KNOWLEDGE GRAPHS FOR PHARMA

A PERSPECTIVE FROM THE PhUSE PROJECT

CLINICAL TRIALS DATA AS RDF

- Tim Williams

Pistoia Alliance

2019-01-24

WHO I AM

- Statistical Systems Analyst
- Raleigh, North Carolina

PhUSE

- Steering Committee, Computational Sciences Symposium (CSS)
- Co-lead, Clinical Trials Data as RDF*
- Co-lead, Analysis Results Model (RDF Data Cubes) (2016)
- Instructor, Linked Data Hands-on Workshop*

Perspective: Late Phase Clinical Trials

Pharmaceutical Users Software Exchange

- Membership: >8,700 spanning 30 countries
- Annual Conferences:
  - EUConnect
  - USConnect
- Single Day Events
- Computational Sciences Symposium (CSS)
  - A *working* conference

PhUSE SEMANTIC WEB (LINKED DATA) PROJECTS

Recent Work:

- **CDISC Foundational Standards in RDF**
- CDISC Conformance Checks
- Reusing Medical Summaries for Enabling Clinical Research
- Analysis Results and Metadata (RDF Data Cube)
- Regulatory Guidance in RDF
- Clinical Program Design in RDF
- CDISC Protocol Representation Model in RDF

PhUSE SEMANTIC WEB (LINKED DATA) PROJECTS

Today's Discussion

- Clinical Trials Data as RDF
- New Project: Includes Non-clinical + Clinical

Pharmaceutical Users Software Exchange

Mission:

Provide a welcoming, neutral platform for creating and sharing ideas... exploring innovative methodologies, techniques, and technologies.

Working Groups Mission:

...open, transparent, and collaborative forum in a non-competitive environment

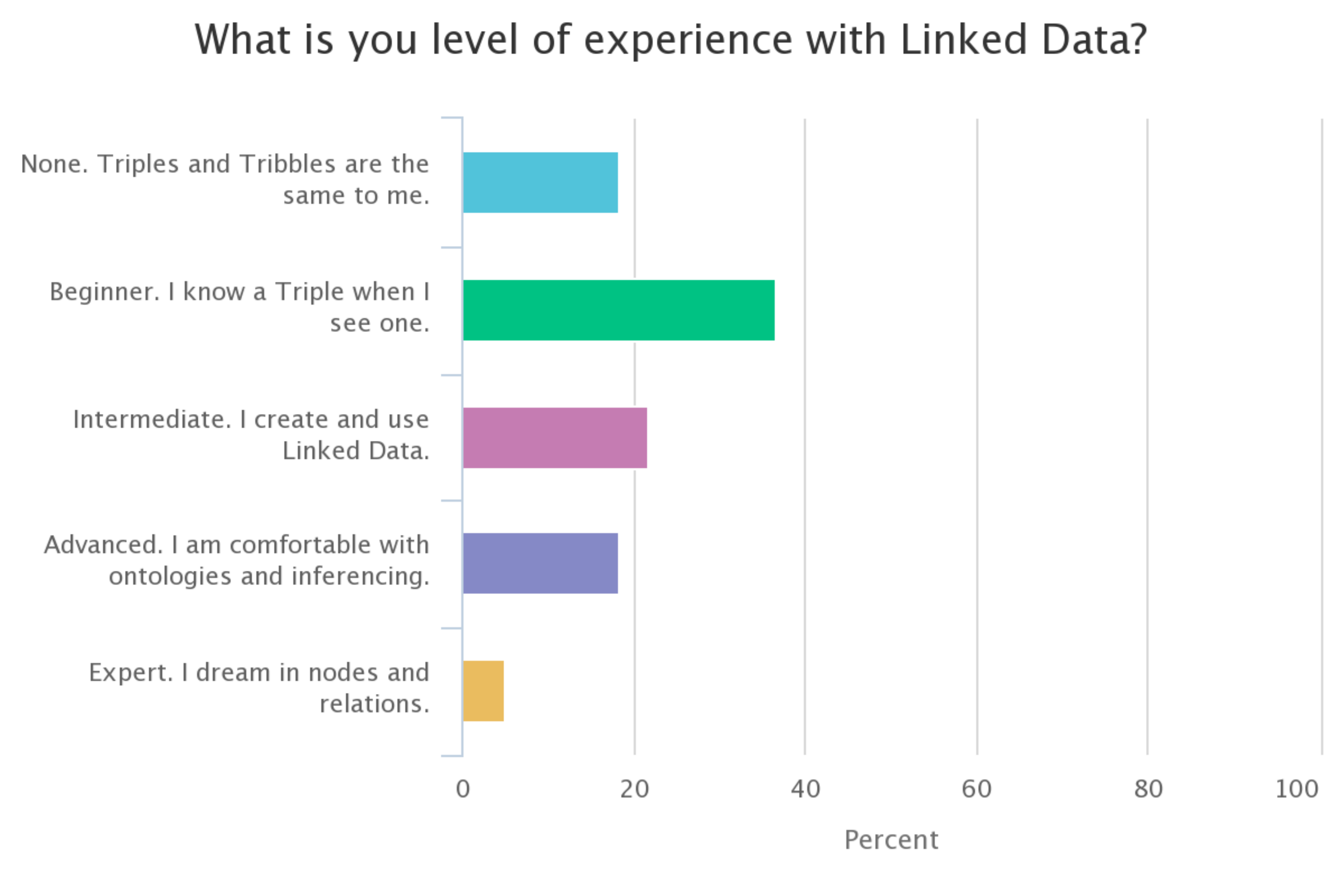
Pistoia Alliance

MISSION STATEMENT

To lower barriers to innovation in Life Sciences R&D through pre-competitive collaboration

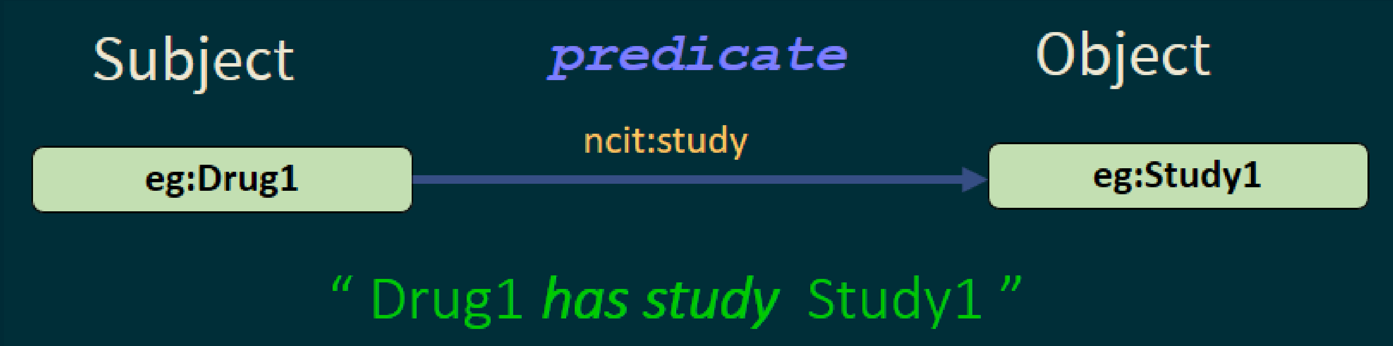
POLL

WHAT IS YOUR LEVEL OF EXPERIENCE WITH LINKED DATA?



TERMINOLOGY

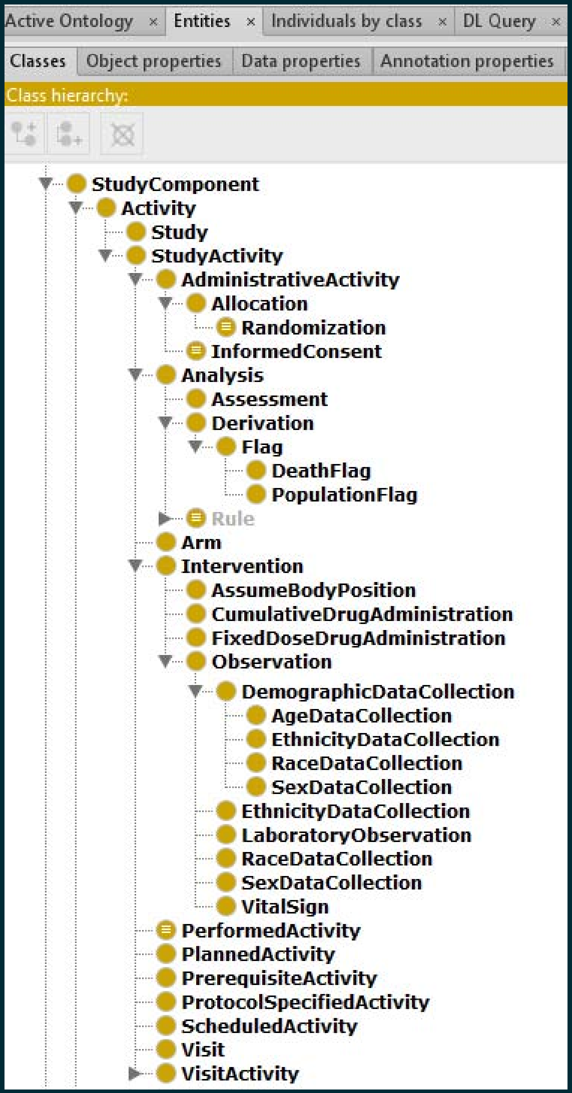
RESOURCE DESCRIPTION FRAMEWORK (RDF)



- Unique identifiers
- Define once, use many

TERMINOLOGY: KNOWLEDGE GRAPH*

RDF
+
Ontology
+
Reasoner (inferencing)
+
Rules
(SPIN, SHEX, SHACL)



- StudyComponent
  - Activity
    - Study
    - StudyActivity
      - AdministrativeActivity
        - Allocation
          - Randomization
        - InformedConsent
      - Analysis
        - Assessment
        - Derivation
          - Flag
            - DeathFlag
            - PopulationFlag
        - Rule
      - Arm
      - Intervention
        - AssumeBodyPosition
        - CumulativeDrugAdministration
        - FixedDoseDrugAdministration
        - Observation
          - DemographicDataCollection
            - AgeDataCollection
            - EthnicityDataCollection
            - RaceDataCollection
            - SexDataCollection
          - EthnicityDataCollection
          - LaboratoryObservation
          - RaceDataCollection
          - SexDataCollection
          - VitalSign
      - PerformedActivity
      - PlannedActivity
      - PrerequisiteActivity
      - ProtocolSpecifiedActivity
      - ScheduledActivity
      - Visit
      - VisitActivity

* my definition

FOUNDATIONS FOR THE CTDasRDF PROJECT

5 STAR OPEN DATA PRINCIPLES

Web, open license, +/- format

Structured, machine readable

Non-proprietary format

URIs

Linked to other data

@NovasTaylor

F.A.I.R DATA PRINCIPLES

- Findability
  - F1. globally unique, persistent id
  - F2. rich metadata
  - F3. searchable source
  - F4. metadata specify data id
- Accessibility
  - A1. retrievable by id using standard protocol
  - A1.1 protocol open, free, universal
  - A1.2 protocol allows authentication
  - A2 metadata avail. when data is not
- Interoperability
  - I1. formal, accessible, shared, broadly applicable language
  - I2. uses FAIR vocabularies
  - I3. qualified references to other data
- Reusability
  - R1. plurality of accurate and relevant attributes
  - R1.1 clear and accessible usage license
  - R1.2 provenance
  - R1.3 meets domain-relevant standards

More about F.A.I.R @ Pistoia Alliance: Ready, Set, GoFAIR 31 July 2018

https://www.pistoiaalliance.org/pistoia-alliance-debates-webinar-series/

HOW DOES PHARMA FARE ON F.A.I.R.?

