Knowledge Graphs

Shaping our Data Future

Accenture Lunch & Learn
2020-12-08

Presenter

Tim Williams

- Knowledge Graph Project Lead, PHUSE
- Lead Statistical Solutions Analyst, UCB Biosciences

“Connecting things to other things since ~ 2013”

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

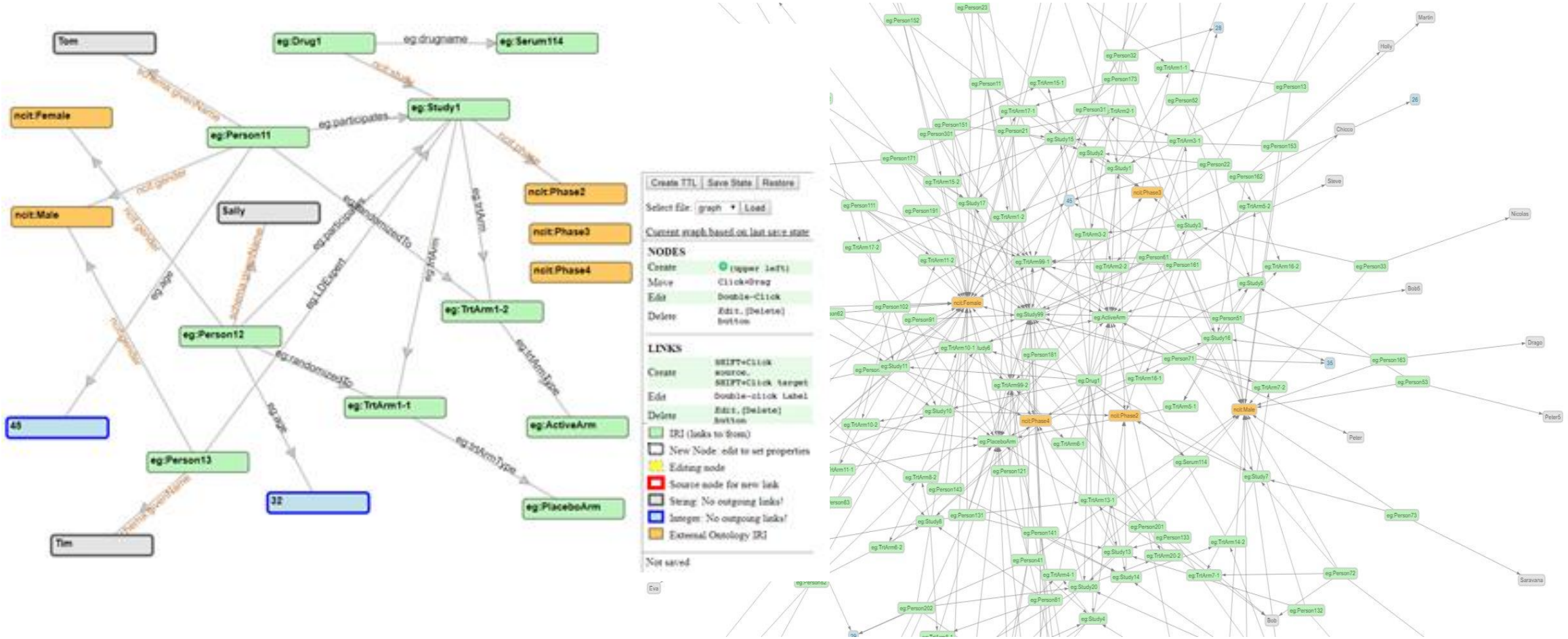
Pharma and biotech collaboration in the pre-competitive space:

