

How Knowledge Graphs will Transform the Pharmaceutical Industry

Tim Williams



- Knowledge Graph Project Lead
 - PHUSE
- Statistical Solutions Lead
 - UCB Biosciences, Raleigh N.C.
- Past Experience
 - ~ 20 years experience in Pharma
 - Systems Admin, Systems Validation & Deployment, Research Associate (Epidemiology), Programmer.

Perspective : Late Phase Clinical Trials

Opinions are my own

Why Transform?

Time

- 10 years from discovery to marketplace
- Clinical trials 6 to 7 years

Cost for successful drug

- \$314M to \$2.8B (median \$985M)*
 - Includes cost of failures
- Success Rate
 - Less than 12%

* <https://www.biospace.com/article/median-cost-of-bringing-a-new-drug-to-market-985-million/>

What if we could make a 1% improvement?

- Cost Savings
 - \$3M to 28M per successful drug
- Re-invest
 - Research and Development
 - New insights, discovery, revenue streams
 - New Technology
 - New efficiencies
 - Patient-value initiatives

Outline

- **Current State**
- **Future State**
- **PHUSE Project**
- **Industry Examples**
- **Implementation Strategy**
- **Q&A**

Terminology

- **Knowledge Graph**

- Entities and relations in a defined schema
- *Machine-readable, semantic, extensible model*

- **Linked Data**

- Resource Description Framework (RDF)

- **Semantic Web**

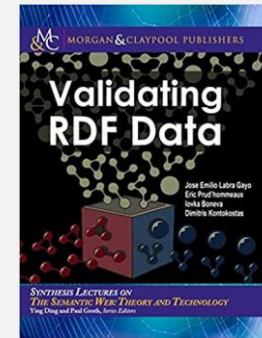
- Linked Data (as RDF), Web Ontology Language (OWL), SPARQL, SHACL, etc.



Complexity? Flexibility.

“People think RDF is a pain because it is complicated. The truth is even worse. RDF is painfully simplistic, but allows you to work with real-world data and problems that are horribly complicated.”

- attributed to Dan Brickley, Libby Miller. In: “Validating RDF Data”



Current State

The pharmaceutical industry needs a technology transformation.

Data Landscape



*Gartner

Code Landscape

What does the code do?









- Data Manipulation (majority)
 - Merging, subsetting, categorization...
- Statistical Procedures
- Output formatting

Problems

- Errors. 15-50 defects 1000 lines*
- Application & Skill set lock-in
- Logic and metadata in code
- Add context and complexity

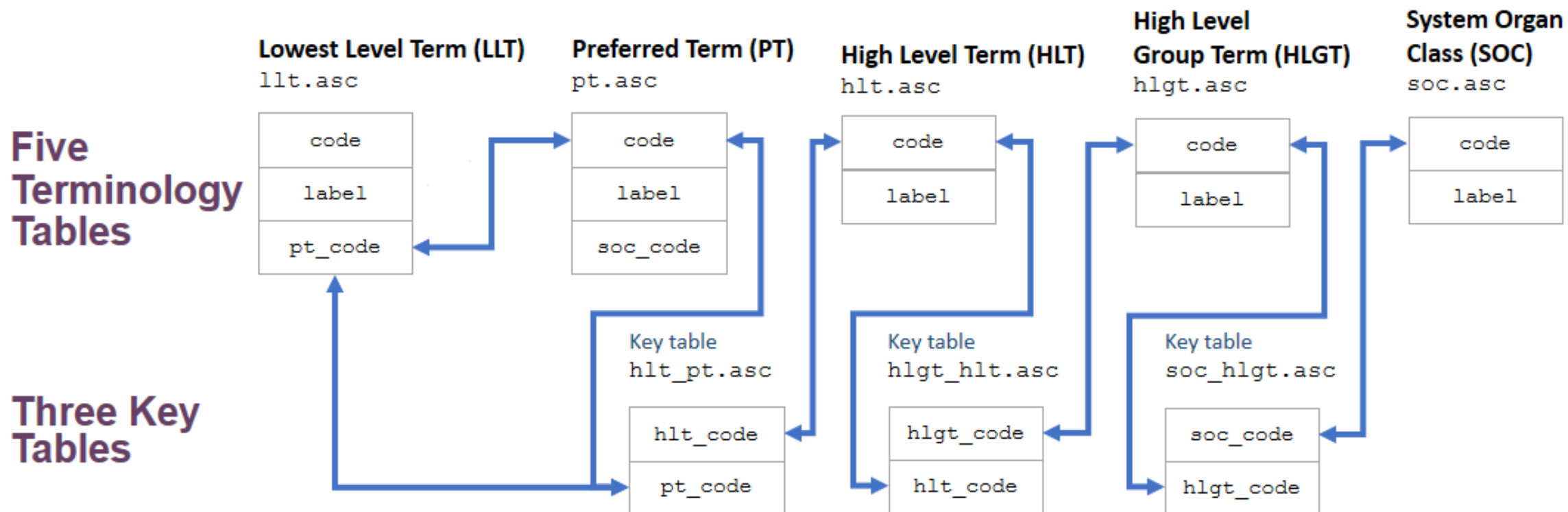
* "Code Complete" – Steve McConnell

Dictionaries, Taxonomies, Coding...

- Clinical Data Interchange Standards Consortium (CDISC) [Controlled Terminology](#) 
- Logical Observation Identifiers Names and Codes ([LOINC](#)) 
- WHO's International Classification of Diseases   World Health Organization
- [ICD-9/ICD-10](#)
- Medical Subject Headings ([MeSH](#)) 
- Systematized Nomenclature of Medicine ([SNOMED](#)) 
- WHO Drug Dictionary ([WHODrug](#)) 
- Medical Dictionary for Regulatory Activities ([MedDRA](#)) 