My view from late-phase clinical trials

- Findability
  - F1. globally unique, persistent id

    Human Study Subject "Bob"
    - PharmaCo
      - Study 1, Drug A
      - Study 2, Drug B
    - DrugCo
      - Study 3, Drug C

Merge Bob's data.

HOW DOES PHARMA FARE ON F.A.I.R.?

- Accessibility
  - A2 metadata available when data is not

    Data changes form during its journey from collection to analysis.

    Biostatisticians and Medical Writers do not have easy access to the metadata from data collection and transformation process.

❌

HOW DOES PHARMA FARE ON F.A.I.R.?

- Interoperability
  - I1 : shared language for knowledge
    - Lack : language
    - Lack : representation
    - F.A.I.R. adoption is helping!

  Core Challenge
  - Current: Modeled our data to the industry standards for submission to regulatory authorities
  - Future: developing models of the process and the entities in the data.

✗

HOW DOES PHARMA FARE ON F.A.I.R.?

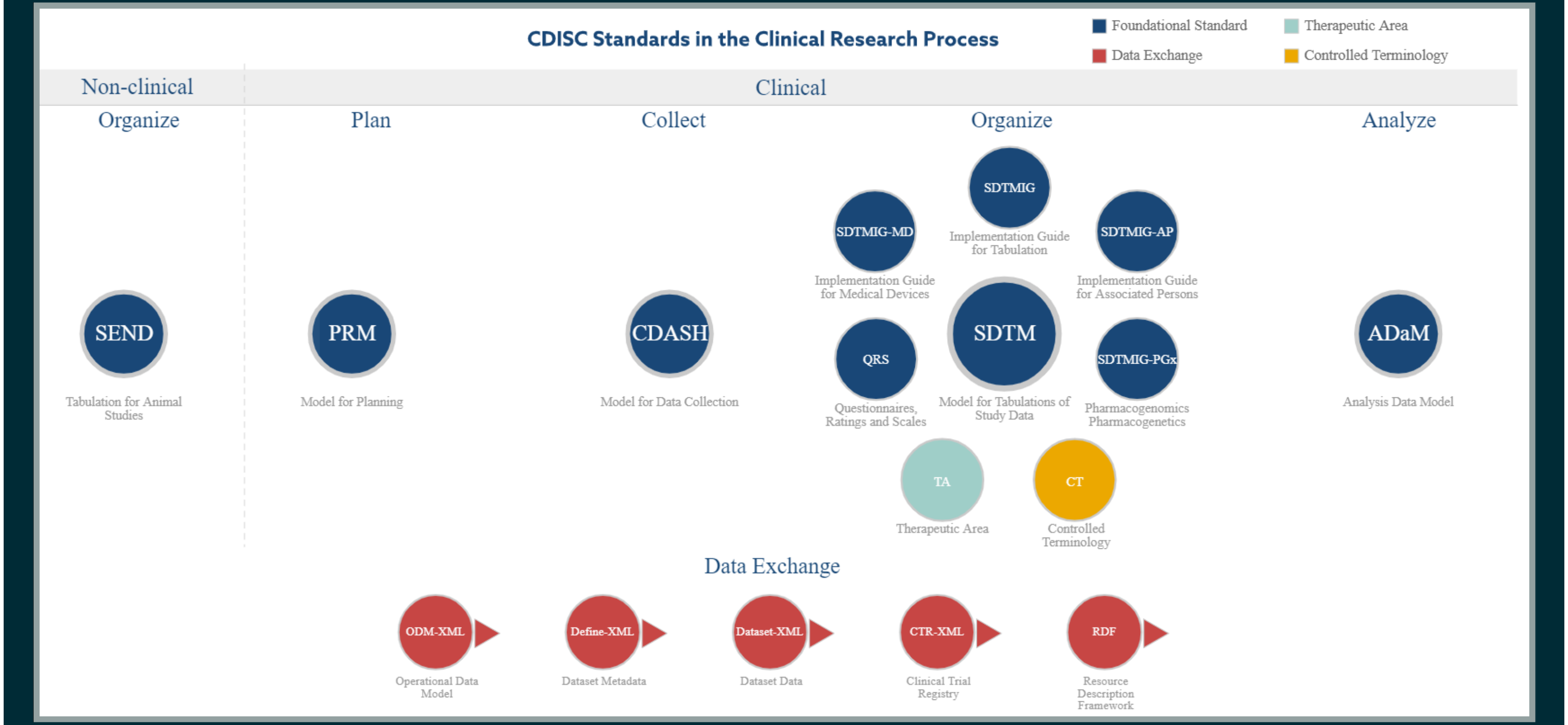
- Reusability
  - R1.3 : data meet domain-relevant community standards

      Positive:
    - We have standards!

      Negative:
    - We have standards! - That must be improved...

Clinical Data Interchange Standards Consortium
~1997

www.cdisc.org
Standards Overview



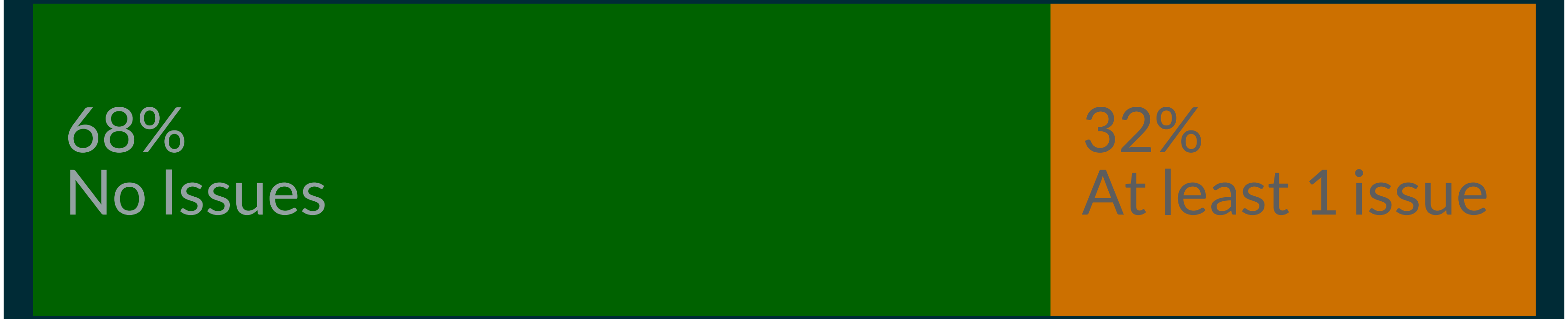
https://www.cdisc.org/standards

STANDARD FOR EXCHANGE OF NONCLINICAL DATA (SEND)

> *Beyond SEND: Leveraging Nonclinical Data to Drive Translational Research Forward*
> *- Pistoia Alliance*

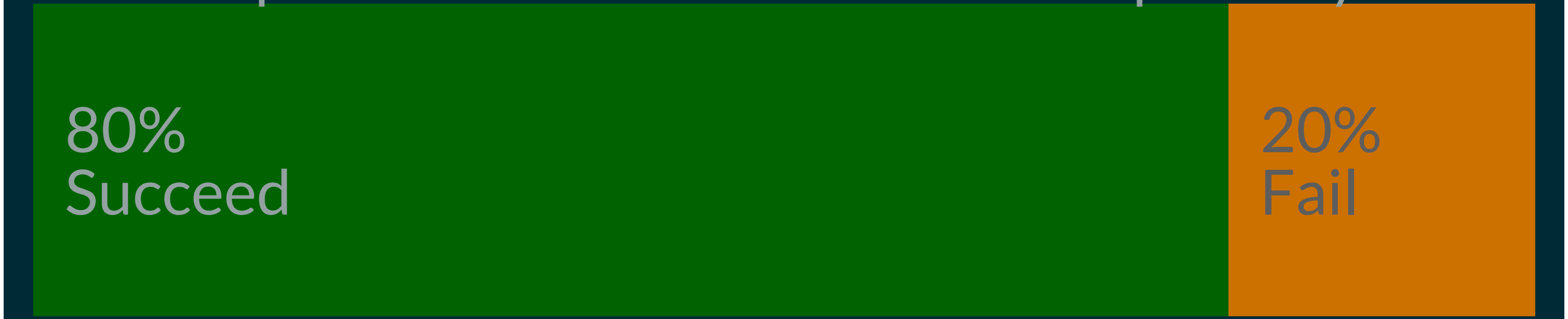
SUBMISSIONS TO FDA

Conformance Criteria



SUBMISSIONS TO FDA

Uploads to Janus Clinical Trials Repository



Why is this happening?

CDISC SDTM DOMAINS

STUDY DATA TABULATION MODEL

> *"A standard structure for data submitted to a regulatory authority."*
>
> *- https://en.wikipedia.org/wiki/SDTM*

> *"23 defined domains within six broad categories."*
>
> *- SDTM 3.1*

- Demographics (DM)
- Vital Signs (VS)
- Adverse Events (AE)
- ...

SELECT PROBLEMS IN CDISC SDTM

> *"Domains represent discrete categories" - CDISC*

Reality: They do not.

- Example: Demographics domain (DM)
  Also contains
  - Study ID
  - Treatment Arm Information (arm, coded value for arm)

SELECT PROBLEMS IN CDISC SDTM

- Multiple approaches to represent Medical conditions
  - Medical History (MH)
  - Adverse Events (AE)
  - Clinical Events (CE)

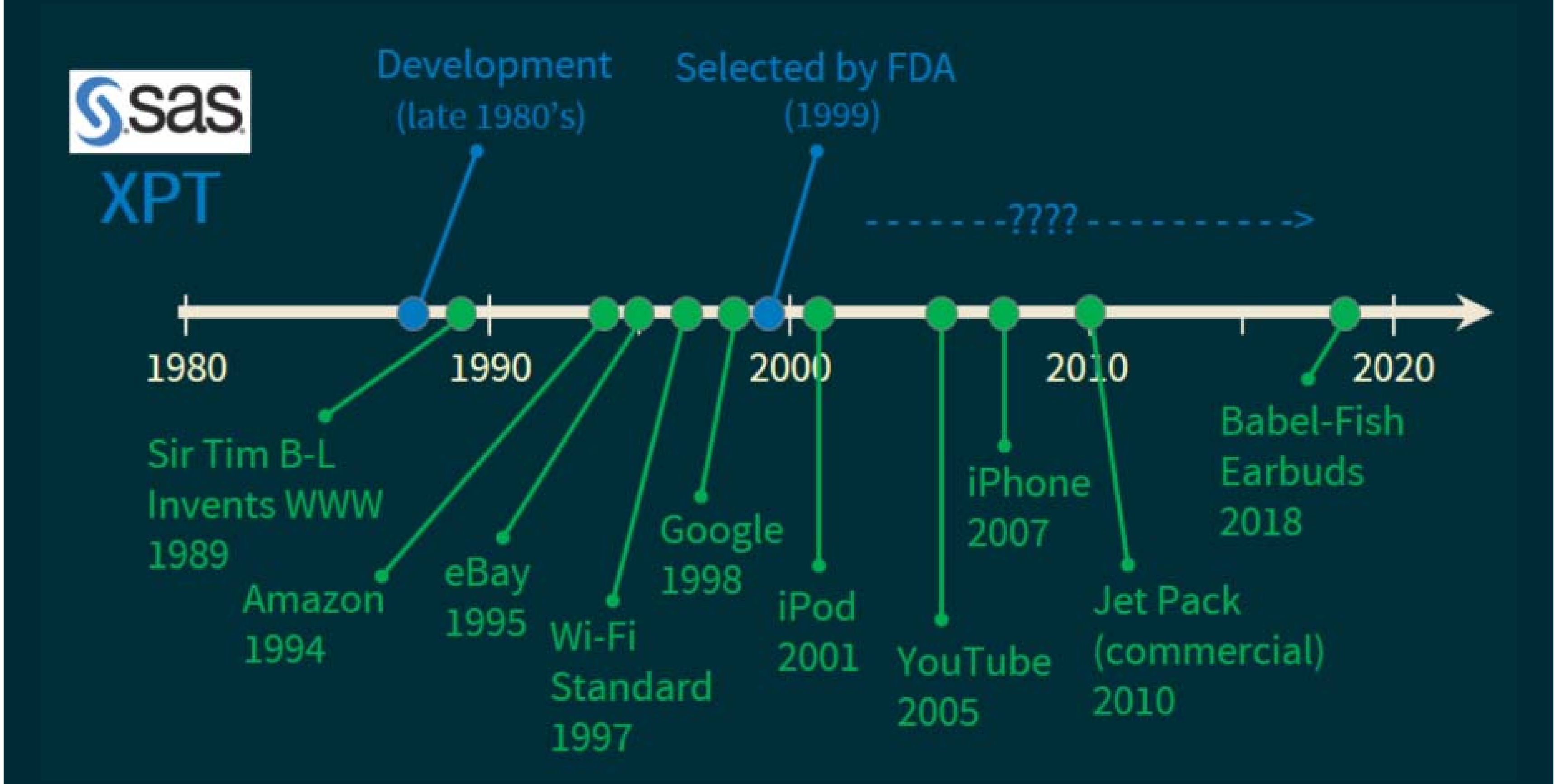
- Multiple locations for same/similar information

  Death Information:
  - Demographics (DM)
  - Disposition (DS)
  - Adverse Events (AE)

A consequence of row-by-column files.