- Standards
- Technology
- Regulations

https://www.phuse.global

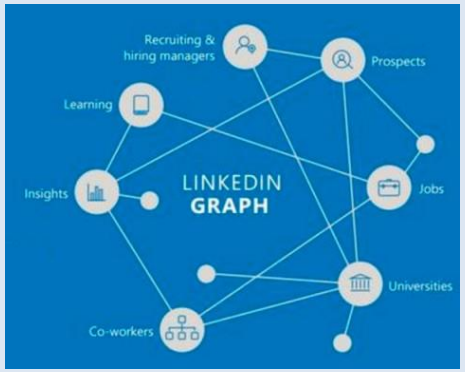
***Tim W. as co-lead**

PHUSE Linked Data Workshop



Outline

- Why Now?
- What is a Knowledge Graph?
- Data has *Shape*
- Validation has *Shape*
- Strategies for Implementation
- Open Discussion



Walmart

ebay

amazon

BEST BUY

Roche

BAYER

Boehringer Ingelheim

AMGEN

AstraZeneca

NOVARTIS

Amazon Neptune

neo4j

CAMBRIDGE SEMANTICS

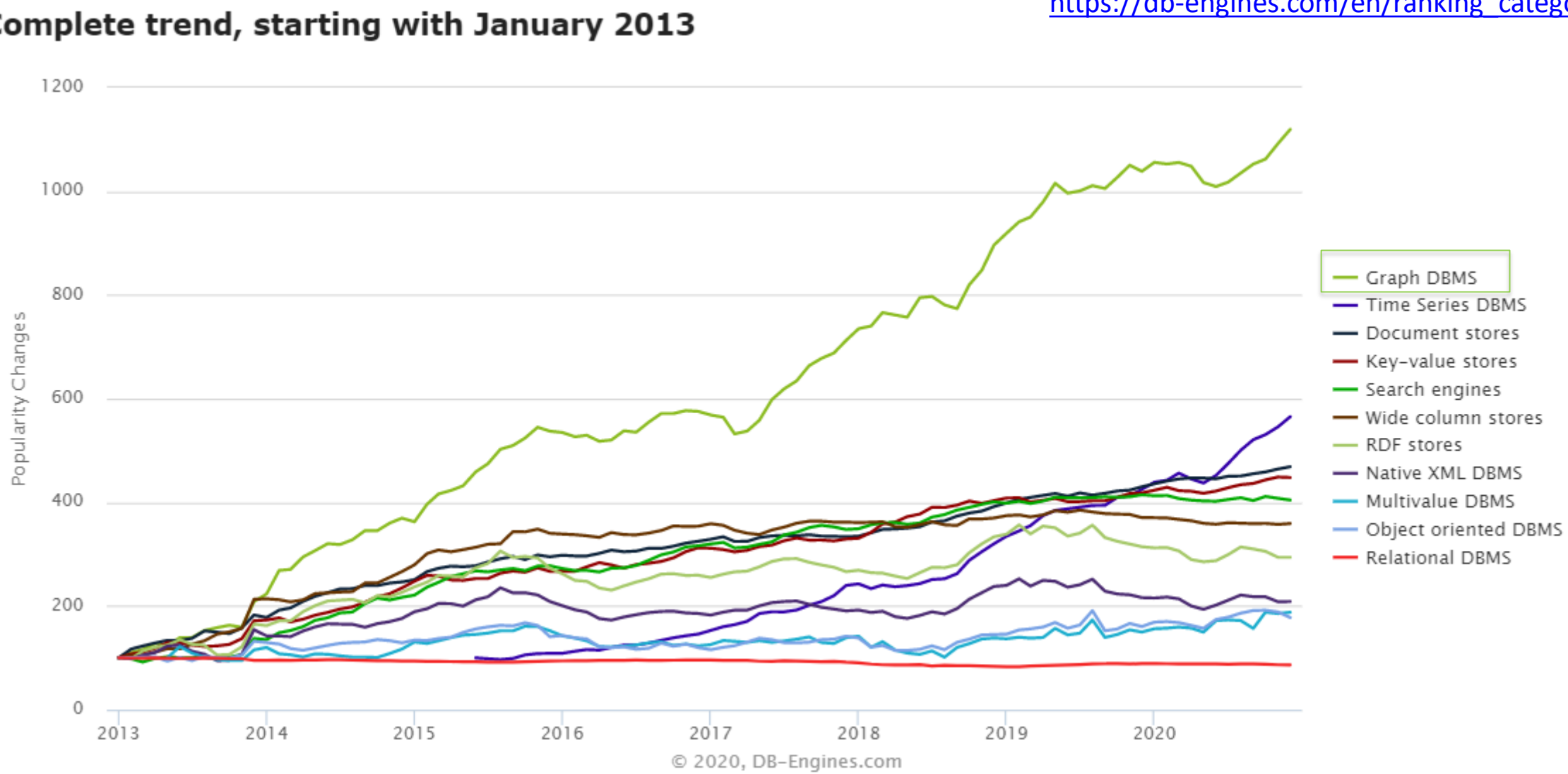
ArangoDB

data.world

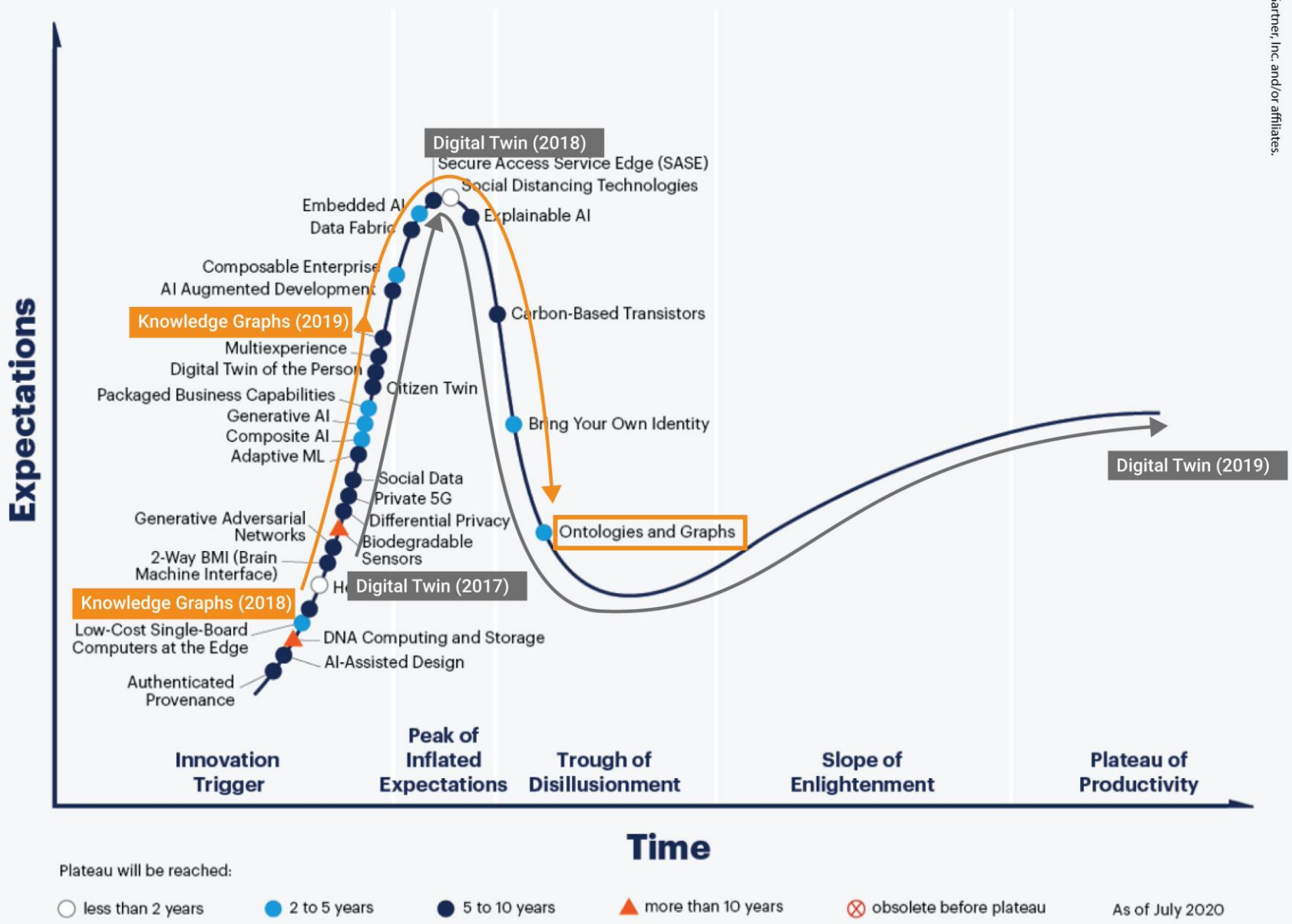
Stardog

Database Popularity Changes by Category

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



Hype Cycle for Emerging Technologies, 2020



FAIR Data Principles

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

Findable F1. Globally unique, persistent id	Merge data across diverse sources. Quality: Define once, reuse.
Accessible A2. Metadata available when data is not	Source data is not available. Who created it?
Interoperable I1. Shared language for knowledge	Share and interpret each other's data (both person and machine). Common format, structure, meaning
Reusable R1.3 (Meta) data meet domain-relevant community standards	Meet minimal standards for your data community (data type, organization, file format, documentation, vocabulary…) to facilitate re-use.

