Toward F.A.I.R. Pharma

PhUSE Linked Data Initiatives Past and Present

Semantics @ Roche
2019-04-04 9:30-10:00

Outline

I. Introduction
   - What is PhUSE?

II. How FAIR is (late phase) Pharma?

III. Toward FAIR Pharma with PhUSE

Tim Williams

- Statistical Solutions Lead
- UCB Biosciences, Raleigh North Carolina

PhUSE

- Steering Committee, Computational Sciences Symposium (CSS)
- Co-lead
  - Analysis Results Model (RDF Data Cubes) (2016)
  - Clinical Trials Data as RDF (2018)
  - Going Translational with Linked Data (present)
- Instructor: Linked Data Hands-on Workshop

PhUSE

Pharmaceutical Users Software Exchange

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

PhUSE

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

PhUSE 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

II. How FAIR is (late phase) Pharma?

Magic mirror on the wall…

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

How Does Pharma Fare on FAIR?

Findability

F1. globally unique, persistent id

Human Study Subject "Bob"

- PharmaCo
  - Study 1, Drug A
  - Study 2, Drug B
- DrugCo
  - Study 3, Drug C

PhUSE Project:
"Study URI"

Merge Bob's data from all studies.

How Does Pharma Fare on FAIR?

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 processes.

How Does Pharma Fare on FAIR?

Interoperability

I1. shared language for knowledge

Lack

- *Knowledge representation*
- *language*

Linked Data / Knowledge Graph adoption is helping!

Core Challenge : Change of mindset

- **Currently:** Data modeled to industry standards for submission to regulatory authorities
- **Future:** Models of the process and the entities in the data.

How Does Pharma Fare on FAIR?

Reusability

R1.3 data meet domain-relevant community standards

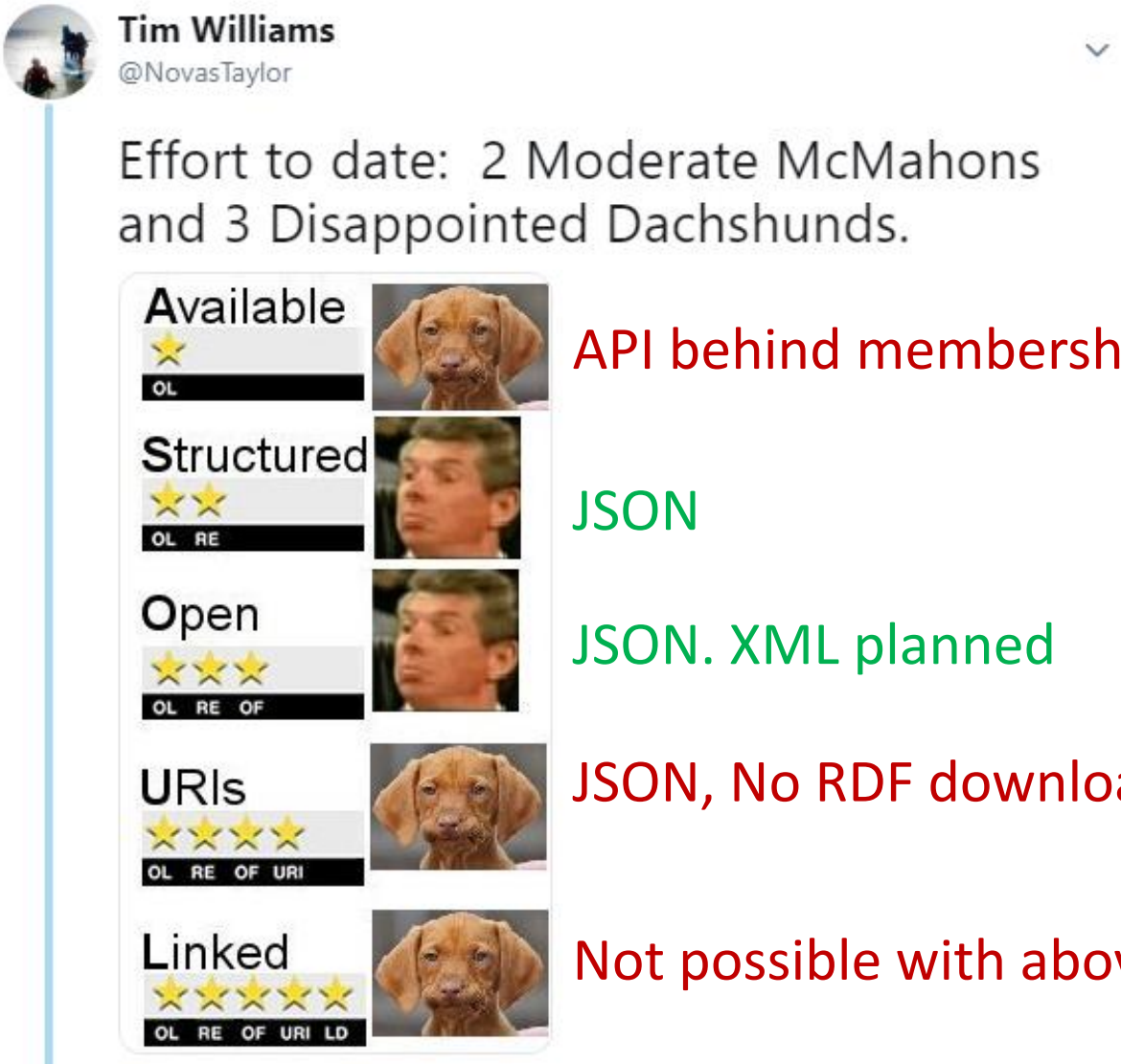
Positive:

- We have standards!