+ Company-specific

MedDRA: ASCII Text Files



MedDRA as RDF

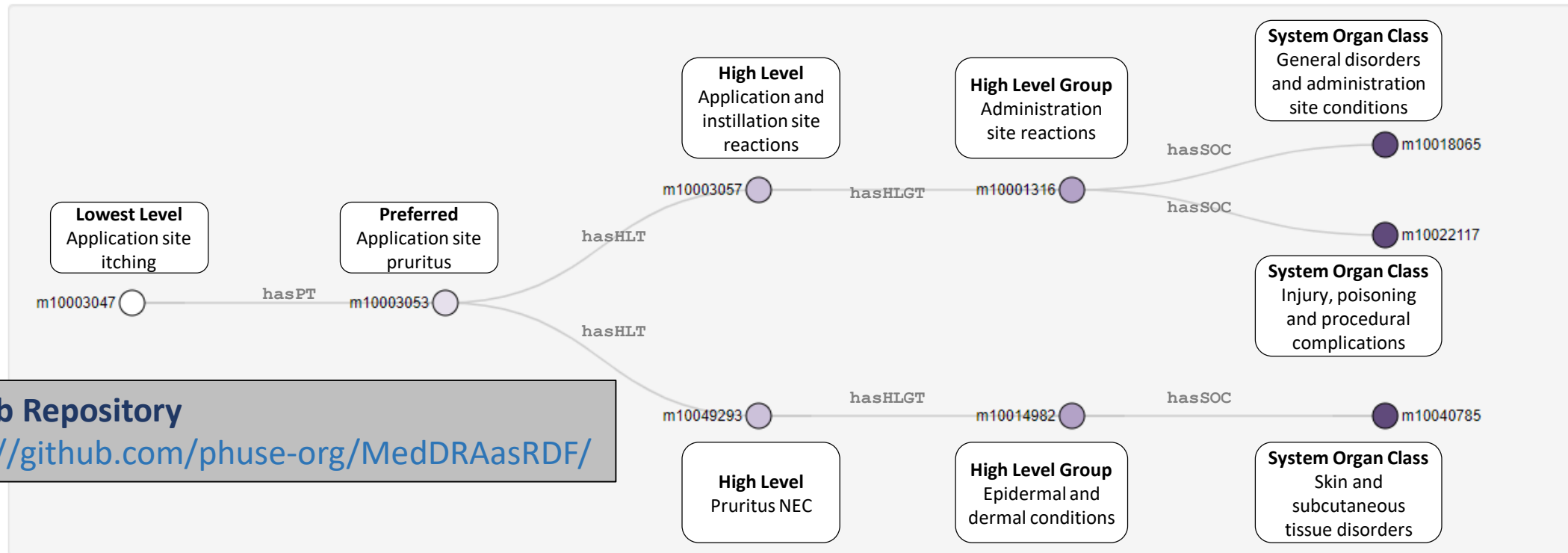
The Query

MedDRA Tracing from Lowest Level Term (LLT) to System Organ Class(SOC)

LLT :

Application Site Itching

```
PREFIX meddra: <https://w3id.org/phuse/meddra#>
PATHS ALL
START ?s = meddra: input$rootNode
END ?o
VIA ?p
```



GitHub Repository

<https://github.com/phuse-org/MedDRAasRDF/>

Data Standards & Models

Pharmaceutical Industry

Health Care

Observational Health Data Sciences and Informatics (OHDSI)  **OHDSI**

- [OMOP Common Data Model](#)

Health Level 7 (HL7) **HL7**

- HL7 Fast Healthcare Interoperability Resources ([HL7 FHIR](#))  **HL7® FHIR®**
<https://www.hl7.org/fhir/rdf.html>
- Microsoft Common Data Model
- [Healthcare extension](#)  Microsoft

Research

Clinical Data Interchange Standards Consortium (cdisc.org/standards) 

- Biomedical Research Integrated Domain Group Model ([BRIDG](#))
- Standard for Exchange of Nonclinical Data ([SEND](#))
- Clinical Data Acquisition Standards Harmonization ([CDASH](#))
- Study Data Tabulation Model ([SDTM](#))
- Analysis Data Model ([ADaM](#))

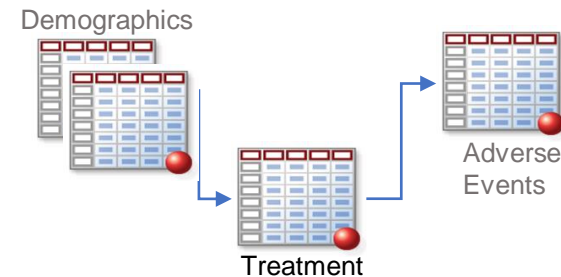
Industry Standards. Example SDTM

Study Data Tabulation Model

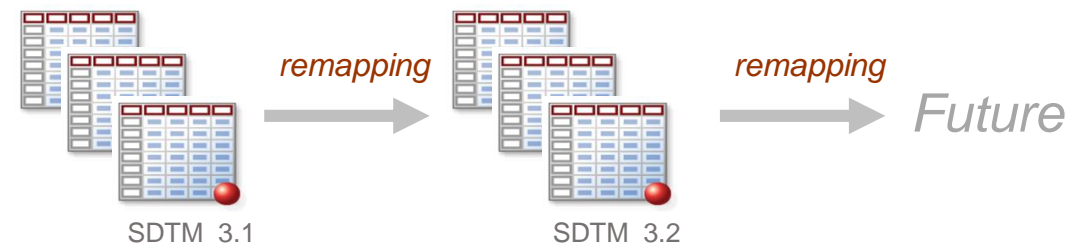
“...defines a standard structure for human clinical trial (study) data tabulations and for nonclinical study data tabulations that are to be submitted as part of a product application to a regulatory authority...”

- en.wikipedia.org/wiki/SDTM

- Domain Silos
- ≠ Clinical Trial Process Model



- \$ Millions for version conversion, pooling, data integration, data mining



Siloed Systems / Siloed Data

Data silos prevent a complete view of the patient.



Micro/Shadow Silos

“What’s in your spreadsheet?”

- Store metadata
 - Drive statistical program execution
 - Manage analytics for data pools
 - List of project programmers
 - Requests for special analysis & tracking replies
 - *A host of (often unknown) uses...*
-
- 85% to 95% of spreadsheets have serious flaws
 - error rate: 1%-5% per cell*

* as cited in McComb 2019

Resistance to Change

- Risk-averse
 - Time, money, **patient safety**
- Established, validated processes
- Skill set
- Business models support *status quo*
 - Vendor familiarity
 - Recoding data to revised standards
 - Cottage industry
 - Creating and maintaining code for studies & standards

But there is a way forward!

Future State

Data is the foundation for transformation.

Data as the Foundation

Evolving

Use Cases

- Discovery
- Safety
- Marketing
- Approvals
- Regulatory Response
- Publication...

Analytics

- mean
- p-value
- regression
- Bayesian
- Machine Learning
- Natural Language
- AI ...