30-YEAR-OLD FILE TRANSFER FORMAT

SDTM AND ADaM DATA SUBMISSION TO FDA

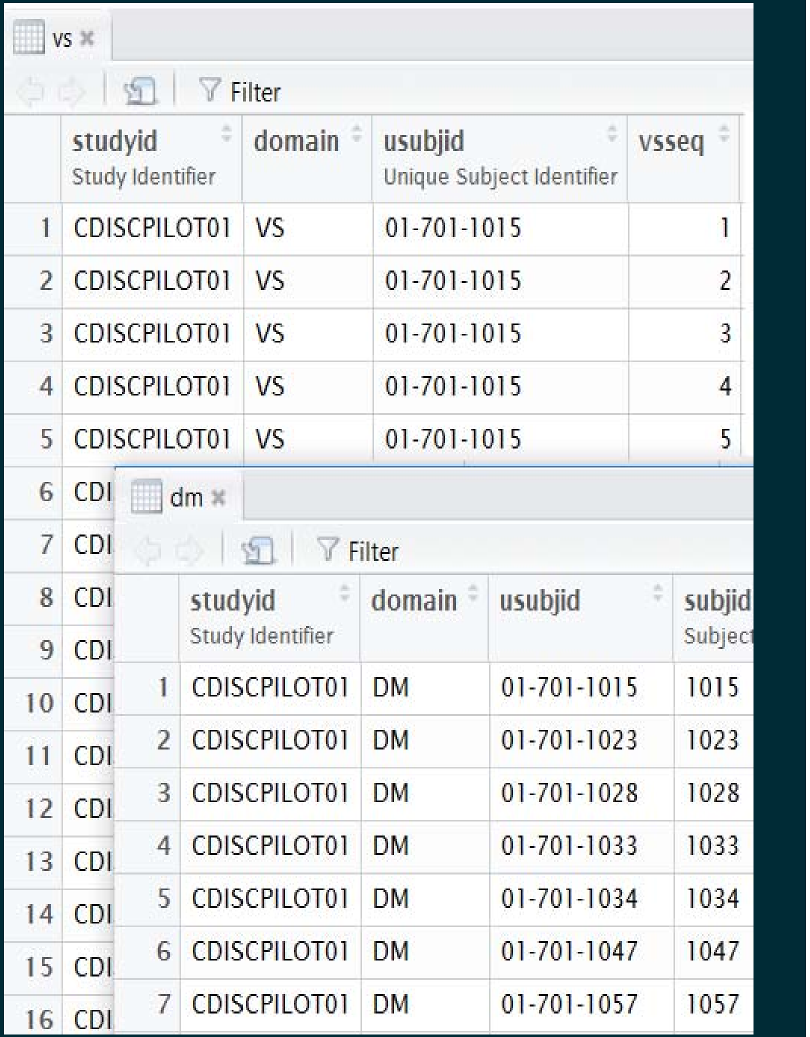


...also used for DATA <u>STORAGE</u>

XPT DATA STORAGE

XPT becomes a source for:

- Submission
- Publication
- Data Pooling

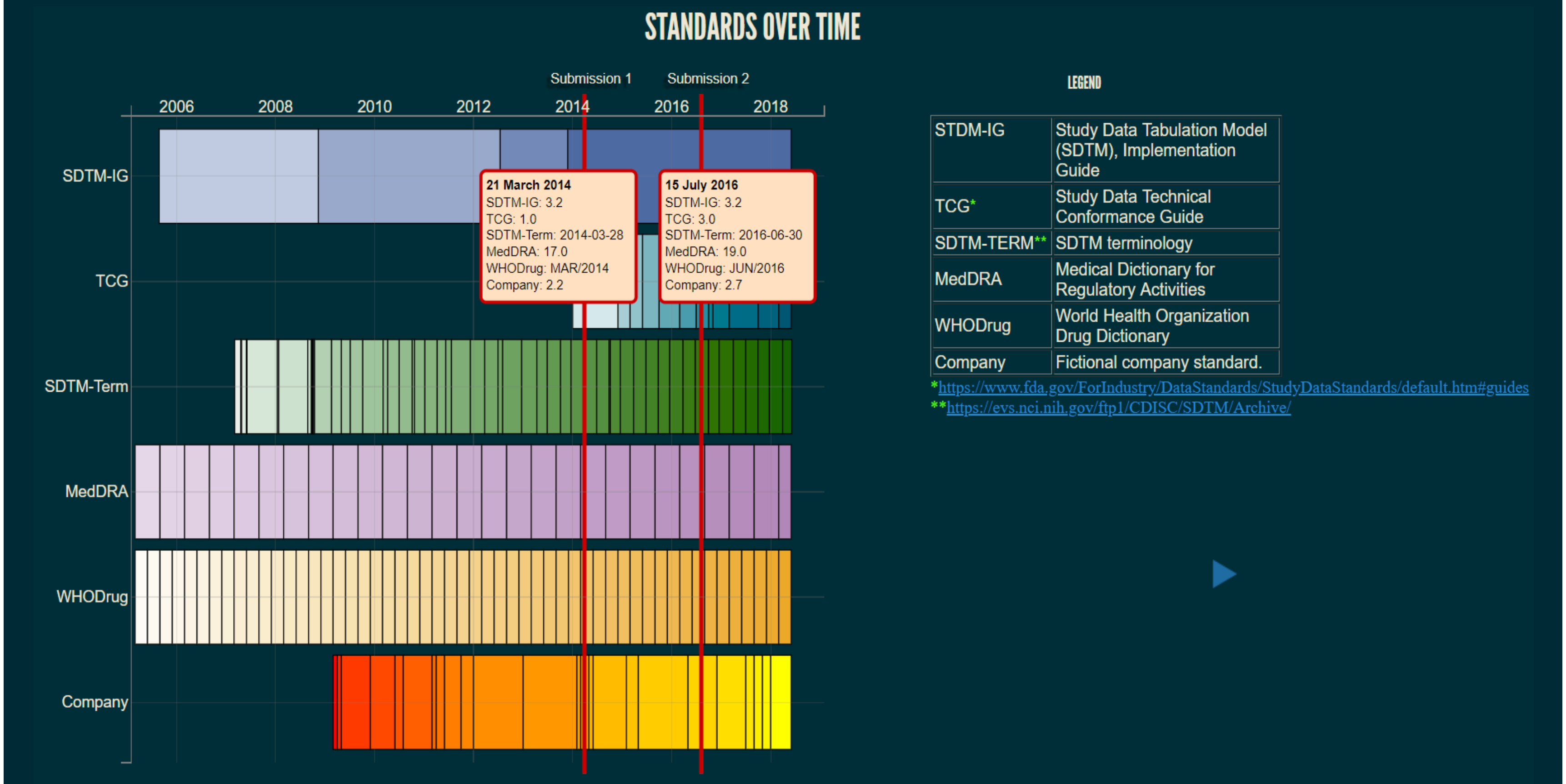


	studyid Study Identifier	domain	usubjid Unique Subject Identifier	vsseq
1	CDISCPILOT01	VS	01-701-1015	1
2	CDISCPILOT01	VS	01-701-1015	2
3	CDISCPILOT01	VS	01-701-1015	3
4	CDISCPILOT01	VS	01-701-1015	4
5	CDISCPILOT01	VS	01-701-1015	5
6	CDI			
7	CDI			
8	CDI			
9	CDI			
10	CDI			
11	CDI			
12	CDI			
13	CDI			
14	CDI			
15	CDI			
16	CDI			

	studyid Study Identifier	domain	usubjid	subjid Subject
1	CDISCPILOT01	DM	01-701-1015	1015
2	CDISCPILOT01	DM	01-701-1023	1023
3	CDISCPILOT01	DM	01-701-1028	1028
4	CDISCPILOT01	DM	01-701-1033	1033
5	CDISCPILOT01	DM	01-701-1034	1034
6	CDISCPILOT01	DM	01-701-1047	1047
7	CDISCPILOT01	DM	01-701-1057	1057

THE VERSIONING PROBLEM

- Standards Change over time
- Version-Conversion
  - Instance data is not version-independent



21 March 2014
SDTM-IG: 3.2
TCG: 1.0
SDTM-Term: 2014-03-28
MedDRA: 17.0
WHODrug: MAR/2014
Company: 2.2

15 July 2016
SDTM-IG: 3.2
TCG: 3.0
SDTM-Term: 2016-06-30
MedDRA: 19.0
WHODrug: JUN/2016
Company: 2.7

STDM-IG	Study Data Tabulation Model (SDTM), Implementation Guide
TCG*	Study Data Technical Conformance Guide
SDTM-TERM**	SDTM terminology
MedDRA	Medical Dictionary for Regulatory Activities
WHODrug	World Health Organization Drug Dictionary
Company	Fictional company standard.