FAIR Data *is* Linked Data *is a* Knowledge Graph

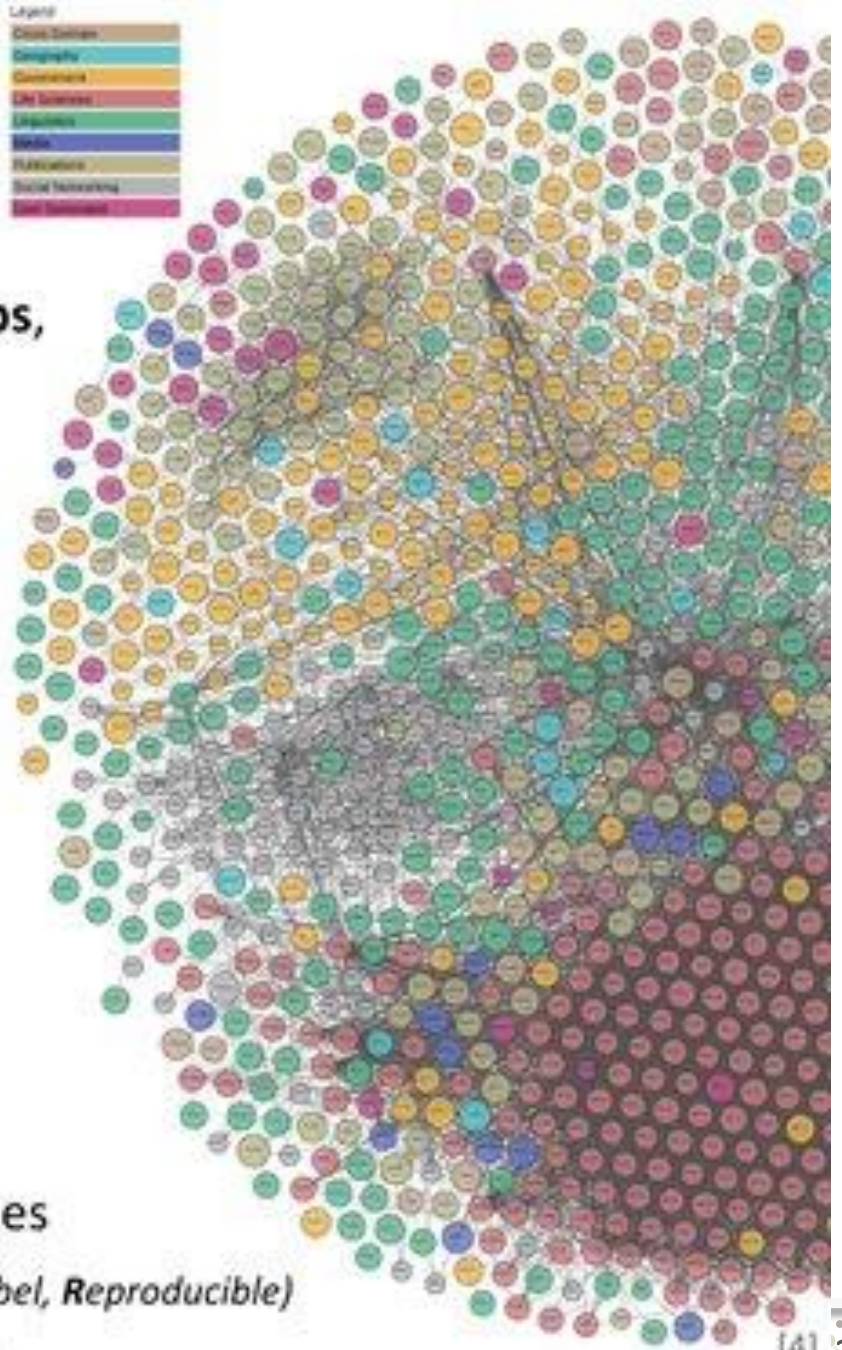
FAIR Data

Myth: FAIR only applies to open data on the web.

Fact: FAIR principles have great benefit “behind the firewall.”

What is a Knowledge Graph?

Knowledge Graphs - A Definition



- A **Graph** consisting of **concepts**, **classes**, **properties**, **relationships**, and **entity descriptions**
- Based on **formal knowledge representations** (RDF(S), OWL)
- Data can be **open** (e.g. DBpedia, WikiData), **private** (e.g. supply chain data), or **closed** (e.g. product models)
- Data can be **original**, **derived**, or **aggregated**
- We distinguish
  - **instance data** (ground truth),
  - **schema data** (vocabularies, ontologies)
  - **metadata** (e.g. provenance, versioning, licensing)
- **Taxonomies** are used to categorize entities
- **Links** exist between internal and external data
- Including **mappings** to data stored in other systems and databases
- *Fully compliant to **FAIR Data principles** (Findable, Accessible, Interoperabel, Reproducible)*

Knowledge Graphs Facilitate Standards

- Shared Definitions and Understanding
  - What is a Tobacco Product?
  - Who is a Tobacco User?
- Coding of
  - Medical Conditions
  - Adverse Events
  - Products, Manufacturers
- Data Classification, Rules, Validation..
- Ontology Driven/Supported