Languages



Standards

- Legacy Standards
- New versions/ Standards

Foundation

Data

Data

Data

Data

From Application-Centric to Data-Centric

Application-Centric

Data-Centric

When

- Historical

- Future

Data
Model

- Planned Use / Analytics
- Submission/Regulatory Requirements
- *Use-case dictates schema*

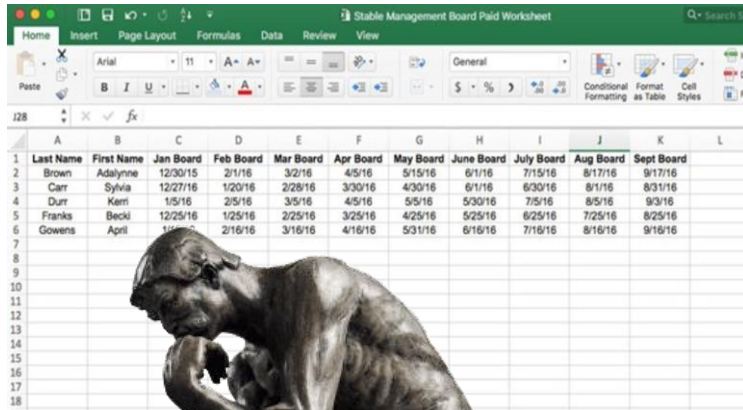
- Real-world Processes
- & Meaning

Designer

- IT Architects,
Database
Administrators

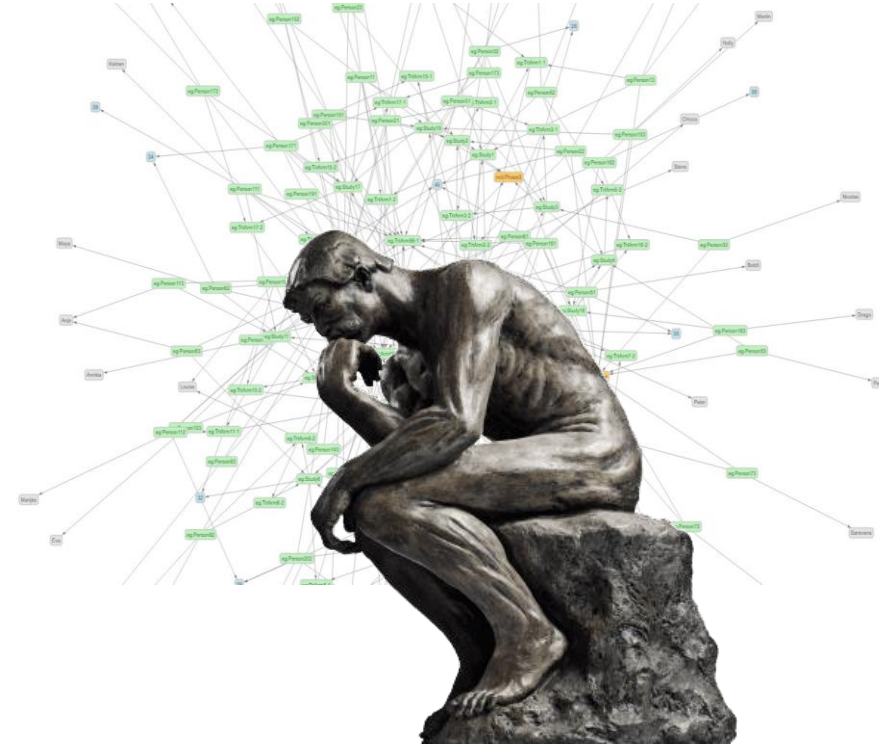
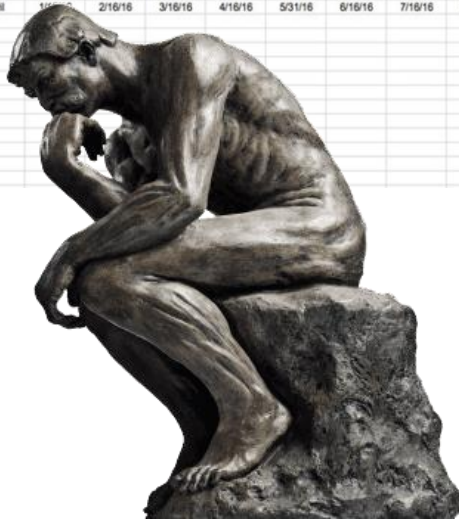
- Researcher, Clinicians, Analysts,
working with IT and technology
experts