*https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm#guides

**https://evs.nci.nih.gov/ftp1/CDISC/SDTM/Archive/

THE VERSIONING PROBLEM

Proposed Solution:

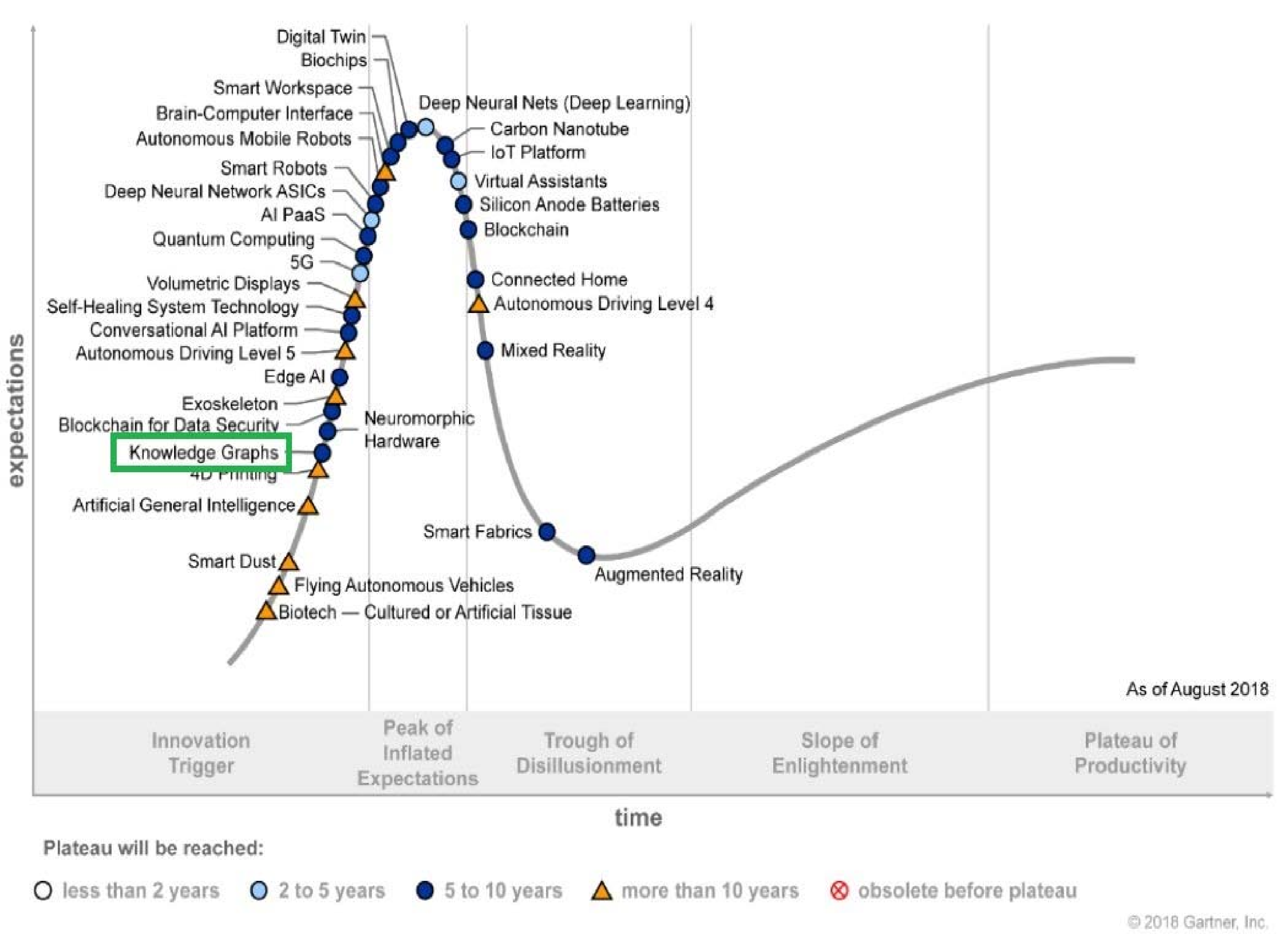
- Data independent of standards version
- *graph data*
- Materialize graph data into versions of the standards

OTHER PROBLEMS

- Quality
- Integration
- Traceability and Provenance
- Metadata
- ***Human Error***
- ...

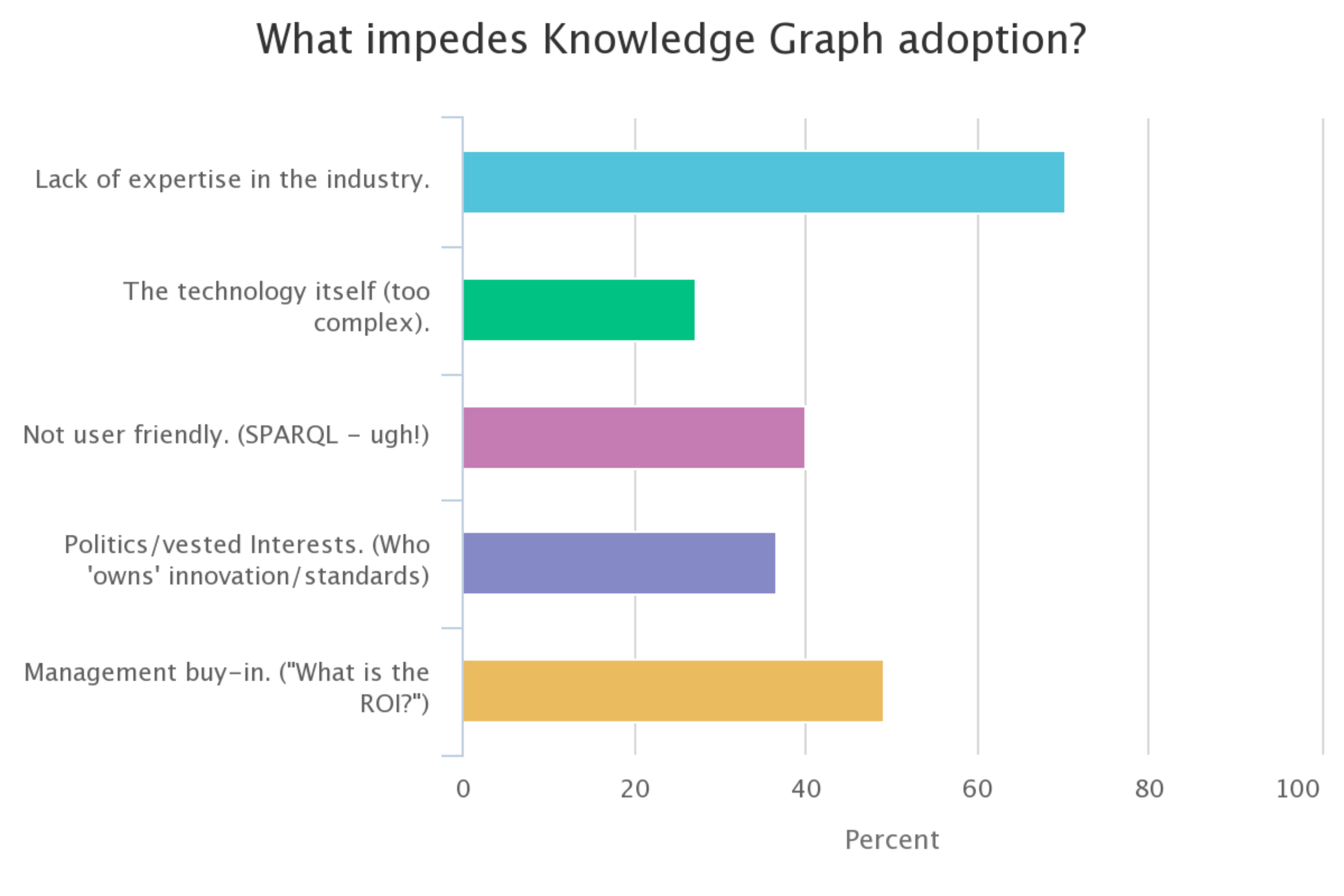
Knowledge Graphs can help us solve these problems!

RECOGNITION (AT LAST)

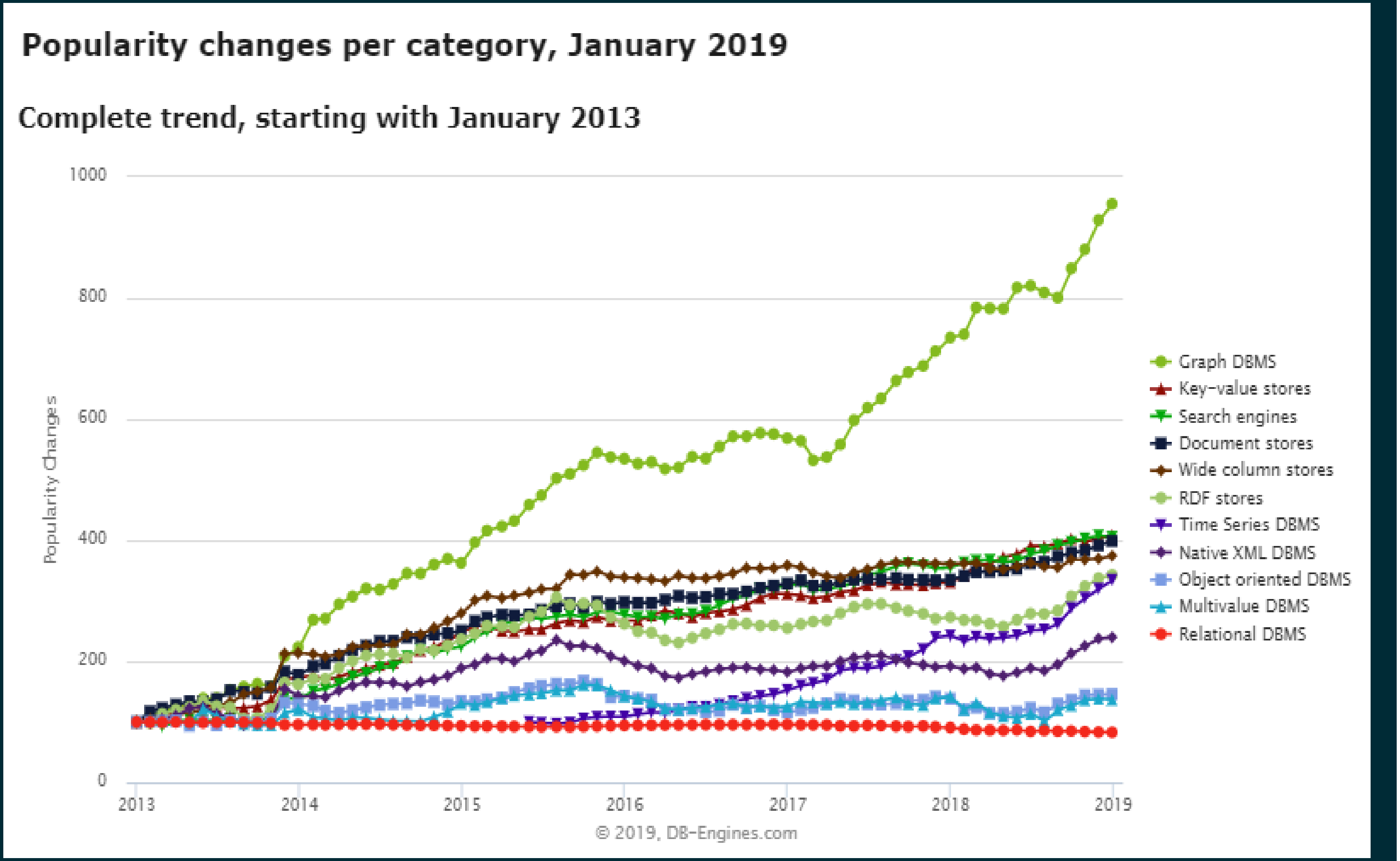


POLL

WHAT IMPEDES KNOWLEDGE GRAPH ADOPTION?



