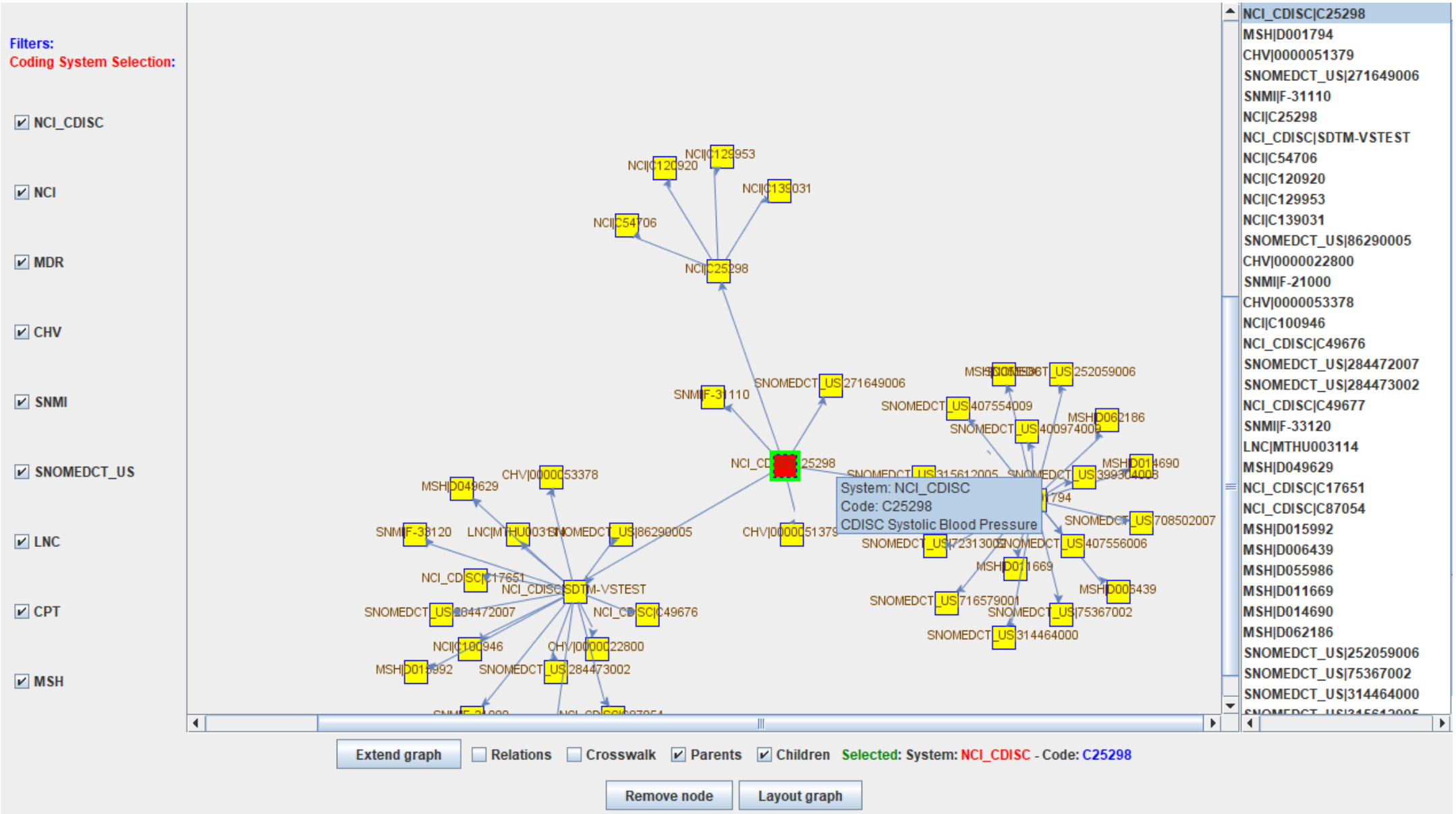
- CDISC-CT is completely disconnected from healthcare-CT
  - EHRs do NOT use CDISC-CT
- CDISC-CT does almost not describe any relations between terms
  - SYSBP with DIABP has the same relation as SYSBP with HEIGHT
- But we have Unified Medical Language System UMLS!
  - Tries to describe relations between all coding systems in the medical world
  - Owned and maintained by the NLM



UMLS Controlled Terminology Explorer

- We are currently developing a software tool to explore relationships between CDISC-CT and CT from healthcare-IT
- Based on UMLS RESTful Web Services
- Generates graphs of relationships
  - Leading to a "knowledge network"
- Still a lot "todo", but it works …

UMLS Controlled Terminology Explorer



A few more CDISC-related projects Jozef is working on …