Thinking about Data

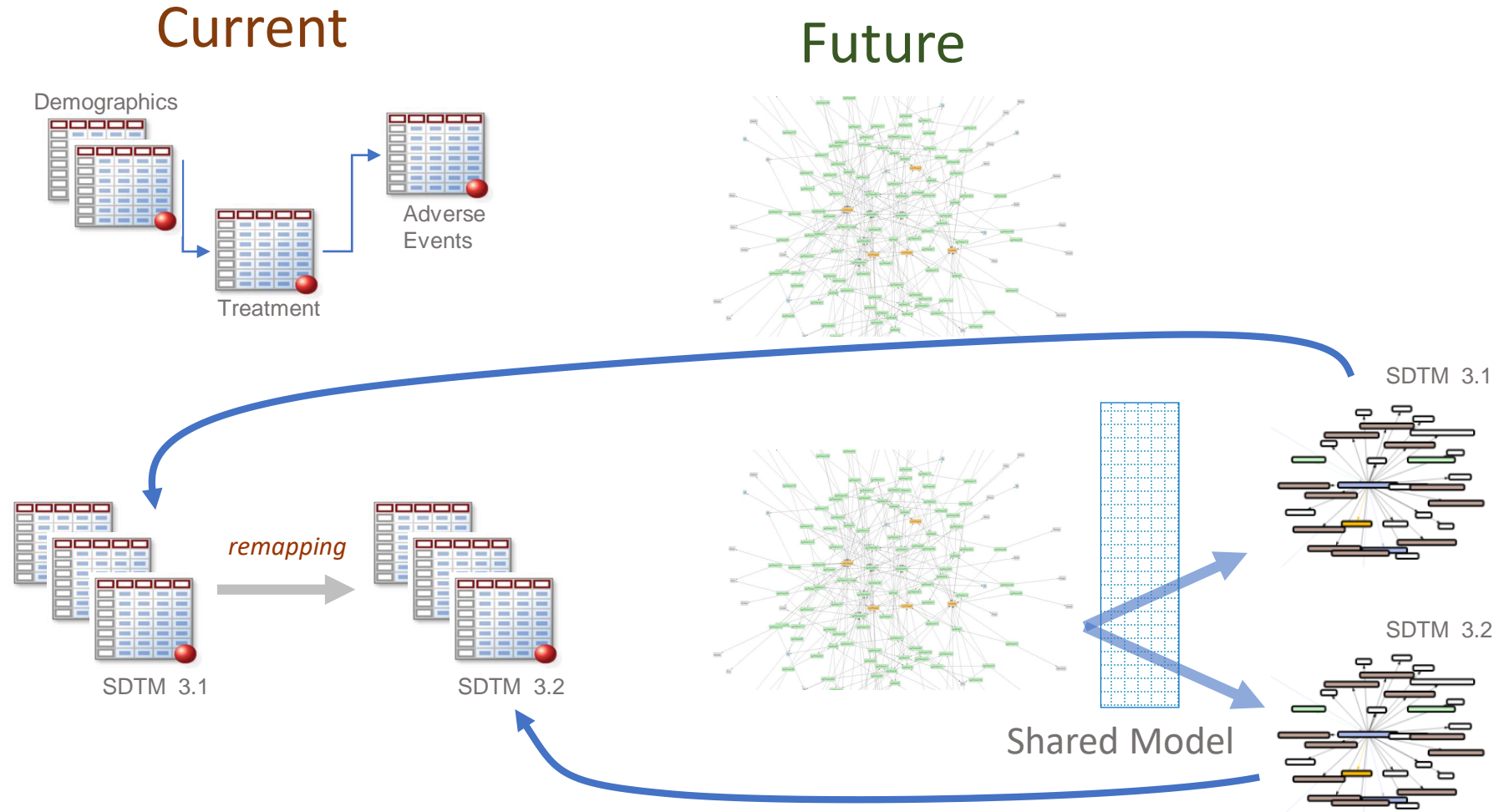


The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Last Name	First Name	Jan Board	Feb Board	Mar Board	Apr Board	May Board	June Board	July Board	Aug Board	Sept Board	
2	Brown	Adalyne	12/30/15	2/1/16	3/2/16	4/5/16	5/15/16	6/1/16	7/15/16	8/17/16	9/17/16	
3	Carl	Sylvia	12/27/15	1/20/16	2/28/16	3/30/16	4/30/16	6/1/16	6/30/16	8/1/16	8/31/16	
4	Durr	Kerr	1/5/16	2/5/16	3/5/16	4/5/16	5/5/16	5/30/16	7/5/16	8/5/16	9/3/16	
5	Franks	Becki	12/25/15	1/25/16	2/25/16	3/25/16	4/25/16	5/25/16	6/25/16	7/25/16	8/25/16	
6	Gowens	April	1/1/16	2/1/16	3/1/16	4/1/16	5/31/16	6/1/16	7/1/16	8/1/16	9/1/16	
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												

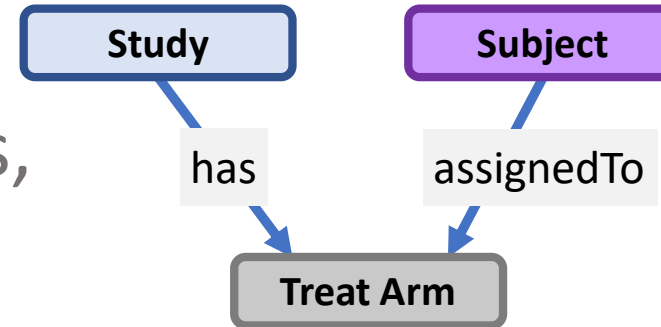


Transition from Relational to Graph



Knowledge Graph Data–Centric Model

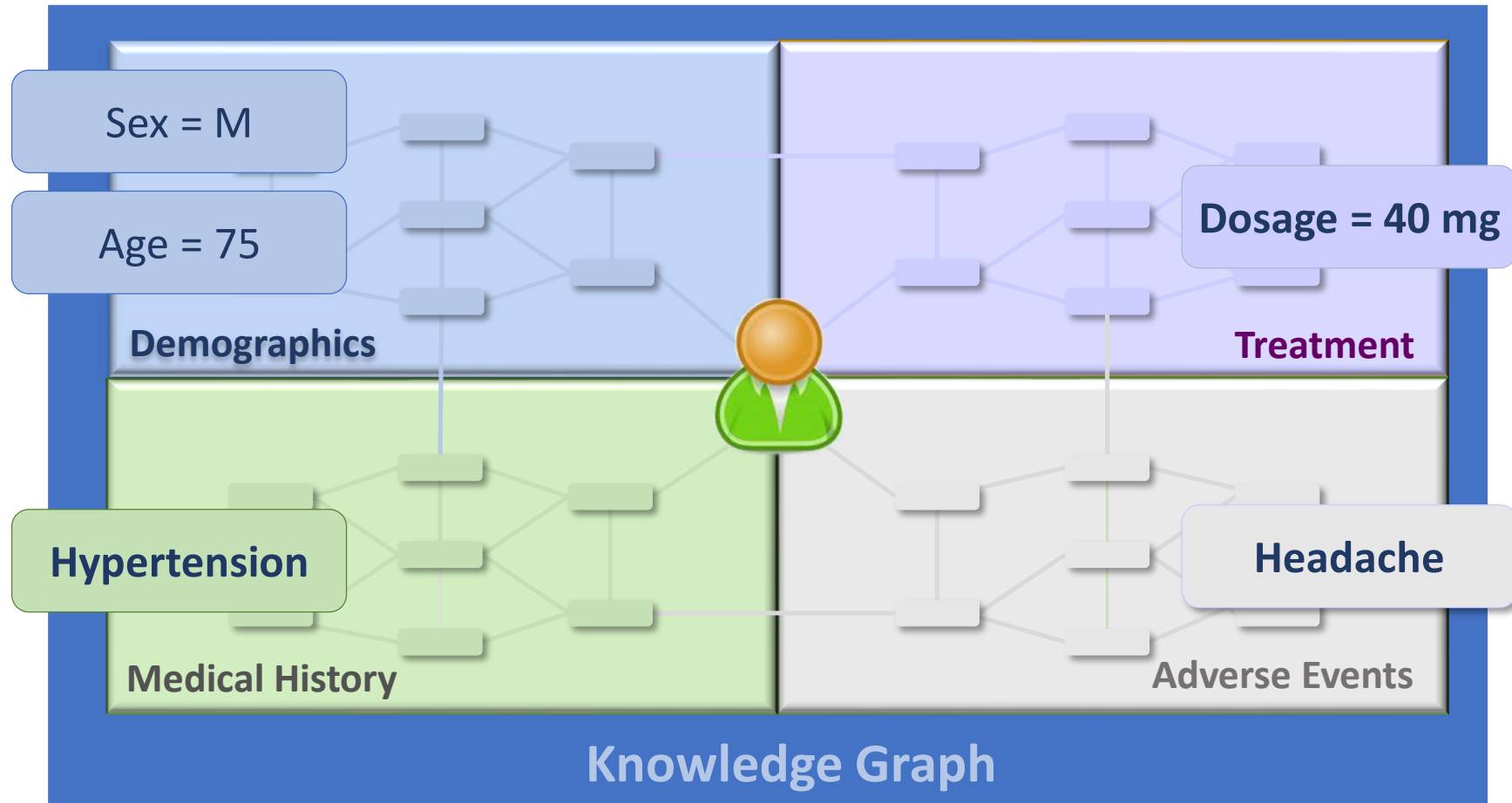
- Use-case neutral
- Real-world processes, entities, relationships