Knowledge Graphs Provide

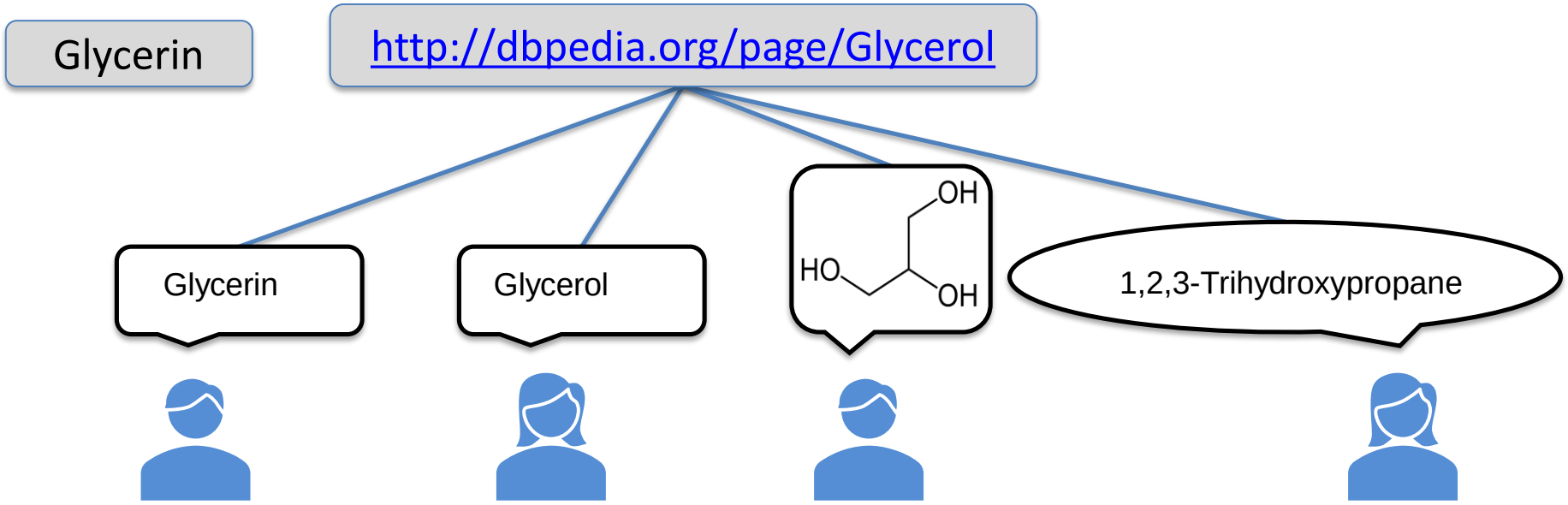
- Fewer, simpler data models
- Less
  - Data manipulation
  - Code
  - Manual data conversion & recoding
- Built-in:
  - Data integration
  - Metadata
  - Validation
- Flexible, incremental model building
- Follow-your-nose approach to information discovery

KG's Facilitate Common Understanding

- Devices
- Manufacturers
- Products
- Ingredients

RDF

- Uniform Resource Identifier (URI)
- Internationalized Resource Identifier (IRI)



- Common terminology
- Link to other Knowledge Graphs