CHALLENGE 1: *EXPERTISE*



Popularity changes per category, January 2019

Complete trend, starting with January 2013

© 2019, DB-Engines.com

https://db-engines.com/en/ranking_categories

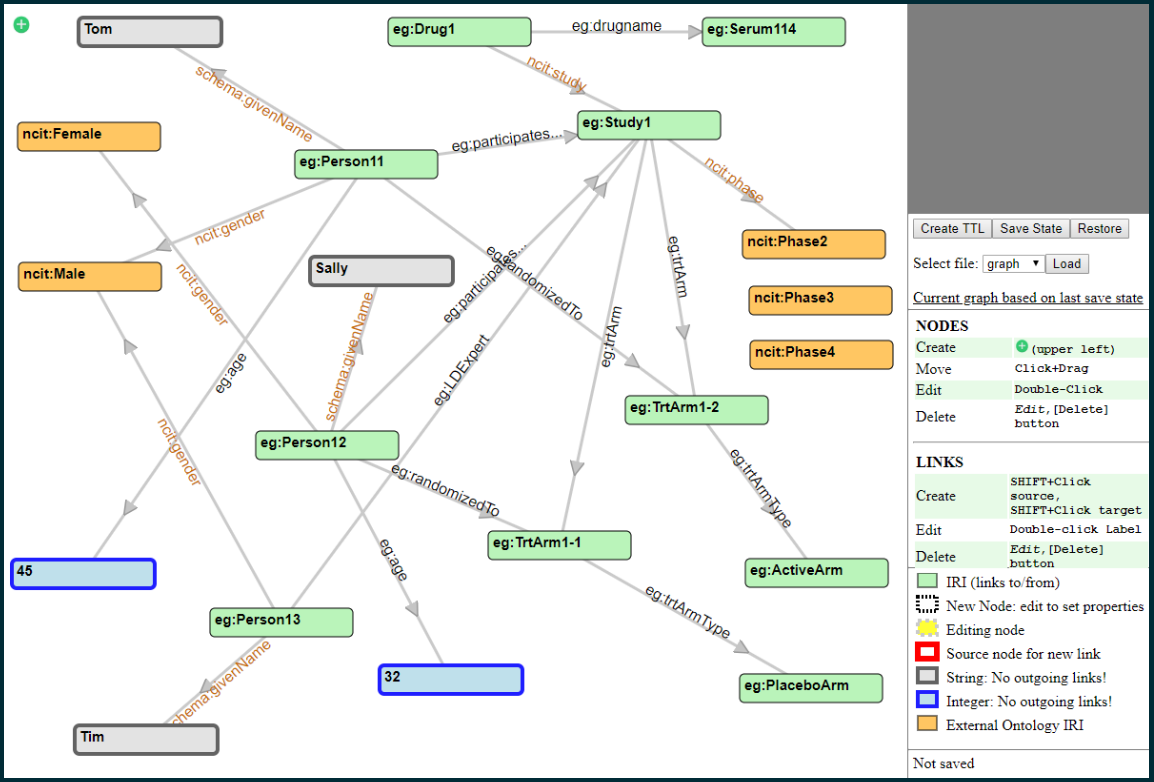
SKILLS DEFICIT

IS IT MORE PRACTICAL TO TRAIN:

A *Knowledge Graph* expert in *Clinical Trials*?
or
A *Clinical Trials* expert in *Knowledge Graphs* ?

LINKED DATA EDUCATION

HANDS-ON WORKSHOP: GRAPH EDITOR



Current graph based on last save state

NODES

Create	⊕ (upper left)
Move	Click+Drag
Edit	Double-Click
Delete	*Edit*, [Delete] button

LINKS

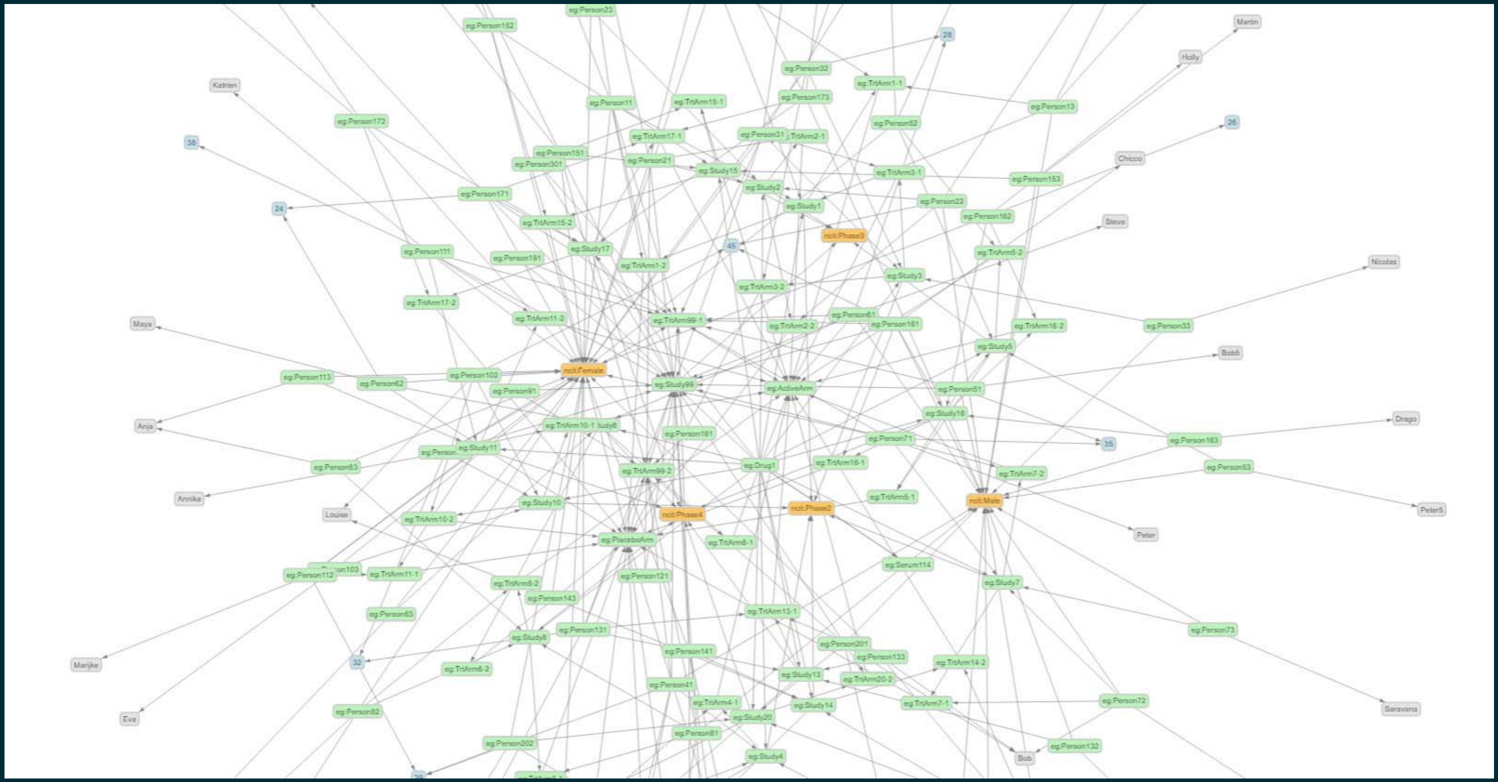
Create	SHIFT+Click source, SHIFT+Click target
Edit	Double-click Label
Delete	*Edit*, [Delete] button

- 🟩 IRI (links to/from)
- ⬚ New Node: edit to set properties
- 🟨 Editing node
- 🟥 Source node for new link
- ▢ String: No outgoing links!
- 🟦 Integer: No outgoing links!
- 🟧 External Ontology IRI

Not saved

Tim Williams (UCB) , Johannes Ullander (A3), PhUSE EU Connect 18, Frankfurt

HANDS-ON WORKSHOP: 21 MERGED STUDIES



PhUSE EU Connect 18, Frankfurt

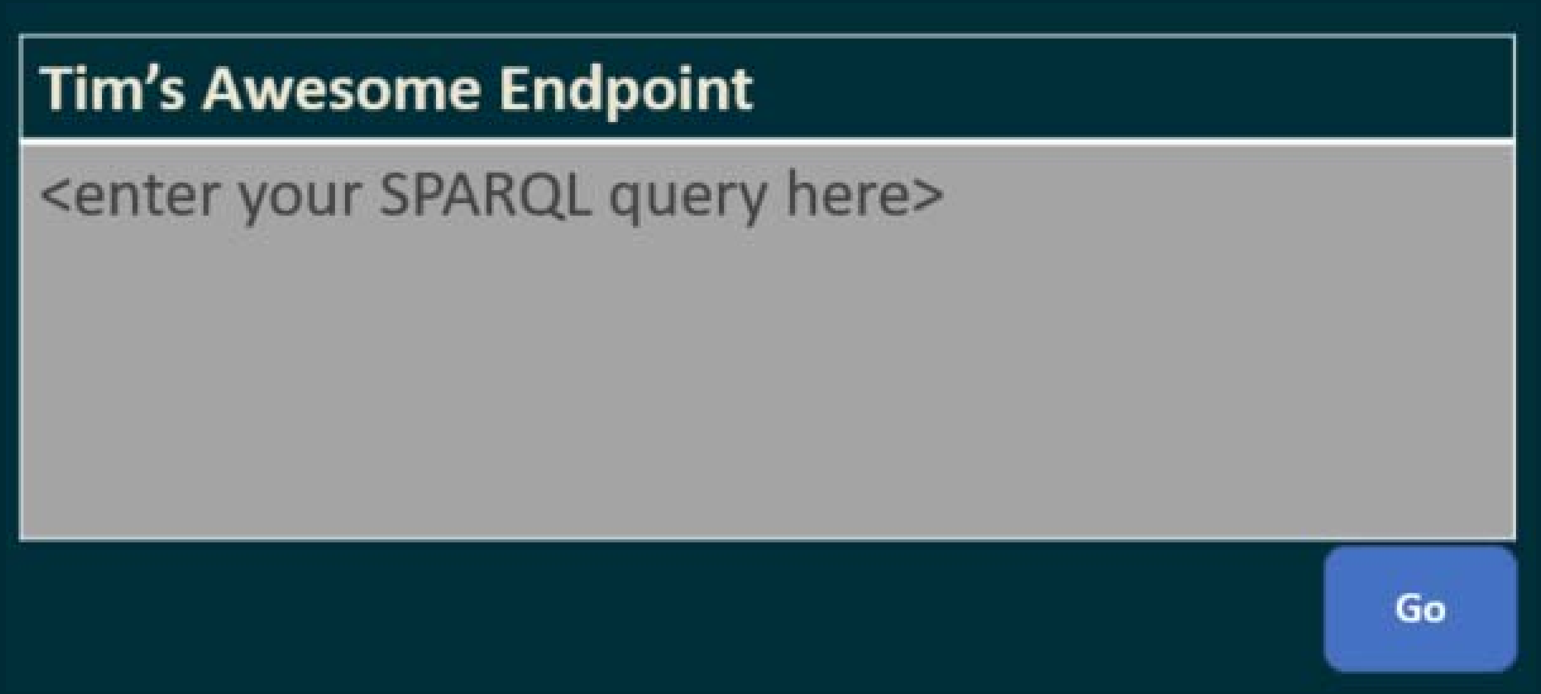
CHALLENGE 2: *COMPLEXITY*

CHALLENGE: *COMPLEXITY*

> *"People think RDF is a pain because it is complicated. The truth is even worse. RDF is painfully simplistic, but it allows you to work with real-world data and problems that are horribly complicated."*
> ***- Dan Brickley and Libby Miller***

CHALLENGE 3: *USER INTERFACES*

USER INTERFACES

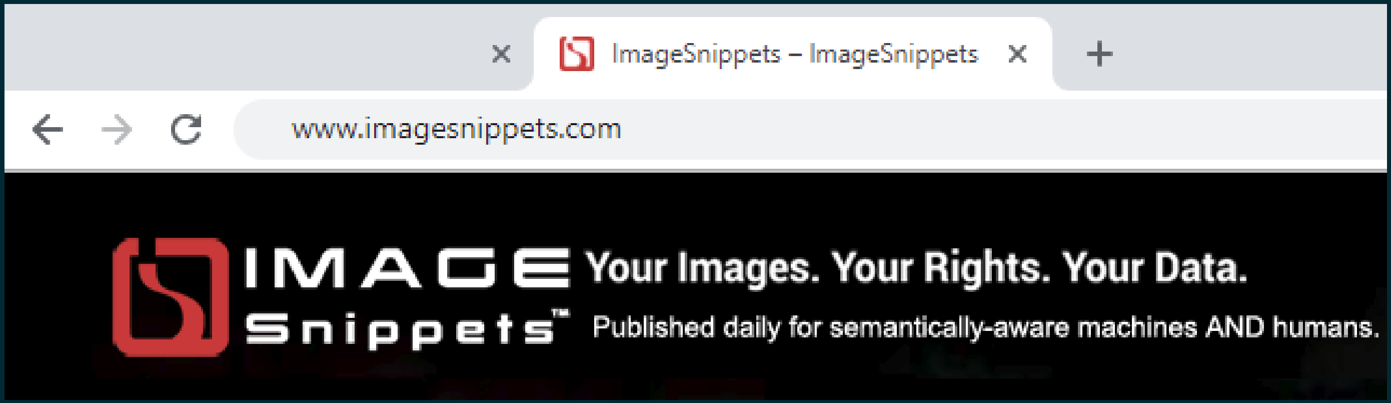


Tim’s Awesome Endpoint

<enter your SPARQL query here>

USER INTERFACES

A FRIENDLY EXAMPLE


ImageSnippets – ImageSnippets
www.imagesnippets.com
IMAGE Snippets™
Your Images. Your Rights. Your Data.
Published daily for semantically-aware machines AND humans.