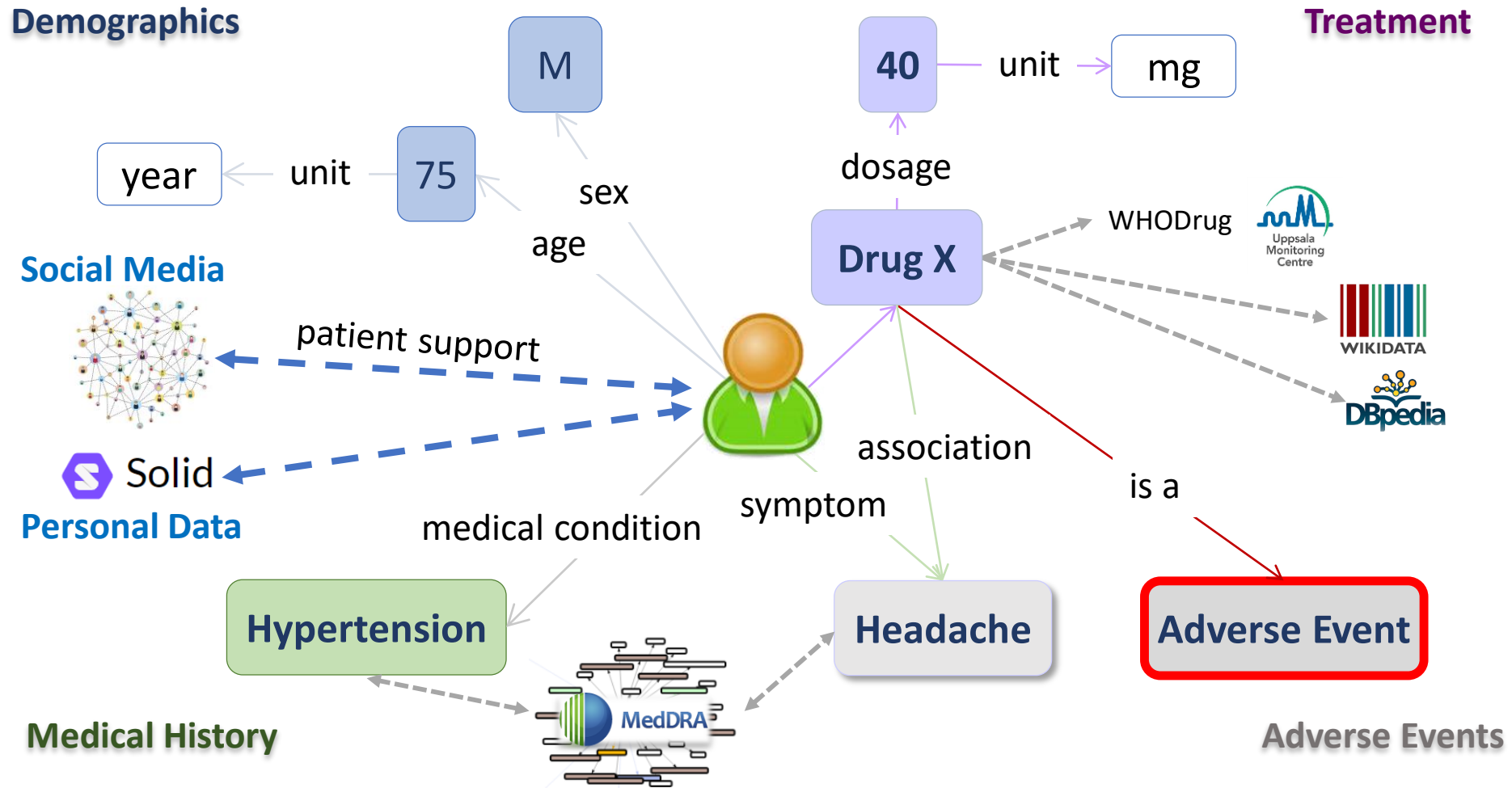
- Knowledge Graph using Resource Description Framework (RDF)



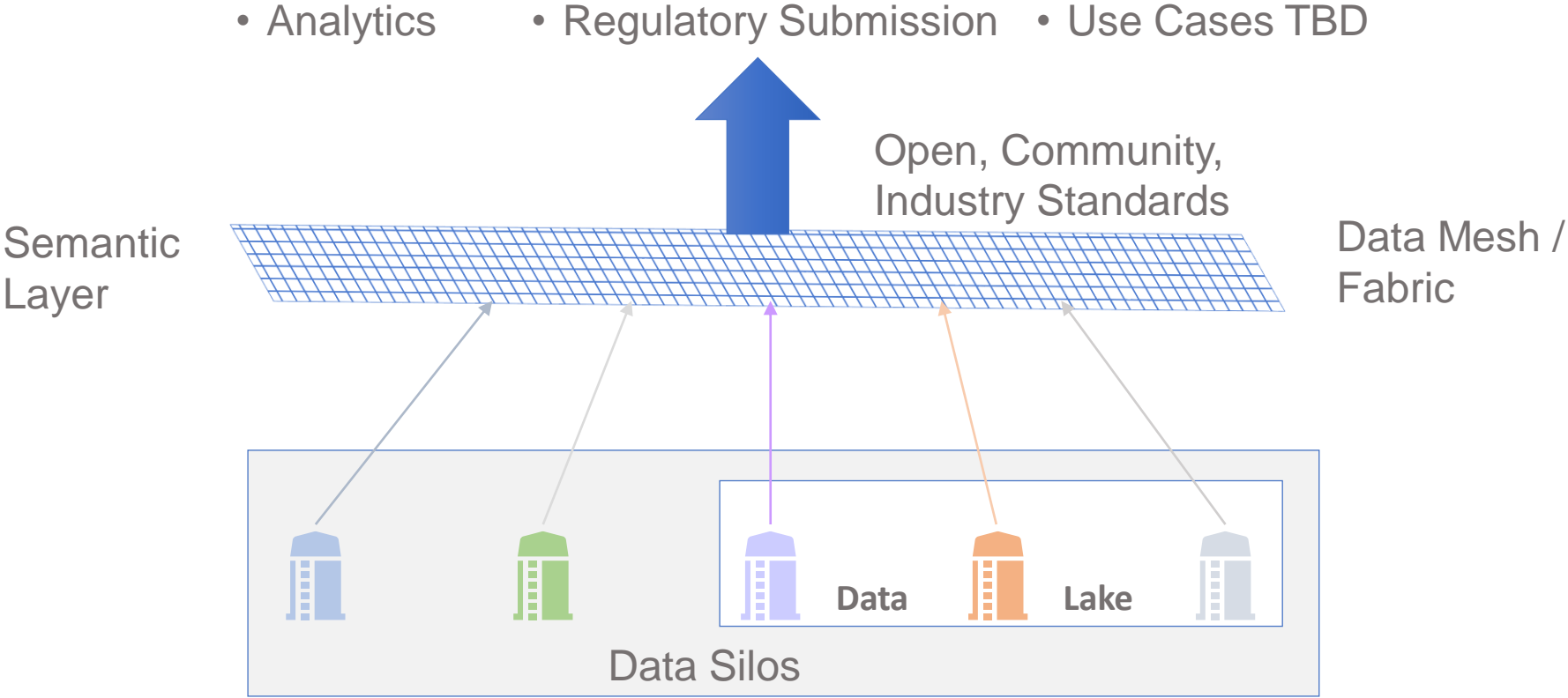
Data-Centric = Patient-Centric



Comprehensive Patient View



Semantic Data Mesh / Data Fabric



PHUSE Project

Study Data Validation & Submission Conformance

What is PHUSE?