Labeled Property Graph or RDF Triple Store?

We will not have this fight today.

Property Graph versus RDF Graph*

Property Graph : Designed for Graph Analytics

- Clustering, Path Analysis
- Lack standardization, formal semantics, schema language

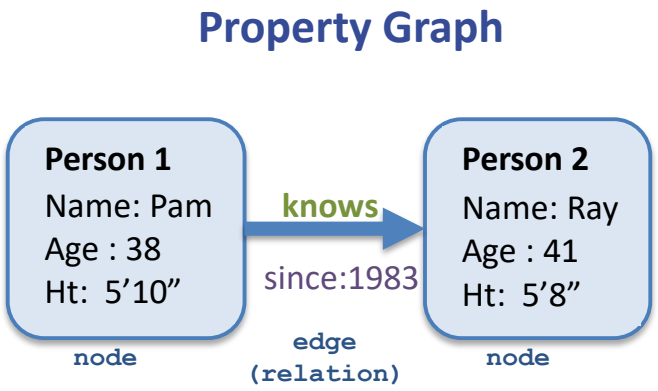
RDF Graph : A Stack of W3C Standards (RDF, OWL, SPARQL, SHACL…)

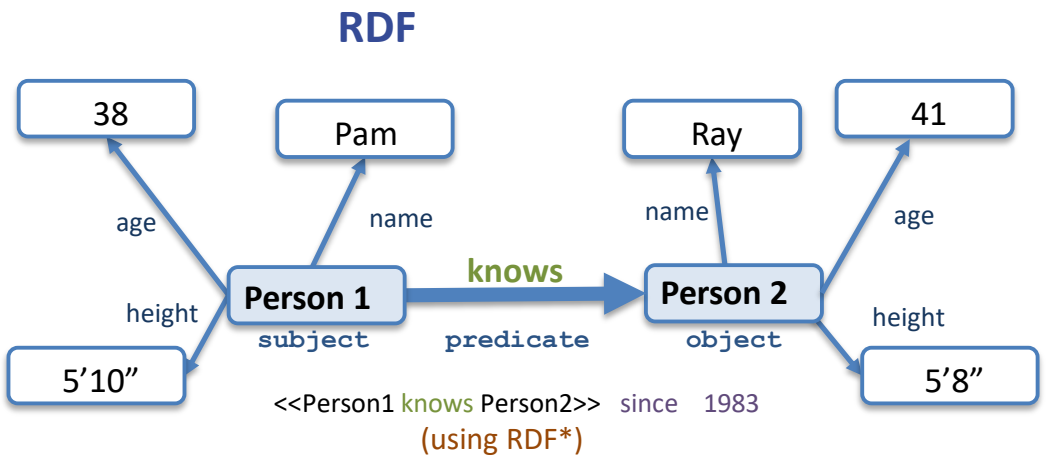
- Global identifiers (URI/IRI) => **Interoperability**
- Formal Semantics => **Common Meaning** (for the person + the machine)
- Standardization => **Unification**
- Validation => **Quality**