Negative:

- We have standards
  - Historically are row x column structure
  - New initiatives are not 5 Star Open or FAIR

New CDISC “Library” Initiative



API behind membership paywall

JSON

JSON. XML planned

JSON, No RDF download

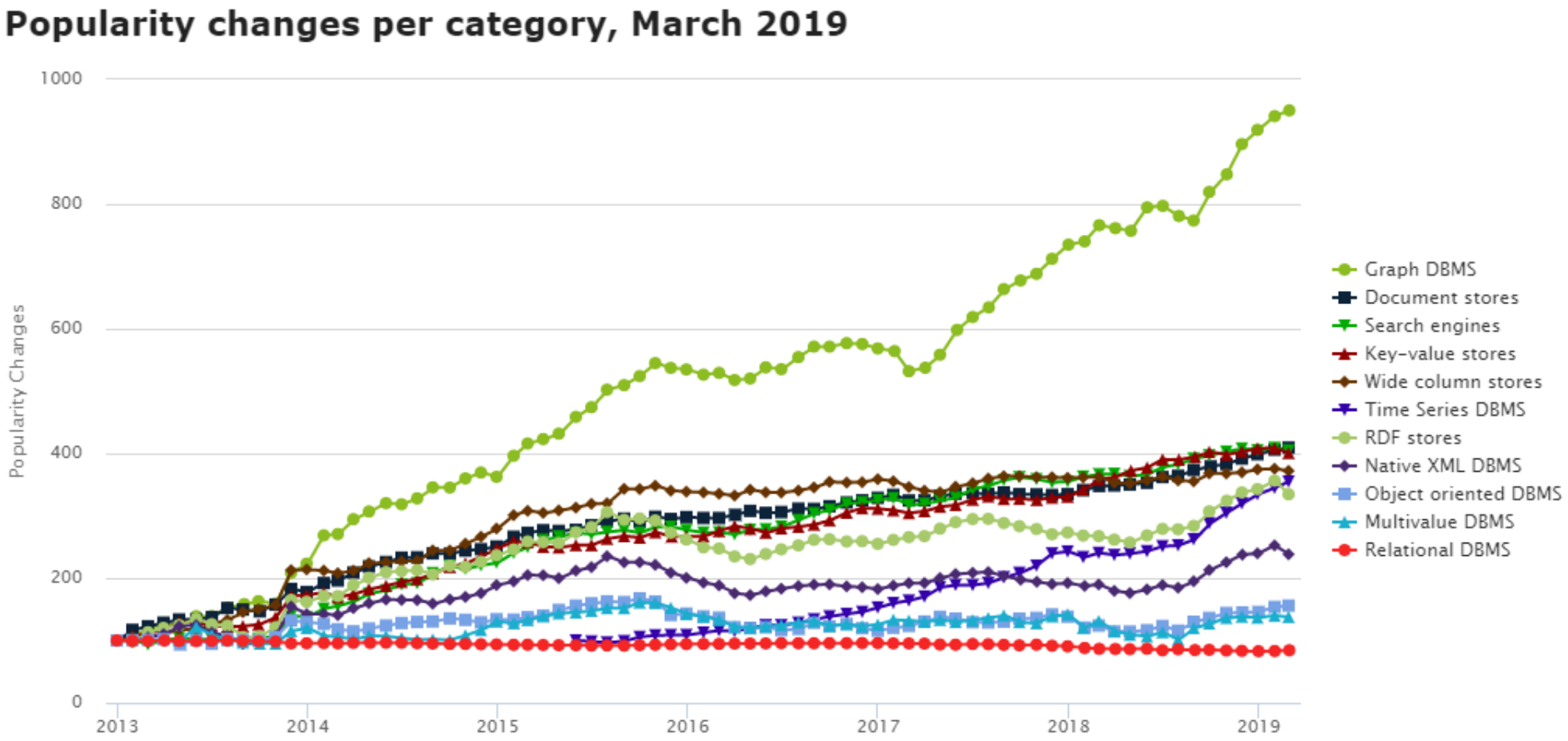
Not possible with above restrictions…

III. Toward FAIR Pharma with PhUSE

Risk-averse Pharma has a skills deficit

Database Popularity

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***?

CSS Workshop

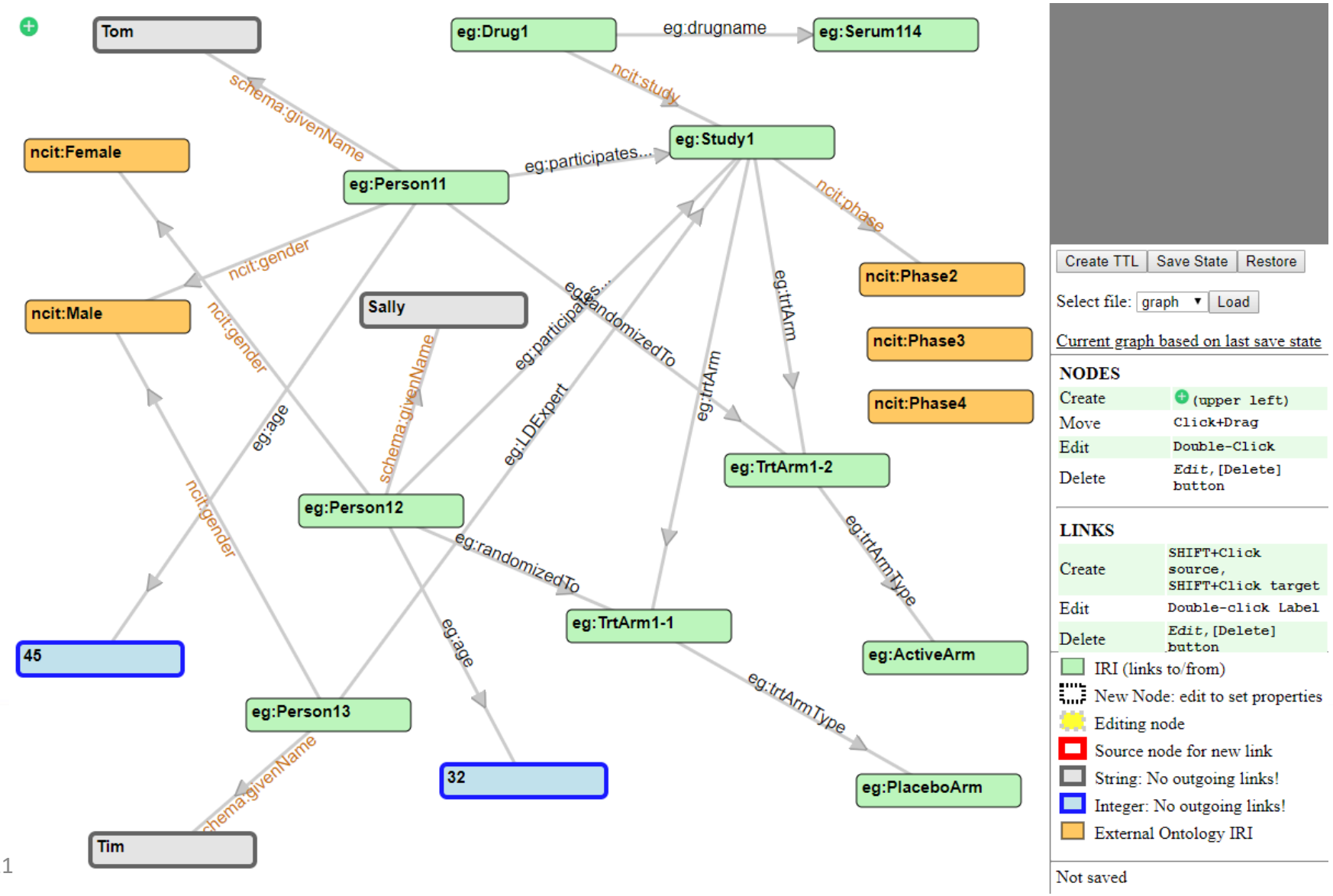
"Let's Make a Knowledge Graph!"