Mission

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

Working Group Mission

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

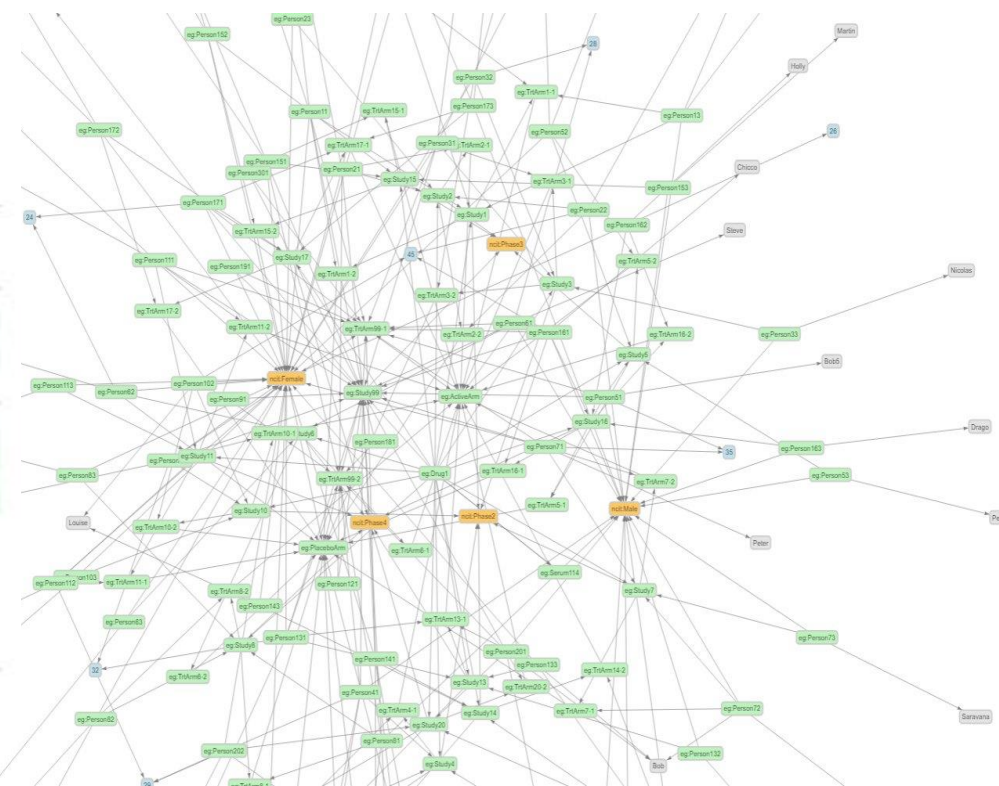
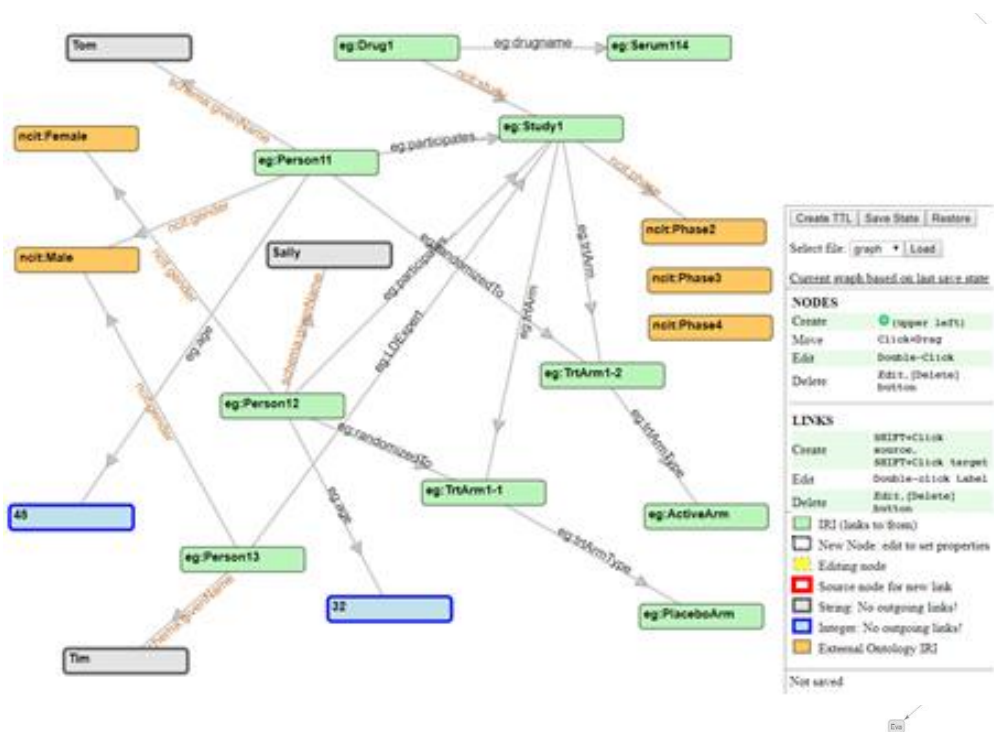
<https://www.phuse.eu/working-groups>

What is PHUSE?

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

Linked Data Workshop

A PHUSE, Stardog Partnership



PHUSE Linked Data Projects

- CDISC Foundational Standards in RDF
- CDISC Conformance Checks
- Reusing Medical Summaries for Enabling Clinical Research
- Regulatory Guidance in RDF
- Clinical Program Design in RDF
- CDISC Protocol Representation Model in RDF
- **Analysis Results & Metadata**
 - *RDF Data Cubes for clinical trial results*
- **Clinical Trials Data as RDF**
 - *Study Data Tabulation Model as Linked Data*
- **Going Translational with Linked Data**
- **Study Data Validation and Submission Conformance**
 - Pre-clinical data + submission metadata

Study Data Validation & Submission Conformance

A PHUSE Linked Data Project

Emerging Trends & Technologies Working Group

FDA Submissions

- 32% data conformance issues* (2016-2018)

Why so high? *[Personal Opinion]*





- A submission usually contains multiple studies
 - Data consistency and integration issues.
- Requirements: lack of clarity and understanding
- Contributing factor:
 - Legacy data structures
 - Lack of consistent identifiers
 - Poorly integrated metadata and validation rules

Application types:

- New Drug Application (NDA)
- Abbreviated New Drug Application (ANDA)
- Biologics License Applications (BLA)
- Commercial Investigational New Drug

* “Update on Technical Rejection Criteria for Study Data.”
- Ethan Chen, CDER. 3 April 2019.