- Margaret Warren

ImageSnippets

http://www.imagesnippets.com

http://www.imagesnippets.com/imgtag/images/NovasTaylor@gmail.com/Brain_CT_scan.html

CHALLENGE 4: *POLITICS*

CHALLENGE: *POLITICS*

Who owns innovation?

- No one.
- You need IT, Business support
- More than enough work for everyone...
- New roles, new expertise

  Challenge is bigger than any one:

  - person
  - department
  - company
  - agency, authority, organization

CHALLENGE 5: *RETURN ON INVESTMENT*

THE KNOWLEDGE GRAPH PROMISE

Positive

- Impact the entire data lifecycle.

Negative

- Impact the entire data lifecycle.

R.O.I UNICORN

Image Attribution: https://bit.ly/2x0Hjmd

THE ROOFSHOT / MOONSHOT MANIFESTO

Roofshot

Incremental impacts

1. Study URI
2. CTD (SDTM) as RDF
3. Open Ontology Development

Moonshot

Invent & apply state-of-the-art

`Knowledge Graph`

Pharma Data Life Cycle

Concept & Image Attribution:

https://rework.withgoogle.com/blog/the-roofshot-manifesto/

ROOFSHOT 1

STUDY URI AS AN INDUSTRY STANDARD

BASED ON:

> *"Study URI" - K. Forsberg, D. Goude. PhUSE EUConnect18.*

Why?

- Easy entry point for Pharma
- Familiar concept: NCT Number (ClinicalTrials.gov)
  - CT.gov must first review and approve Protocol

STUDY URI COMPONENTS

https://data.pharma.abc/clinicaltrial/D3562C00096

1. Global Namespace
2. Resource type
3. Trial ID

Review and comment at: https://github.com/phuse-org/LinkedDataEducation/blob/master/doc/StudyURI.md

Invite comment from FDA, EMA, PMDA, CDISC... *You* !

ROOFSHOT 2

PhUSE PROJECT

CLINICAL TRIALS DATA (SDTM) AS RDF

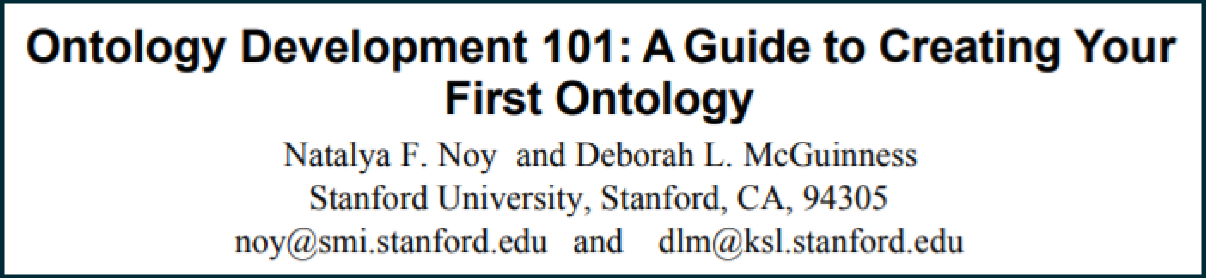
Project Co-Leads:
Dr. Armando Oliva, Semantica LLC
Tim Williams, UCB Biosciences

CORE PRINCIPLES

- Model Clinical Data in a way that makes sense to clinicians and data consumers
- Model what is needed to automate creation of high-quality SDTM
- Re-use existing models and definitions where possible
  - Examples: SDTM Terminology, BRIDG, HL7, ISO, Time ontology...
  - Placeholder classes until other standards are in RDF

THE CTDasRDF APPROACH

A combination classic ontology development:



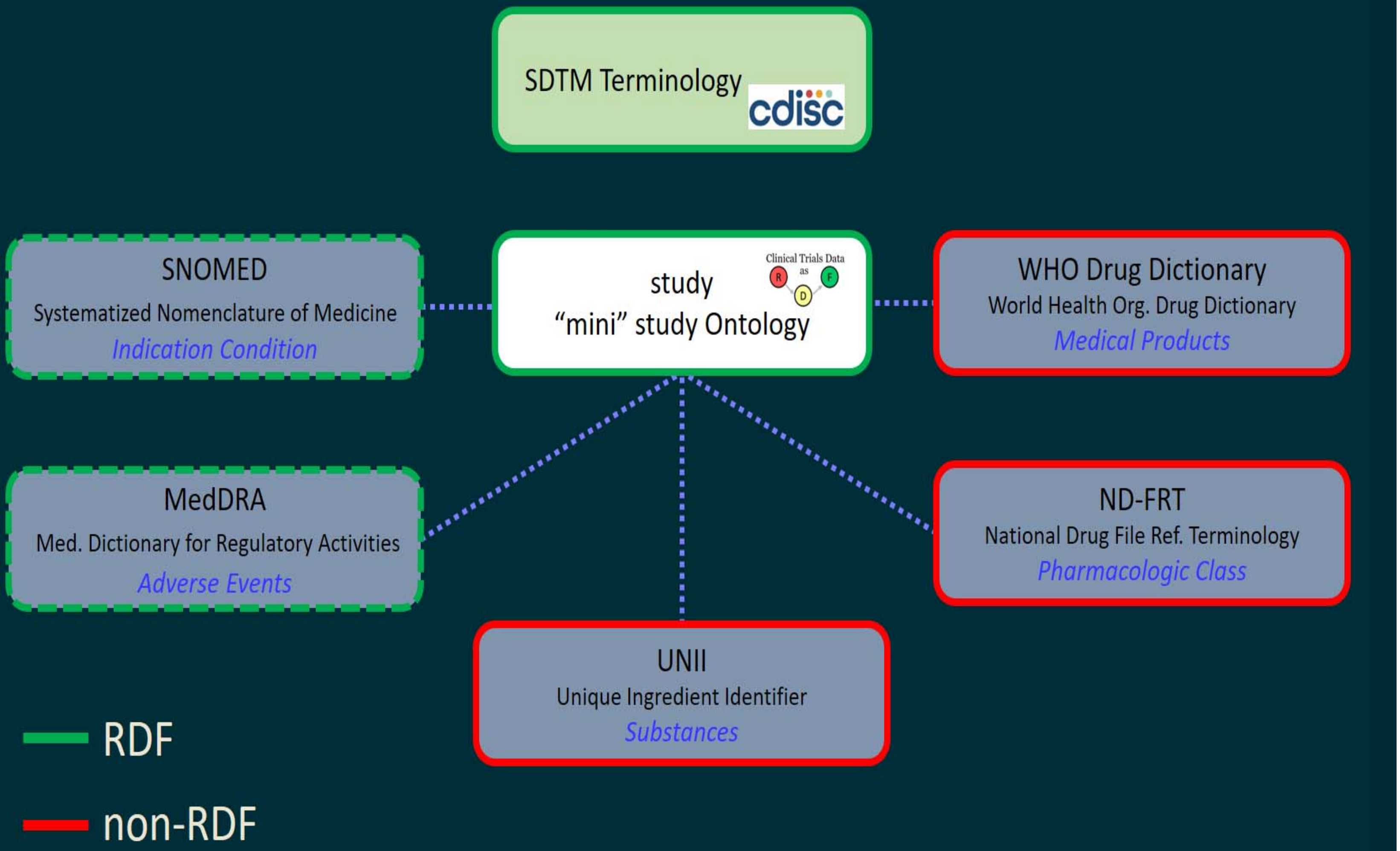
Ontology Development 101: A Guide to Creating Your First Ontology

Natalya F. Noy and Deborah L. McGuinness
Stanford University, Stanford, CA, 94305
noy@smi.stanford.edu and dlm@ksl.stanford.edu

Combined with modeling

- real-world data
- processes (Study Design, Protocol, rules)

LEVERAGE EXISTING STANDARDS



SDTM Terminology cdisc

study
“mini” study Ontology

Clinical Trials Data as R D F

SNOMED
Systematized Nomenclature of Medicine
Indication Condition

MedDRA
Med. Dictionary for Regulatory Activities
Adverse Events

WHO Drug Dictionary
World Health Org. Drug Dictionary
Medical Products

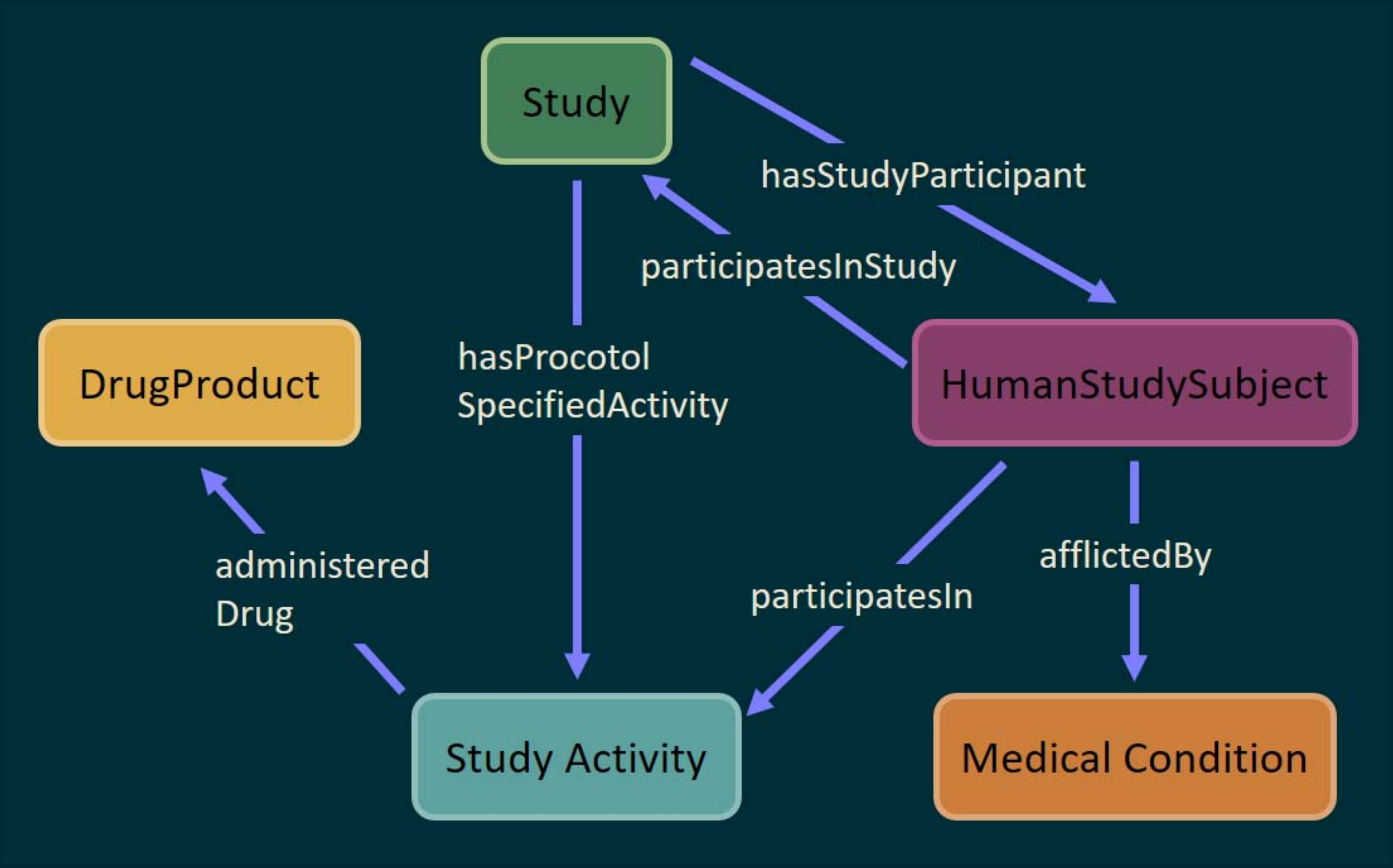
ND-FRT
National Drug File Ref. Terminology
Pharmacologic Class