- Build your own RDF Linked Data Knowledge Graph
- Sunday Evening, June 9th

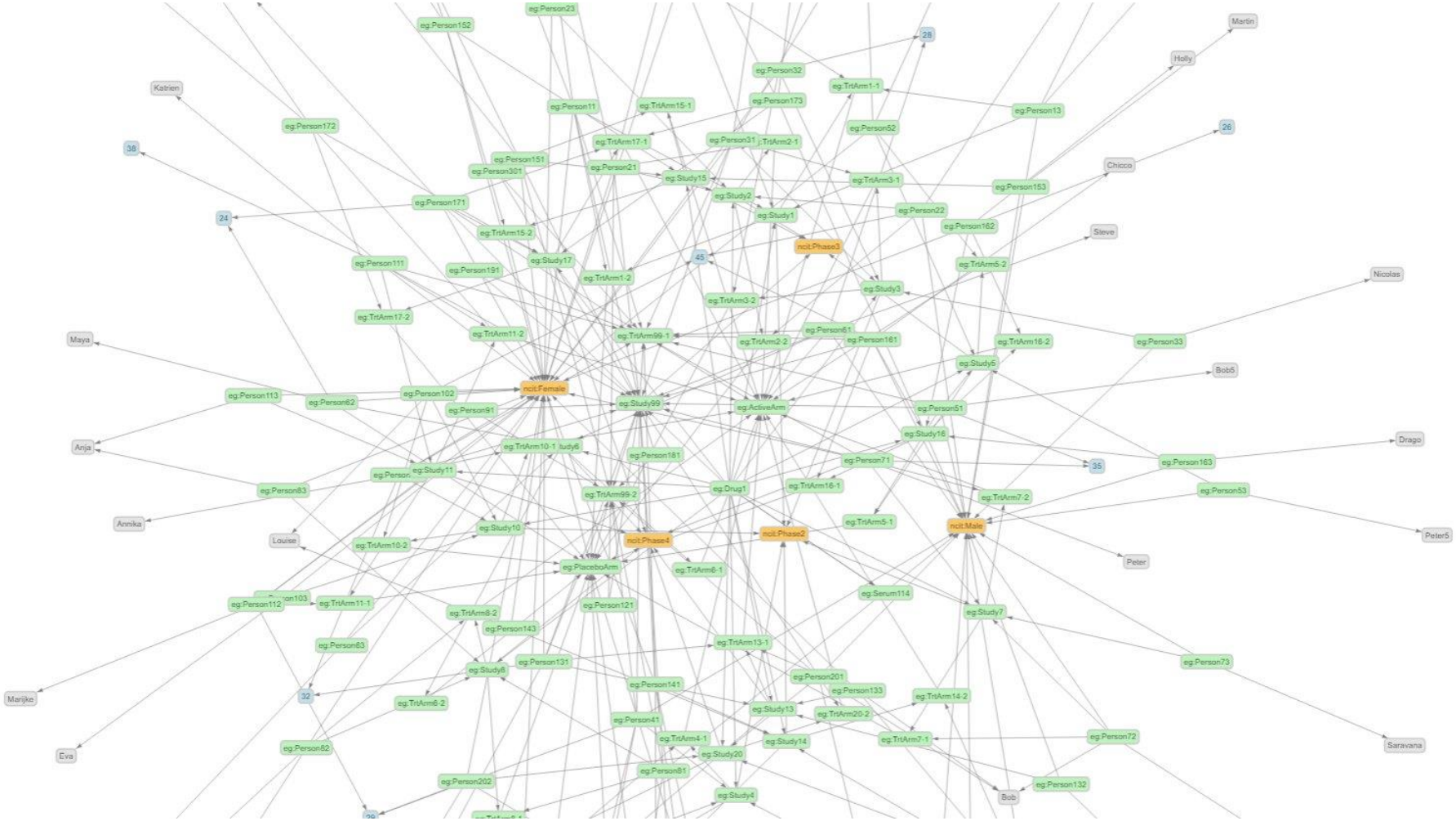
  https://www.phuse.eu/css19
- Introduction to Linked Data concepts

Also at PhUSE EUConnect, Amsterdam (November)

Knowledge Graph Workshop



Workshop : Merged Studies



Industry ~~Enterprise~~ Knowledge Graphs for Pharma

How?

The Roofshot / Moonshot Manifesto

Moonshot

Invent and apply state-of-the-art

- **Enterprise and Industry Knowledge Graphs**
- **Across the Pharma Data Life Cycle**

Roofshot

Incremental impacts in production

1. **Study URI**
2. **Clinical Trial Results Domains**
3. **Open Ontology Development**

Concept & Image Attribution: https://rework.withgoogle.com/blog/the-roofshot-manifesto/

Roofshot 1 : Study URI

Based on:

"Study URI" – K. Forsberg, D. Goude.
PhUSE EUConnect 18

- Easy entry point for Pharma
- Familiar Concept: NCT Number (clinicalTrials.gov), EudraCT number