Project Collaboration

- PHUSE 
- FDA (preclinical submissions) 
- Academia (pending/TBA) 
- Stardog Knowledge Graph 
- + YOU
 - Open Positions: Project Co-lead, Contributors

Study Data Validation

◦ Study Data Validation and Submission Conformance

- FDA Validation Rules for pre-clinical data
- **SHA**pes **C**onstraint **L**anguage (**SHACL**)



Study Subject
SHACL



Study Subject
Data Validation

Submission Conformance

◦ Study Data Validation and Submission Conformance

- Format and Content Completeness
- FDA Standards Catalog
- Increase automation of:
 - Submission metadata collection
 - Data validation
- Collaboration with Academia
- F.A.I.R. (Findable, Accessible, Interoperable, Reusable)

<https://github.com/phuse-org/SENDConform>

Project Website

◦ Study Data Validation and Submission Conformance

<http://bit.ly/SENDConform>

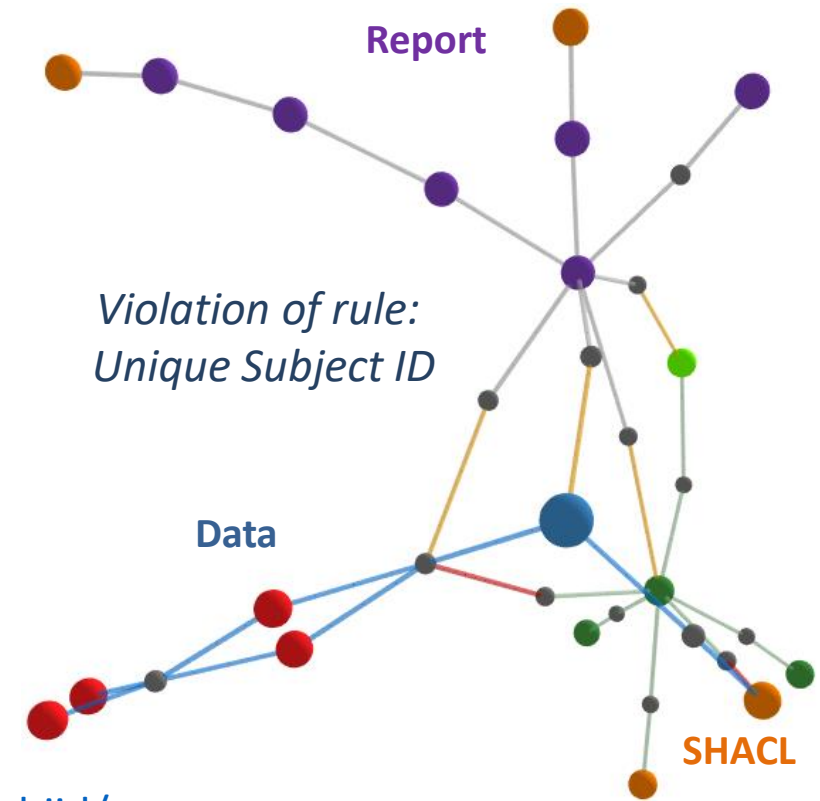
All together:

- Data
- Validation Rules
- Validation Report

Demonstration:

Visualization of USUBJID Rule

<https://phuse-org.github.io/SENDConform/visualization/usubjid/>



Industry Examples

Life Sciences & Beyond

- Life Sciences Linked Open Data

- <https://lod-cloud.net/#subclouds>

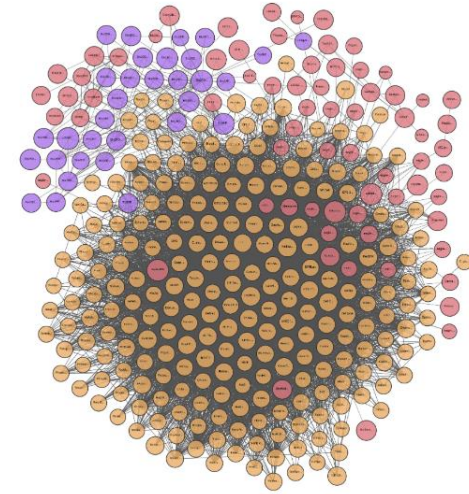
- Wikidata



- DBpedia



- BioPortal  (biomedical ontologies)



...and many others

F.A.I.R Data

◦ Findable ◦ Accessible ◦ Interoperable ◦ Reusable

<https://www.go-fair.org/fair-principles/>



- FAIR Implementation Project & Toolkit

<https://www.pistoiaalliance.org/projects/current-projects/fair-implementation/>

- Ontologies Mapping

<https://www.pistoiaalliance.org/projects/current-projects/ontologies-mapping/>



- FAIR Data Knowledge Graphs – From Theory to Practice

<https://youtu.be/Z0U2O2FjL6w>

Bayer



- Data-Centric Approach
 - Across three Life Sciences Divisions
- FAIR Data Lake
 - Enable future data science on legacy data assets
- Clinical Trial Design
 - Answer complex questions to improve design
- Permanent ID Service
 - Central registration for ID generation and curation
- ...and more.

Information courtesy of:

Dr. Alexander Krupp

Head of Global Data Assets – Pharma 360

Roche



- Common Models for Integration
- Global Data Standards Repository (GDSR)
 - Aligned with (clinical) CDISC Standards
 - Adheres to FAIR Principles
- Roche Terminology Service (RTS)
 - Semantically linked Domain Master for Terminologies
- Clinical Trial Data, Real World Data
- And more...