* Ontotext @ KGC 2020

Property Graph

RDF “Triples”





The Whiteboard model ***is*** the Data Model ***is*** the Data

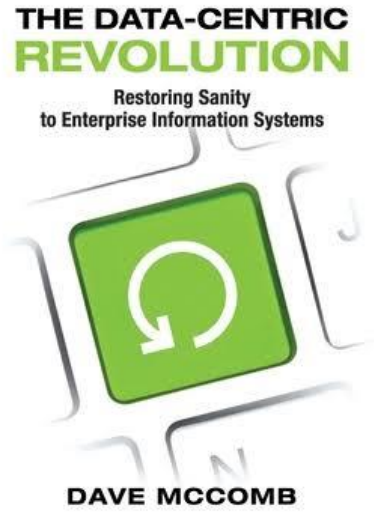
Ontology-based KG Development

1. Define the Model
- Based on an Ontology
- Entities and their Relationships

2. *Instantiate* data to the model

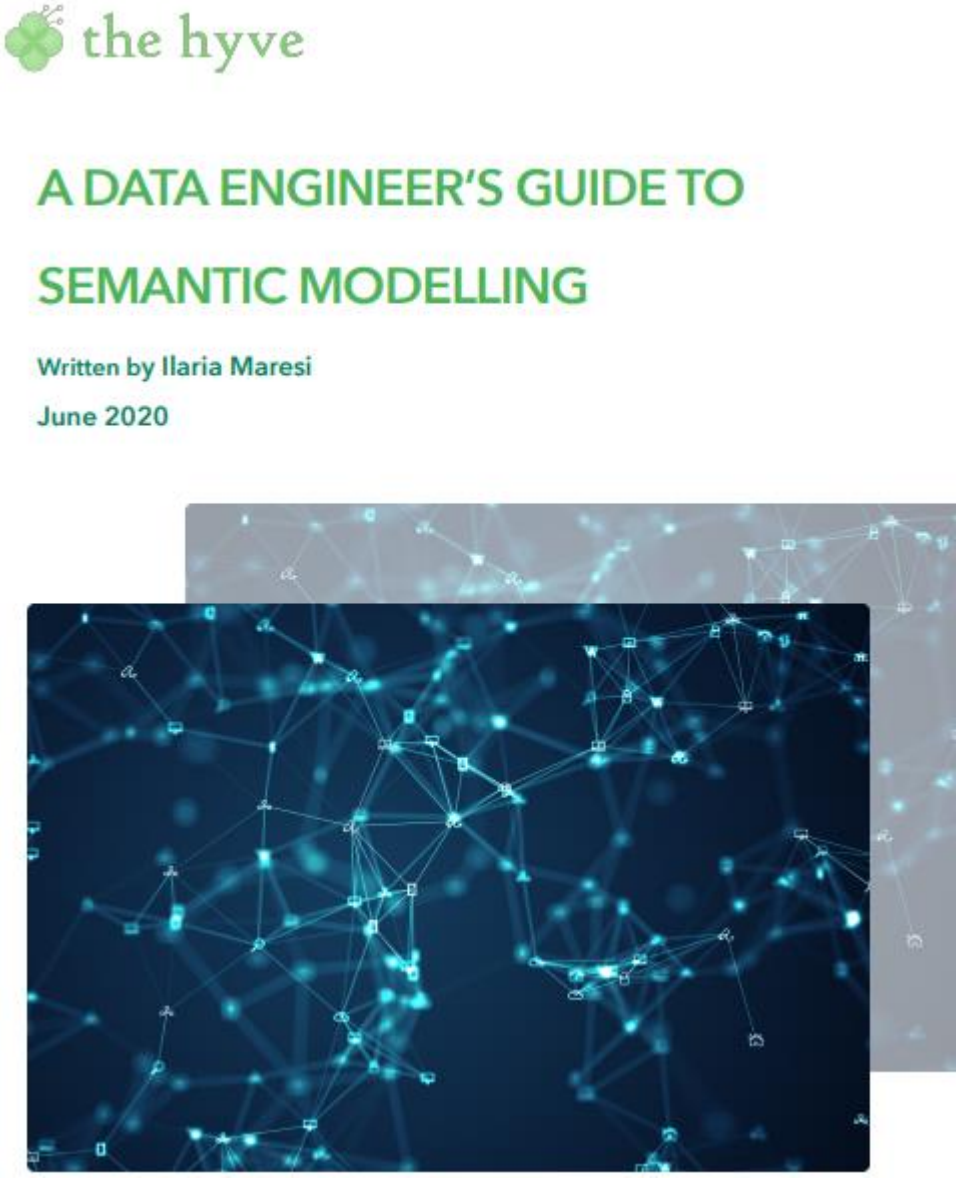
Ontology Development 101: A Guide to Creating Your First Ontology