Study URI

One URI to rule them all,

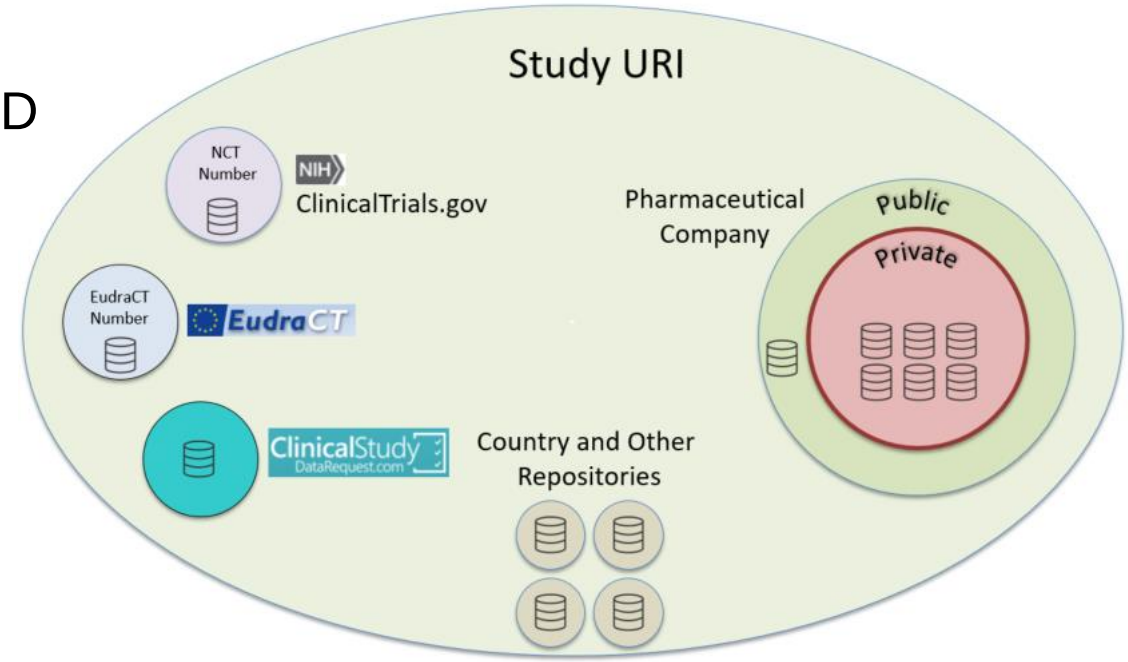
One URI to find them,

One URI to bring them all

and across repositories bind them.

Study URI

- A unique, immutable Study ID
- Links across repositories
- *Returns information*
  - *Context dependent!*



- Value for Patients, Researchers, Agencies, Companies

https://github.com/phuse-org/LinkedDataEducation/blob/master/doc/StudyURI.md

Wikidata as a Study URI Platform?

- Existing Infrastructure
- Community
- Process
- 27k studies already in Wikidata
  http://tinyurl.com/yxqlegy3

Study URI as a WikiProject?

- https://www.wikidata.org/wiki/Wikidata:WikiProjects
- Setup a 3 day workshop with stakeholders?

Roofshot 2 : Clinical Trial Results Domains

1. Clinical Trials Data (SDTM) as RDF (CTDasRDF)

2. Going Translational with Linked Data (GoTWLD)

Clinical Trials Data as RDF

Project Co-leads

Dr. Armando Oliva
Medical Informatics Consulting
Semantica LLC

Tim Williams
Statistical Solutions Lead
UCB Biosciences

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

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

Data Challenges

- Submissions non-conformance*
  - 32% of submissions at least 1 conformance flag
  - 20% of uploads to JANUS fail
- Variability in standards implementation
  - Multiple interpretations of implementation guides
- Version-conversion problems & costs
- Lack of intrinsic metadata
- Challenges linking data and standards
- **Limitations in the row-by-column data model**

* Conservative estimates

Machine-readable, Machine-interpretable Data

With built-in:

- **Meaning** (semantics)

  Interpret

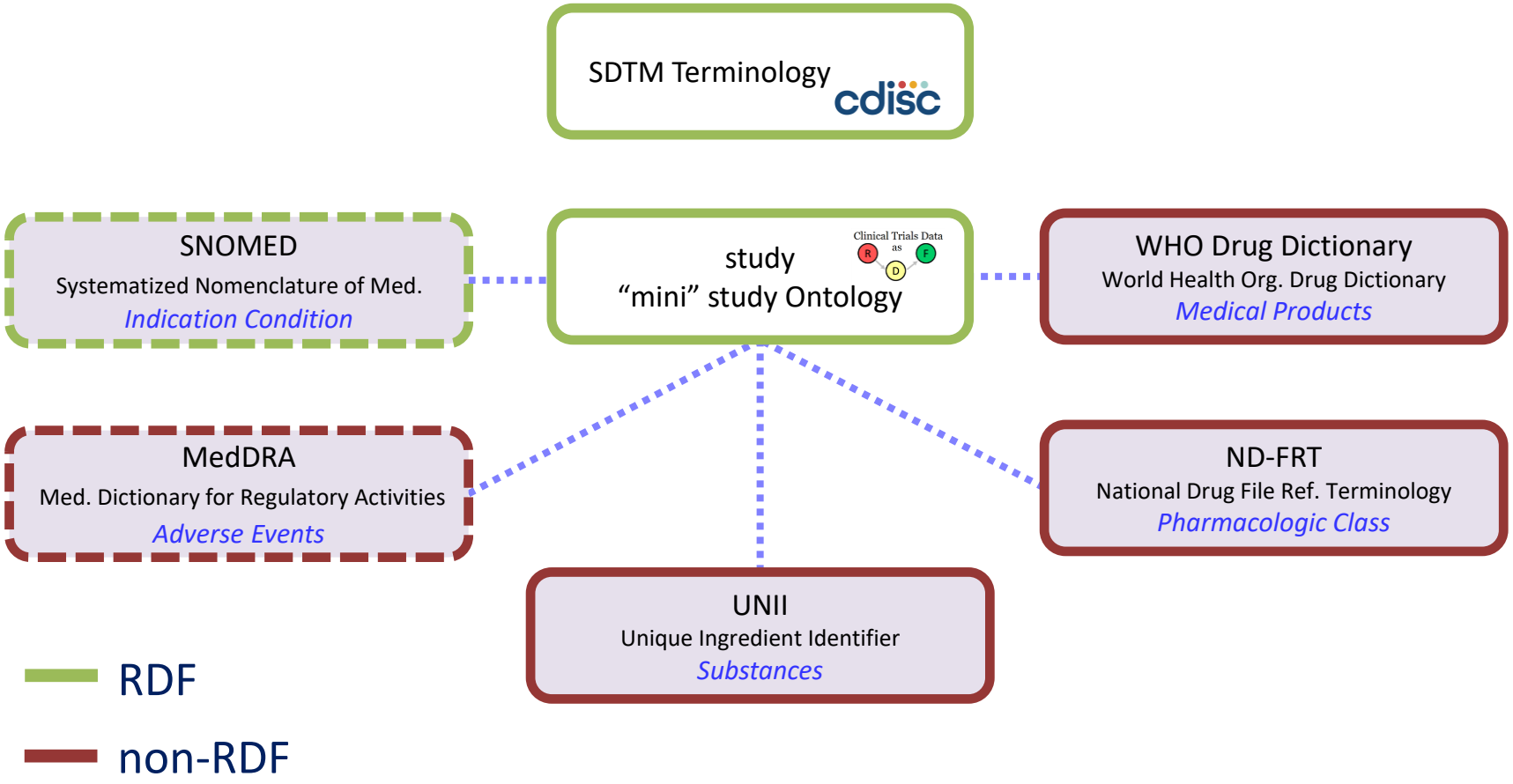
  - Context
  - Content
  - Structure
  - Purpose
- **Rules + Traceability**
  - Validity → Trustworthiness