UNII
Unique Ingredient Identifier
Substances

🟩 RDF

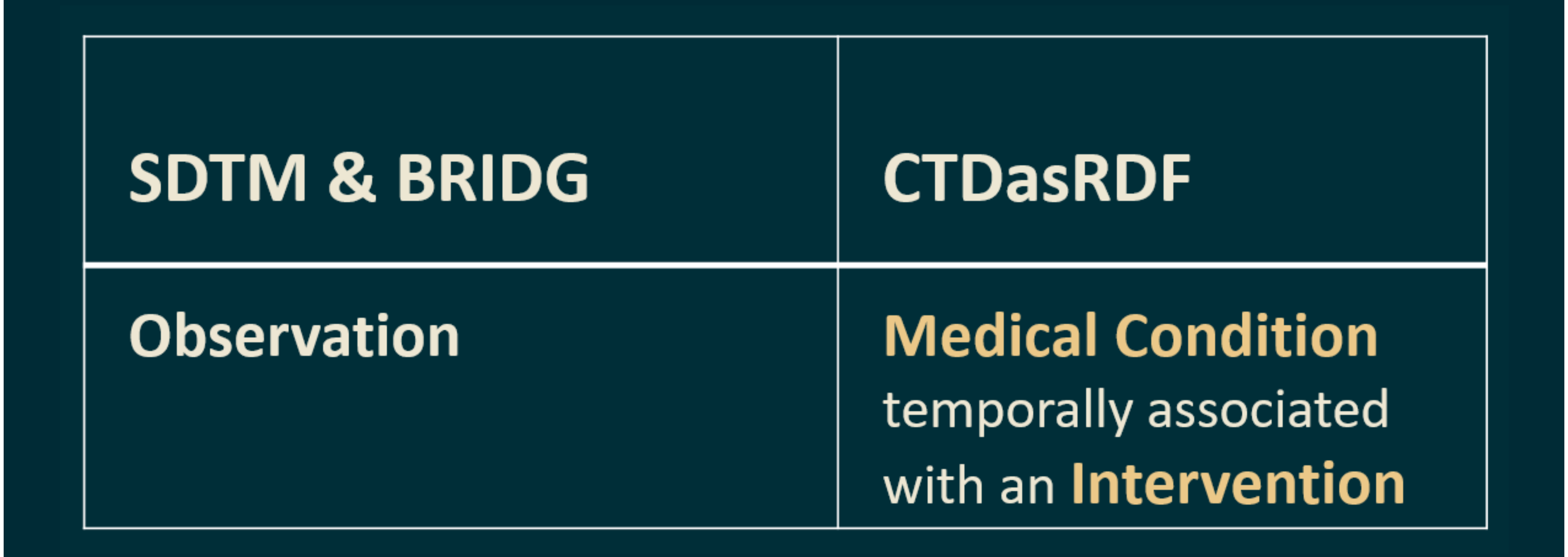
🟥 non-RDF

CORE STUDY 'MINI' ONTOLOGY



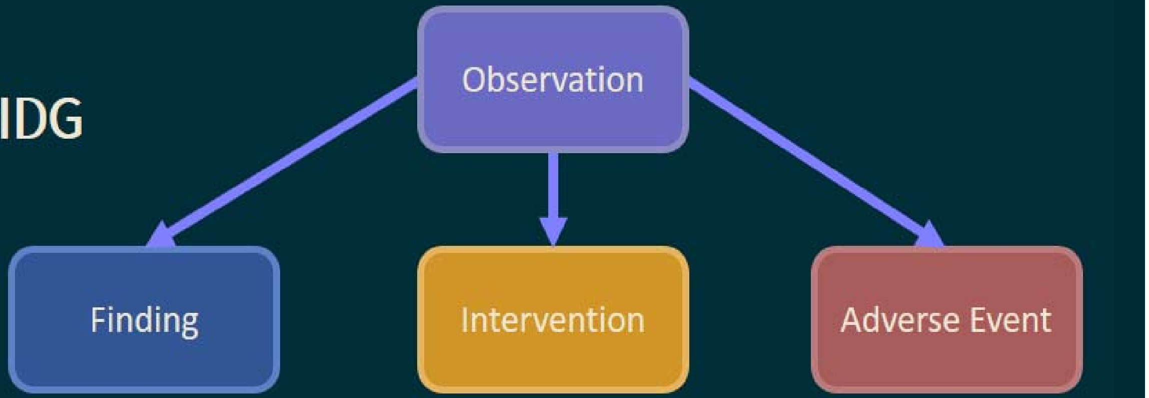
FAMILIAR CONCEPT:
NEW REPRESENTATION

WHAT IS AN ADVERSE EVENT?



SDTM & BRIDG	CTDasRDF
Observation	**Medical Condition** temporally associated with an **Intervention**