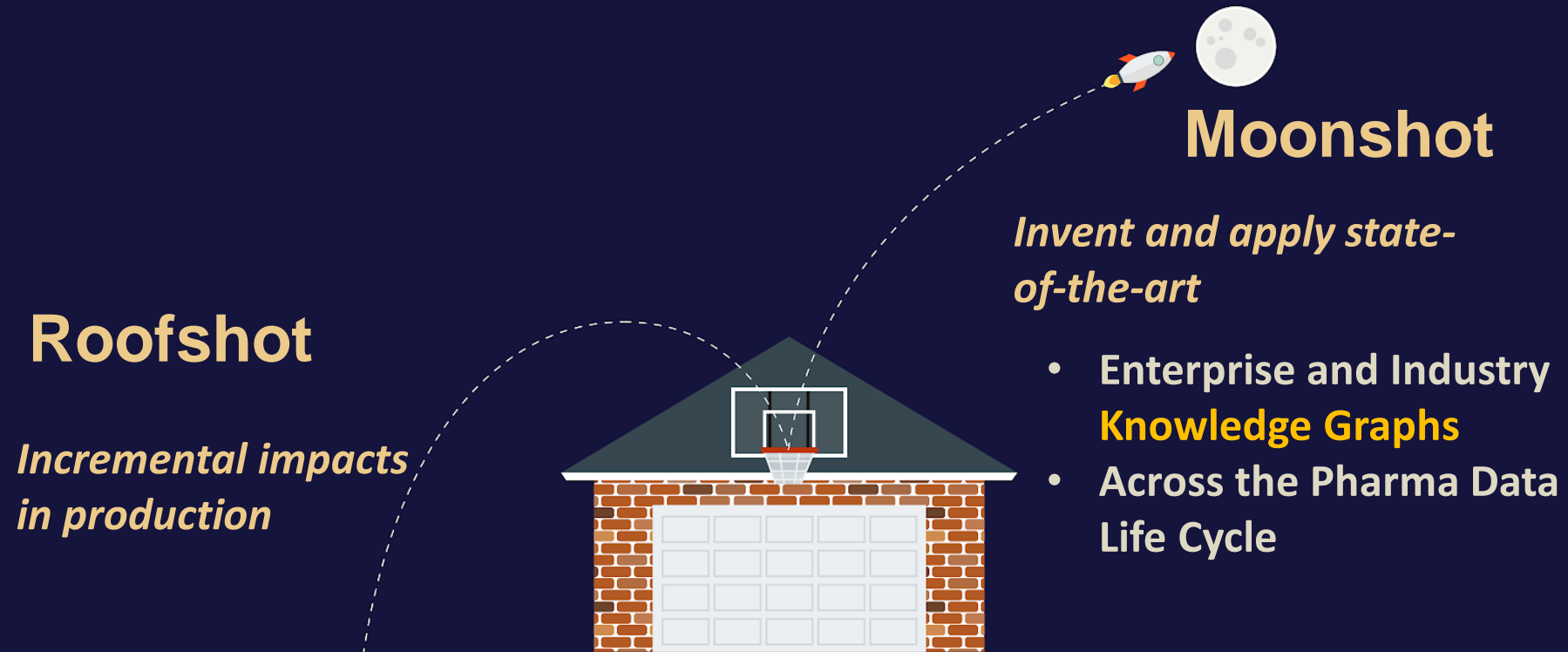
Information courtesy of:

Dr. Javier Fernandez

Medical Data & Information Solutions

Implementation Strategy

The Roofshot / Moonshot Manifesto



Examples

1. Unique Identifiers for Pharma
2. Validation Rules in SHACL for Pre-Clinical Data
3. Open Ontology Development



Industry Knowledge Graphs

Industry Standards & Models

Enterprise Knowledge Graphs

Results Data as RDF

Validation Rules in SHACL

Terminology and Coding

Study Protocol

Study Design

Unique Identifiers for Pharma

Prototype

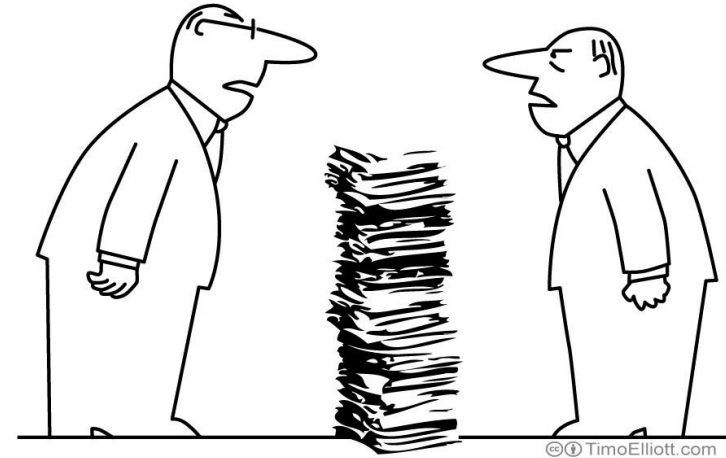
Demonstrations

The Stairway to the Stars Manifesto

Messaging

Data Owner

- *Implied silo*



"No, it's MY data!"

Data Steward

- Creation, oversight, sharing



Messaging

Proof of Concept

“A proof is a proof. What kind of a proof? It’s a proof. A proof is a proof. And when you have a good proof, it’s because it is proven.”

- Former Canadian Prime Minister **Jean Chretien**. 

- The technology is *proven*
- Better choices:
 - *Experiment*
 - *Demonstrate* the **Business Case** to the right audience.

Selling the Knowledge Graph Transformation

Relational



Robotic Process Automation

Machine Learning

Knowledge Graphs

Cloud

Artificial Intelligence

Digital Transformation

Data Mesh

Data Fabric

Relational + Graph



Who will Lead the Transformation?

Favor *status quo*

Legacy Vendors

Traditional Consultants

Contract Research Organizations (CROs)

Legacy Corporate IT

Standards Organizations

Regulatory Agencies

Favor Change

Graph Vendors

Data-centric Consultants

Progressive CROs

Enlightened IT

Standards Organizations (future)

Regulatory Agencies (future)

Research

- Drug Discovery
- Genomics
- Key Opinion Leader ID

Analytics

- Competitive Analysis
- Submissions
- Publishing
- Business
- Risk Management

Knowledge Graphs Provide

- Fewer, simpler data models
- Less
 - Data manipulation
 - Code
 - Manual data conversion & recoding
- Built-in:
 - Data integration
 - Metadata
 - Validation
- Flexible, incremental model building
- Follow-your-nose approach to information discovery

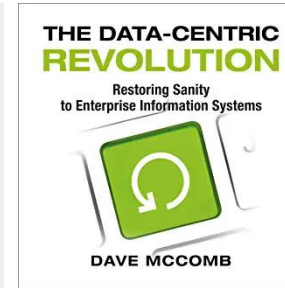
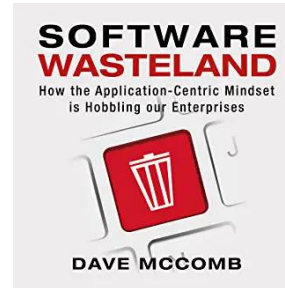
Impact

Paradigm shift to **data-centric** approach

- Higher data quality
- Successful submissions for new medicines
- New insights
- ***Faster delivery of affordable, safe therapies to patients***

Acknowledgements

- Dave McComb



Industry Initiatives

- Dr. Alexander Krupp
- Dr. Javier Fernandez



Thank you!

tim.williams@phuse.eu

Questions?