Project Philosophy

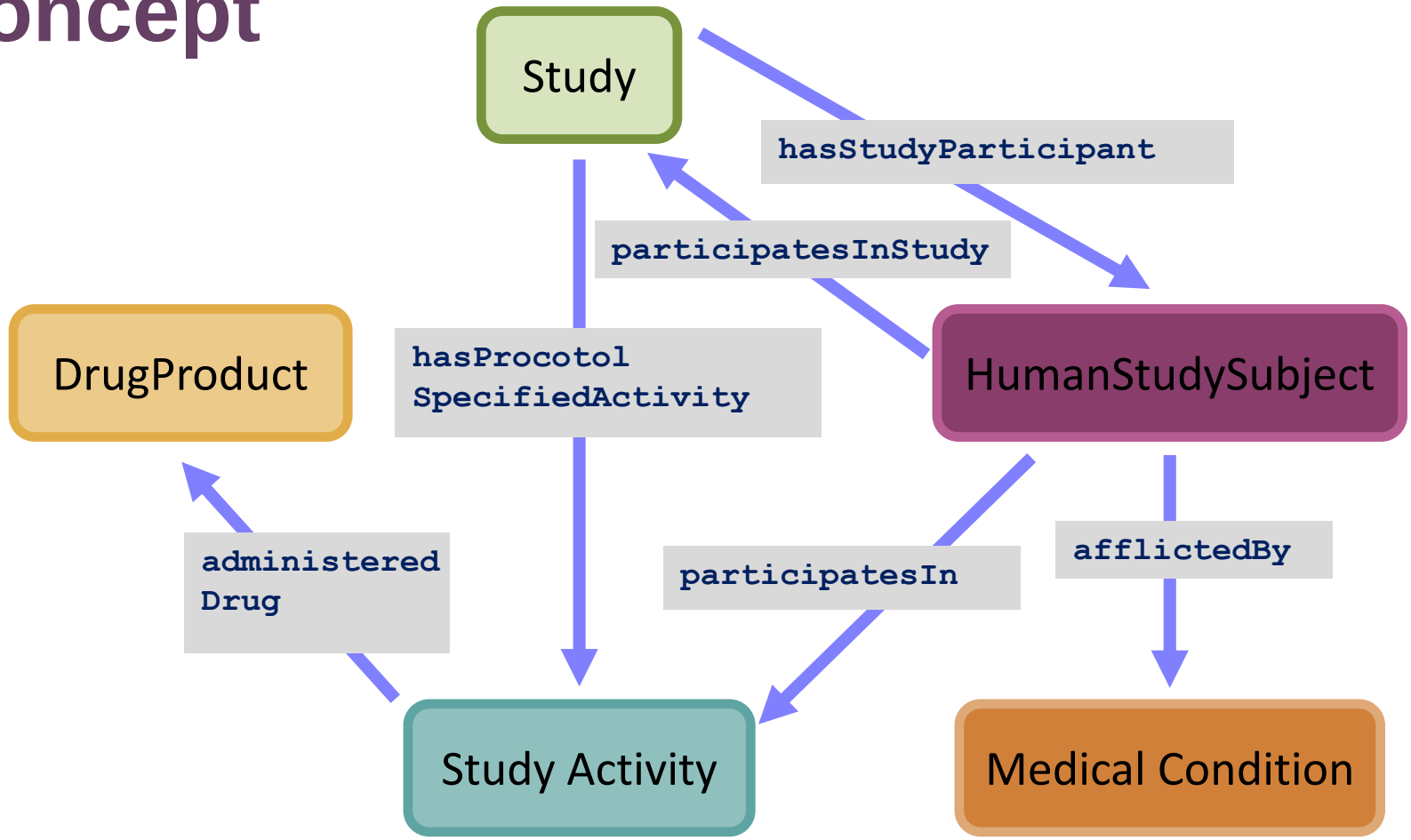
Resource Description Framework (RDF)

- Model the ***concepts*** represented in SDTM
- Map ***instance data*** to the graph model
- Re-use where possible
  - Ontologies, Terminologies
- **Open Source, pre-competitive, cooperative environment**
  - *Results available to everyone*

Re-use Existing Sources



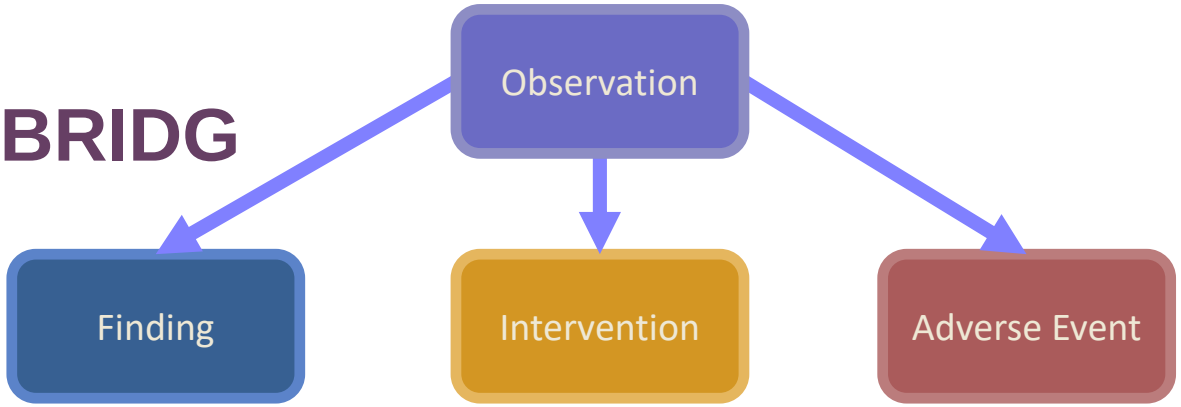
Concept


Study
hasStudyParticipant
participatesInStudy
DrugProduct
hasProcotol SpecifiedActivity
HumanStudySubject
administered Drug
participatesIn
afflictedBy
Study Activity
Medical Condition