SDTM, BRIDG



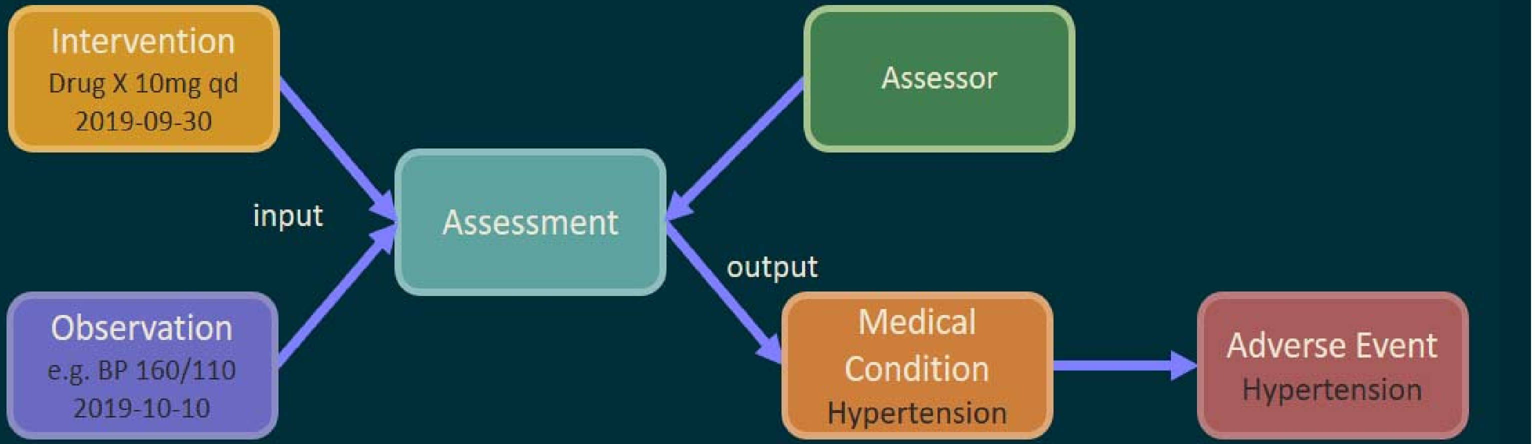
Observation

Finding

Intervention

Adverse Event

CTDasRDF



Intervention
Drug X 10mg qd
2019-09-30

Observation
e.g. BP 160/110
2019-10-10

input

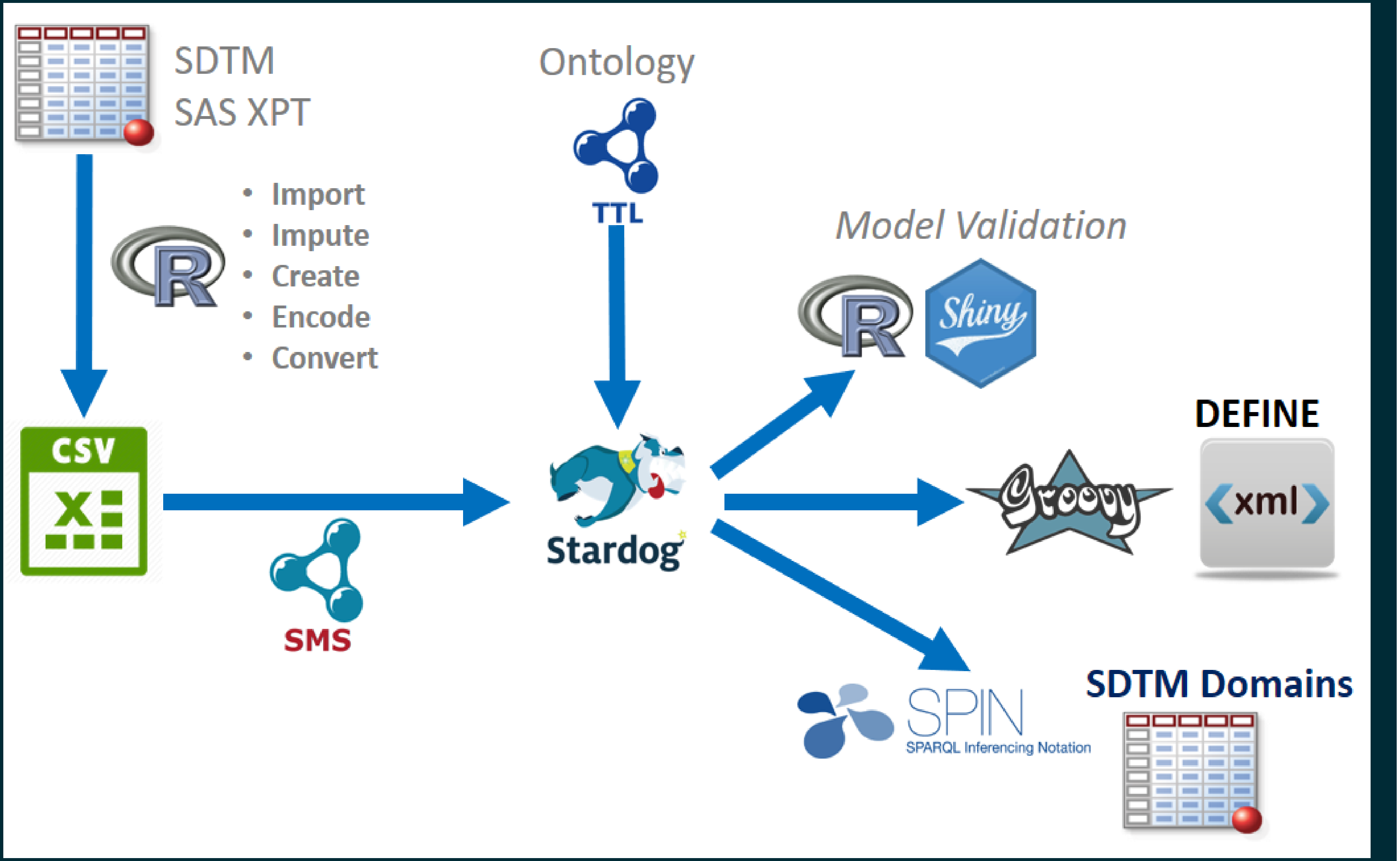
Assessment

Assessor

output

Medical Condition
Hypertension

Adverse Event
Hypertension



SDTM
SAS XPT

- Import
- Impute
- Create
- Encode
- Convert

Ontology

TTL

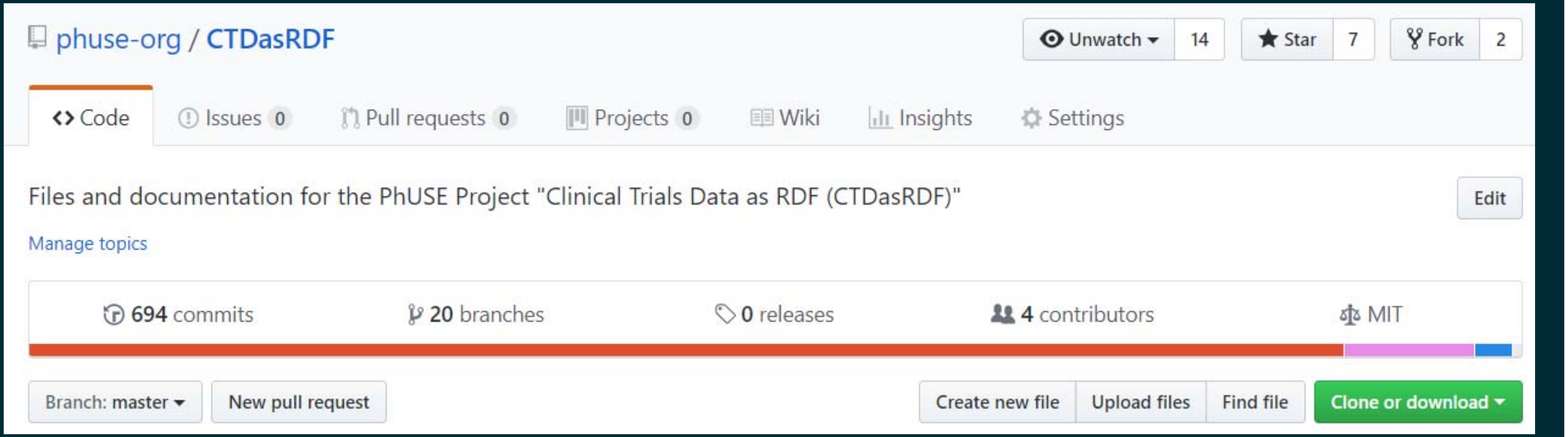
SMS

Model Validation

DEFINE

SDTM Domains

PROJECT WORK ON GITHUB



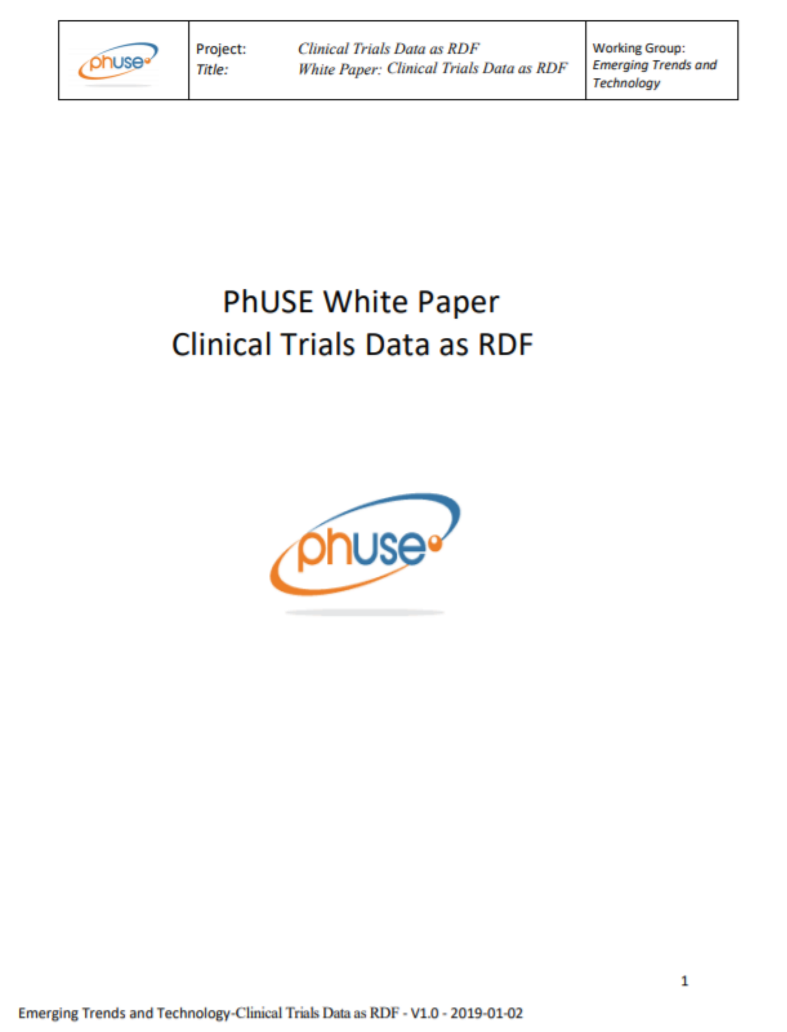
phuse-org / CTDasRDF

Files and documentation for the PhUSE Project "Clinical Trials Data as RDF (CTDasRDF)"

https://github.com/phuse-org/CTDasRDF

Project extending to SEND (2019)

CTDasRDF WHITE PAPER



phuse	Project: Title:	*Clinical Trials Data as RDF* *White Paper: Clinical Trials Data as RDF*	Working Group: *Emerging Trends and Technology*

PhUSE White Paper
Clinical Trials Data as RDF

1

Emerging Trends and Technology-Clinical Trials Data as RDF - V1.0 - 2019-01-02

https://www.phuse.eu/white-papers

NEXT STEPS

New Project

- Extend existing domains (DM, VS, EX, TS) to include non-clinical concepts.
- Expand to new domains: AE and onward
- Development of Study URI concept.

ROOFSHOT 3

COOPERATIVE ONTOLOGY DEVELOPMENT

LEVERAGE EXISTING ONTOLOGIES

> *"We cannot compete with centralized systems unless we collaborate."*
>
> *- Ruben Verborgh*
>
> *Decentralizing the Semantic Web Through Incentivized Collaboration*

Proposal:

- Precompetitive general ontologies
- Extensible to internal, proprietary ontologies

HOW TO OPEN SOURCE ONTOLOGY DEVELOPMENT?

- GitHub?
- Existing pre-competitive organizations?
  - PhUSE
  - TransCelerate
  - Pistoia Alliance

OPEN SOURCE ONTOLOGY

CHALLENGES

- Gate keeper?
- Conflict resolution (approach, code)
- Company:
  - Participation?
  - Contribution?
- Volunteers

ONTOLOGY AVAILABILITY AND CURATION

Leverage existing portals?

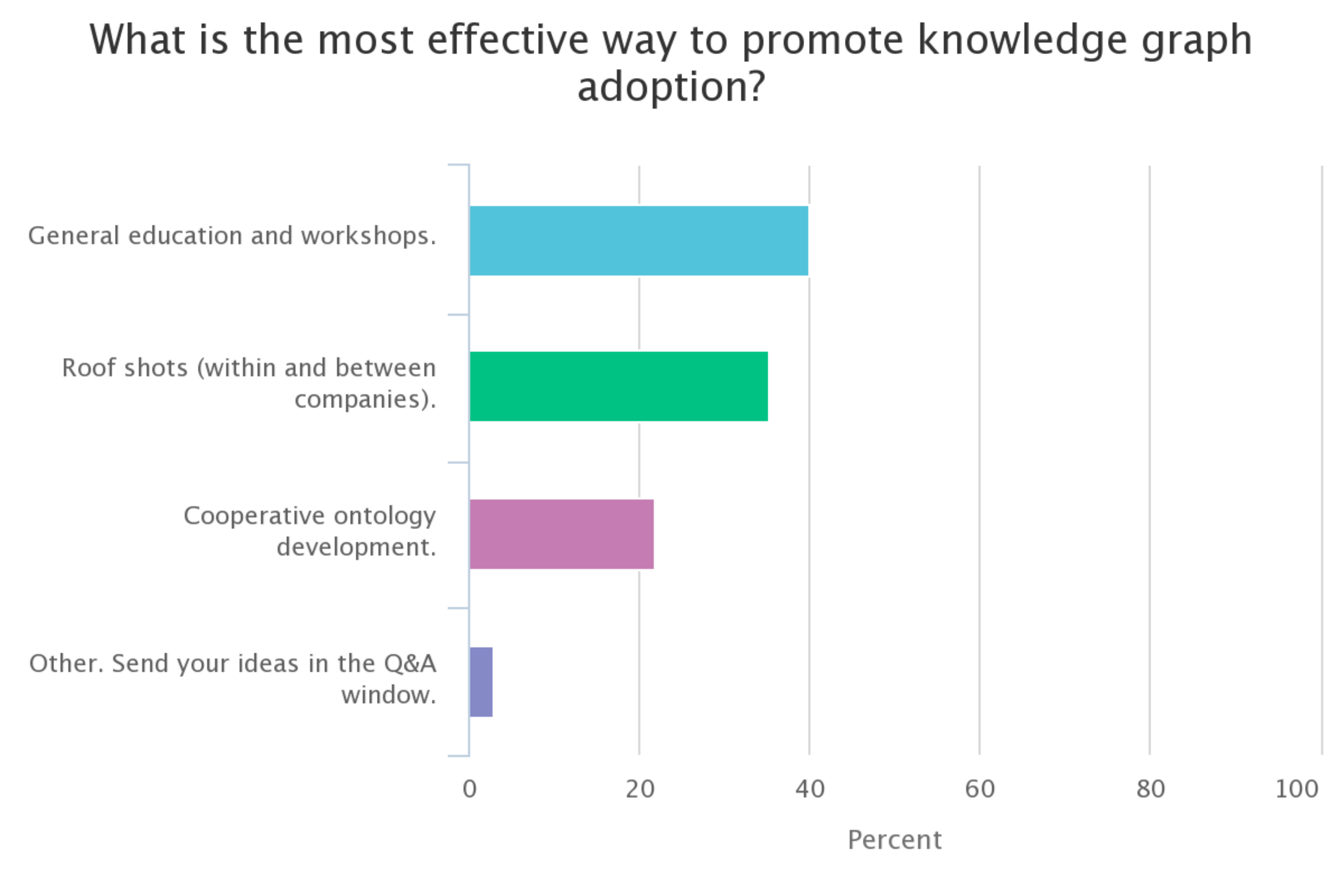
- Open PHACTS
- The OBO Foundry
- BioPortal

ACCESS MUST BE OPEN

Don't hide my OWL behind an API!

FINAL POLL

WHAT IS THE MOST EFFECTIVE WAY TO PROMOTE KNOWLEDGE GRAPH ADOPTION?



Thank you!

CONTACT:

Email: tim.williams@phuse.eu
LinkedIn: https://www.linkedin.com/in/timpwilliams

Twitter : 🐦NovasTaylor