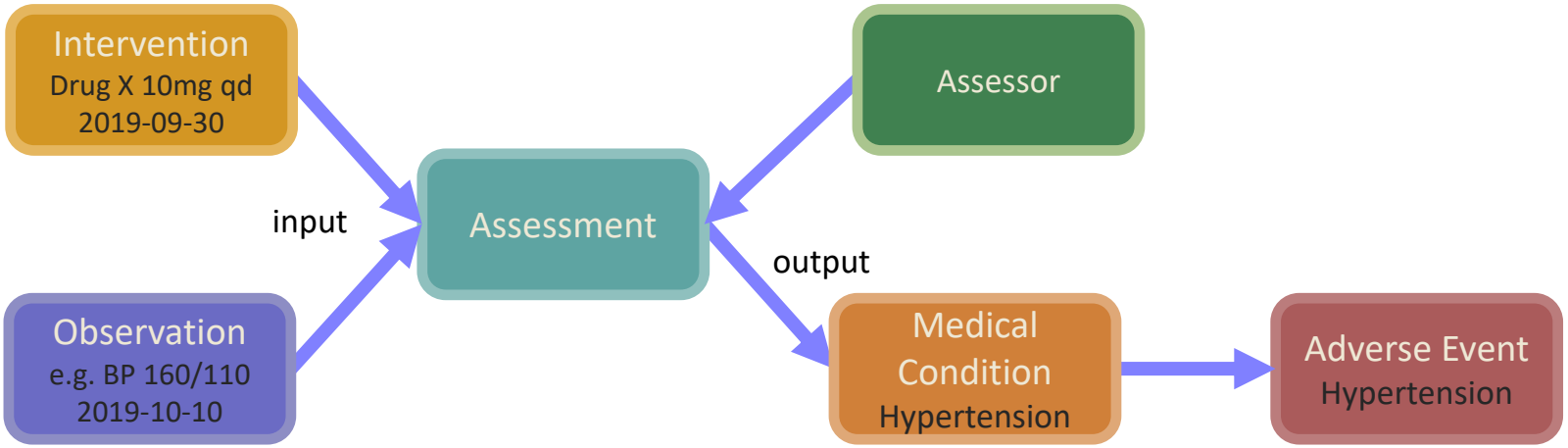
Adverse Event Representation

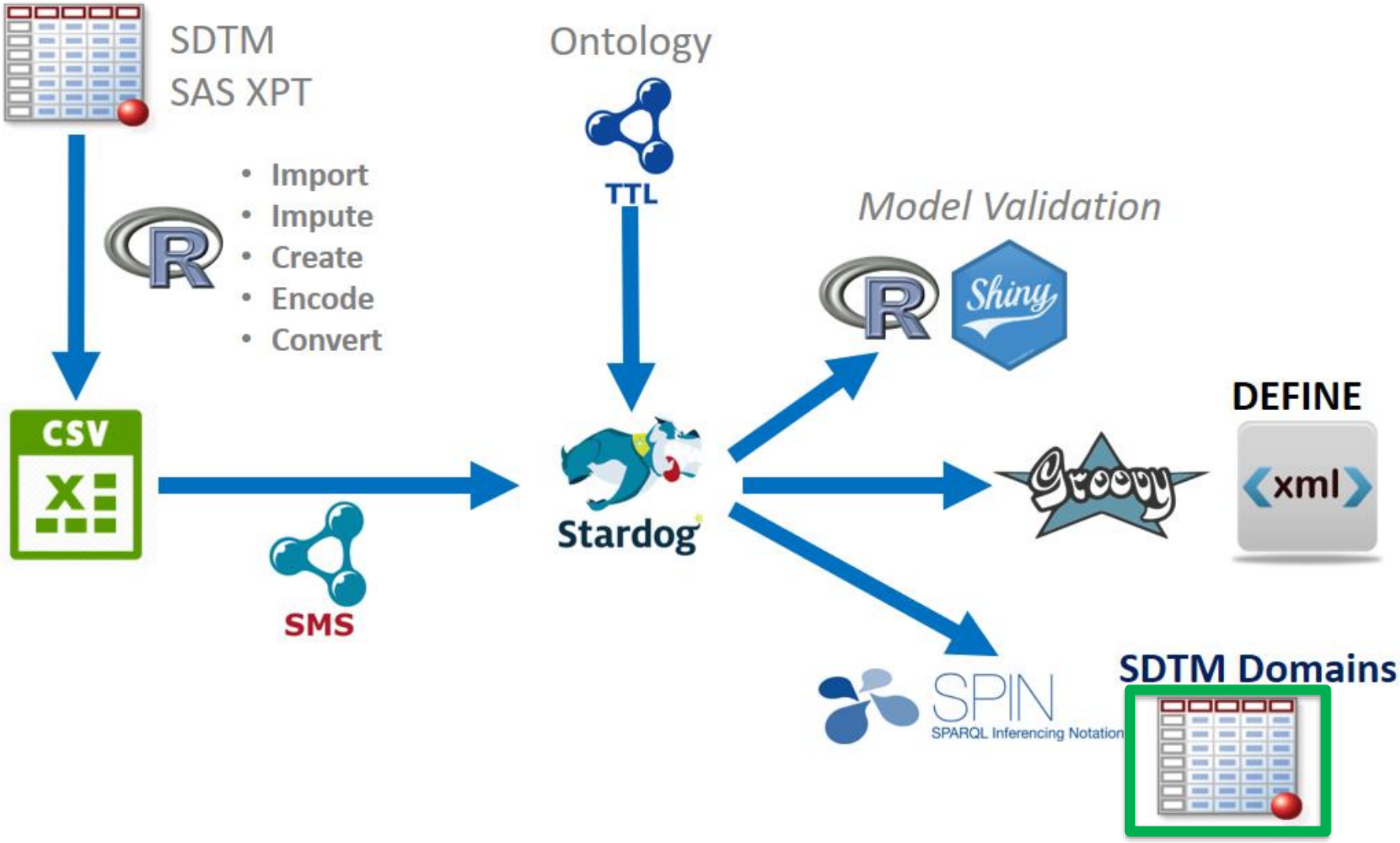
SDTM & BRIDG	Our Model
Observation	Medical Condition temporally associated with an Intervention

SDTM, BRIDG



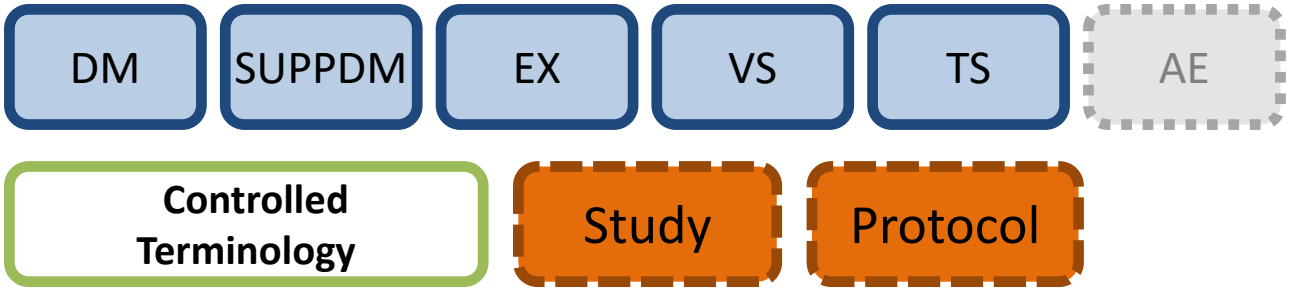
Our Model




SDTM
SAS XPT
Ontology
TTL
Import
Impute
Create
Encode
Convert
Model Validation
Shiny
CSV
DEFINE
xml
Stardog
SMS
Groovy
SPIN
SPARQL Inferencing Notation
SDTM Domains

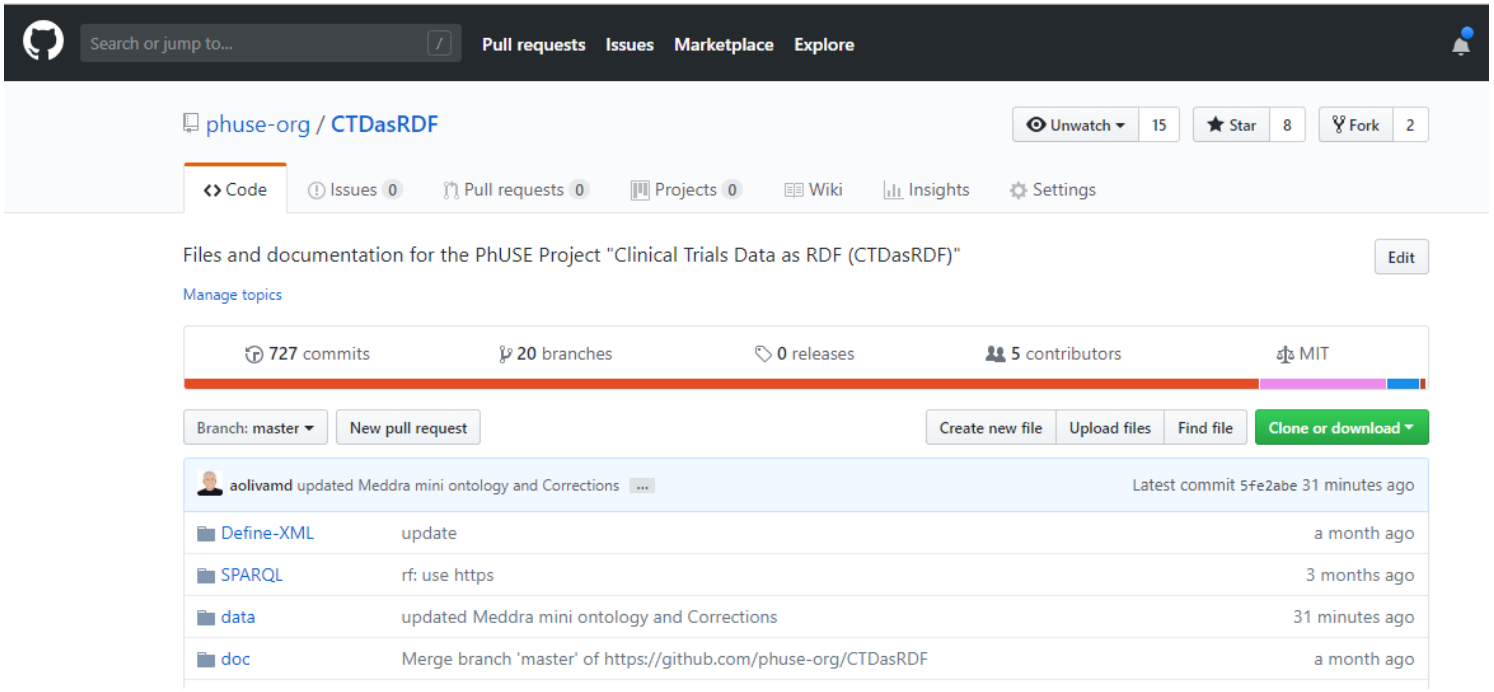
SDTM Modeling and Data Conversion

Clinical Trials Data as RDF



Project on GitHub

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



Going Translational with Linked Data

Going Translational with Linked Data

Project Co-leads

Dr. Armando Oliva
Medical Informatics Consulting
Semantica LLC

Tim Williams
Statistical Solutions Lead
UCB Biosciences

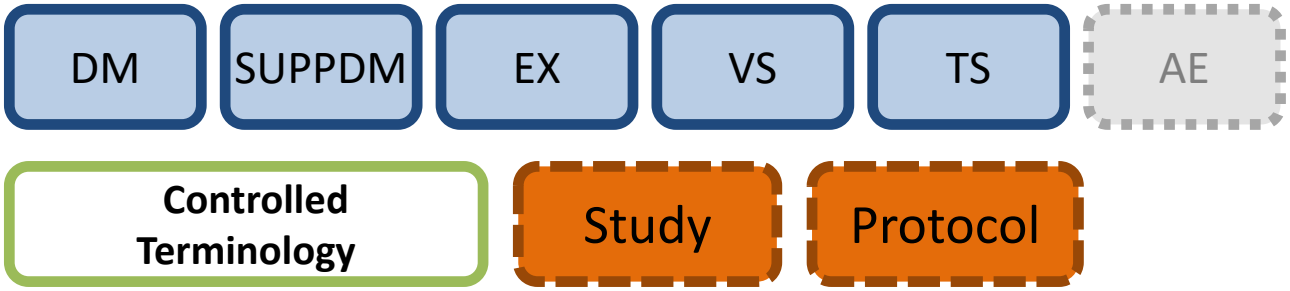
Drashtti Vasant
R&D IT Business Partner
Translational Sciences at Bayer Business Services

Project Focus

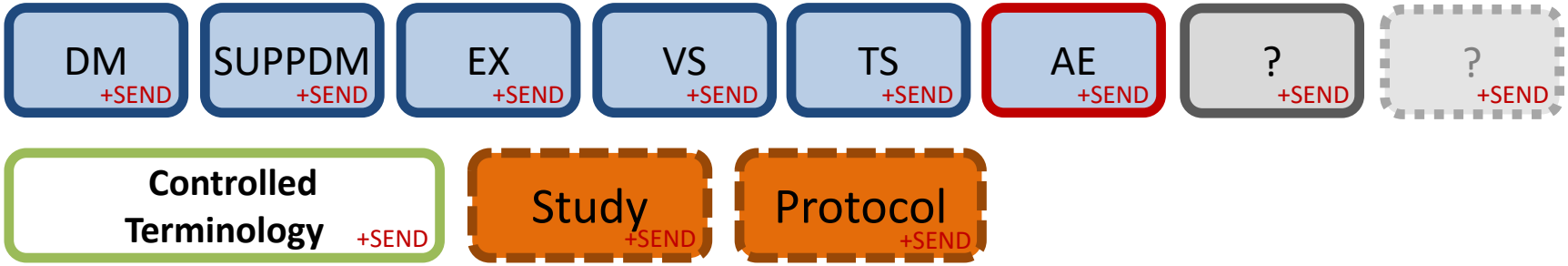
- Ontology Development
- Data Conversion to RDF
- Define XML
- Non-clinical models & data [new]

Modeling and Data Conversion

Clinical Trials Data as RDF



Going Translational With Linked Data



Deliverables

Objective	Timeline
Project Initiation	February 2019
Supporting Ontologies	PhUSE CSS 2020
Instance Data, Define XML	PhUSE CSS 2020
Presentations	PhUSE CSS 2019, 2020 PhUSE EUConnect 2019
Conclusion	PhUSE CSS 2020

Additional Initiatives!

- SPARQL Endpoint for Project Data
  - Query data online, from your desktop

Related Philosophy:
Current : Data ***Submission***
Future : Data ***Sharing***

- MedDRA Modeling and Conversion Process

Roofshot 3: Cooperative Ontology Development

- The challenges cannot be solved by any one person, department, company, agency, or organization.
- ***We must cooperate as an industry.***

Existing Ontologies

How do we ‘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
- OBO Foundry
- BioPortal
- ….others?

Ontologies must be open and Accessible

Don’t hide my OWL !

Conclusion

Linked Data Adoption in Pharma

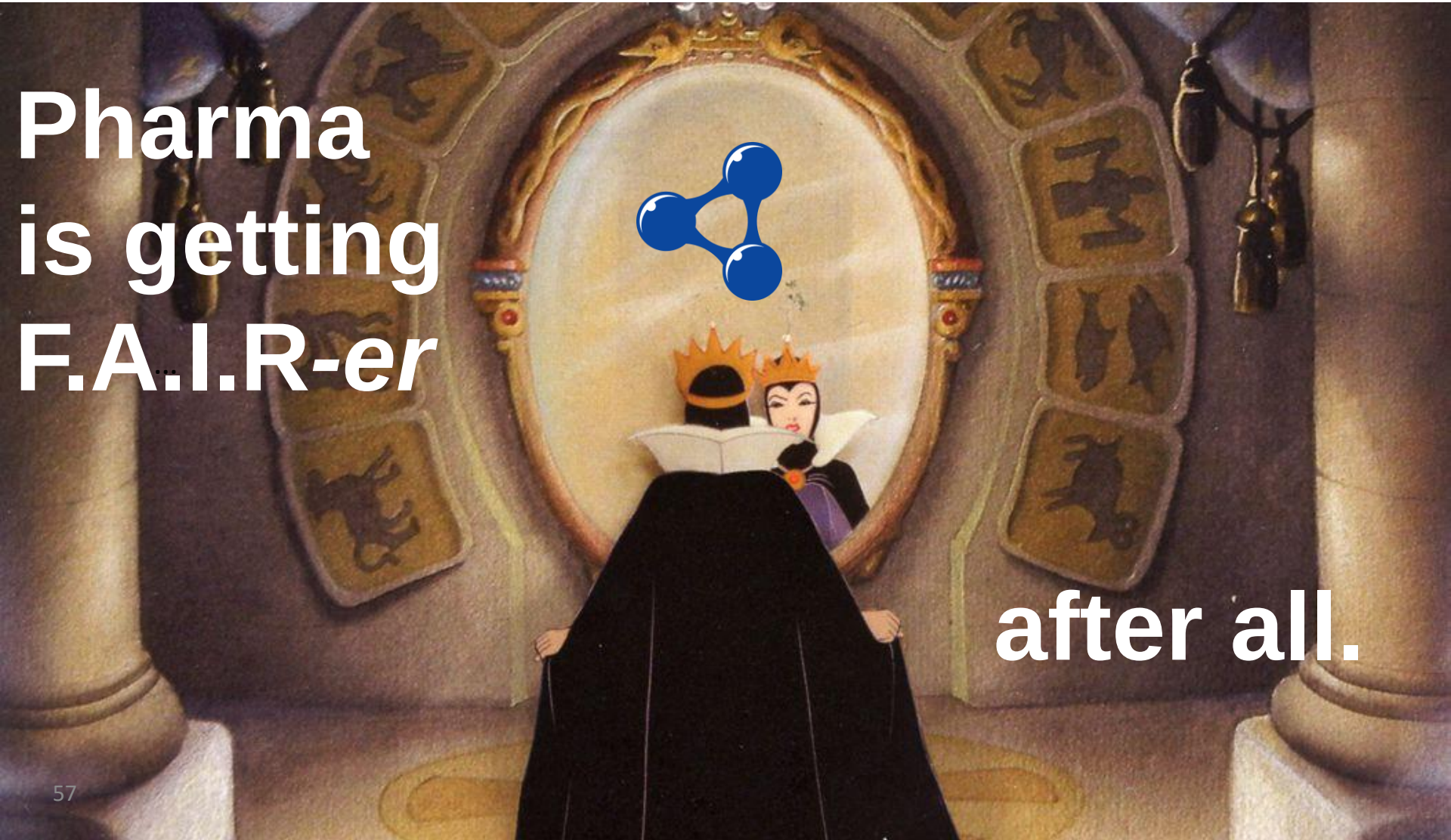
- Companies adopting and adapting CTDasRDF outputs
- Knowledge Graph initiatives at:
  - AstraZeneca
  - Bayer
  - Roche
  - Sanofi
  - ….and more
- External datasets published as RDF
  - EBI RDF platform
  - openPhacts
  - OpenTargets
  - …and more

The Future

Knowledge Graphs are here to stay

"The application of graph processing and graph DBMSs will grow at 100 percent annually through 2022..." – Gartner "Top 10 Data and Analytics Technology Trends for 2019"

Magic mirror on the wall…



Thank you!

Contact

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