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



What is an Ontology?

- **Dictionary**
  - **Terms** and their definitions
- **Taxonomy**
  - **Class** hierarchy
- **Thesaurus**
  - **Relationships** between terms
- Rules and Restrictions
  - Group membership, exclusions, types
  - Employ **reasoner** to ***infer*** values, relations

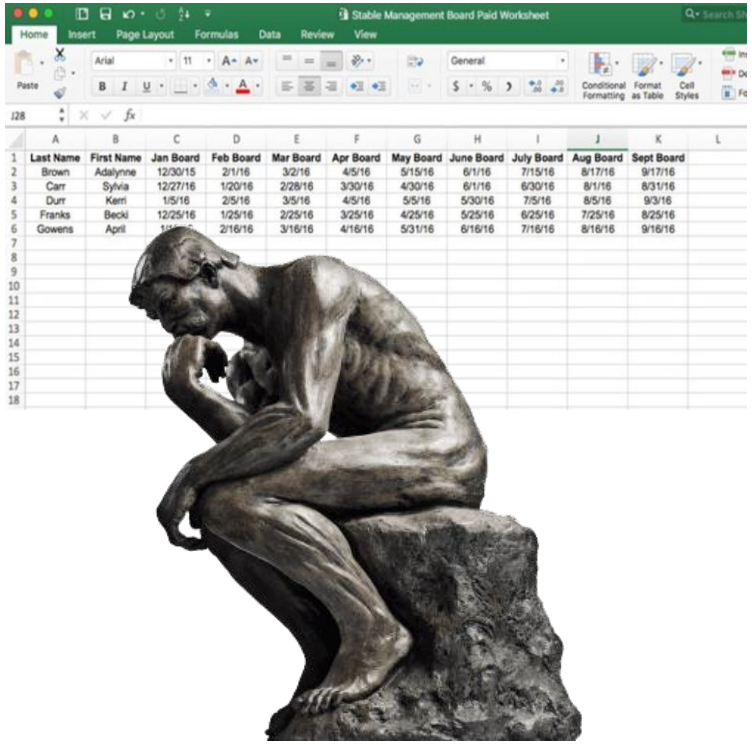


http://blog.thehyve.nl/news/ebook-semantic-model

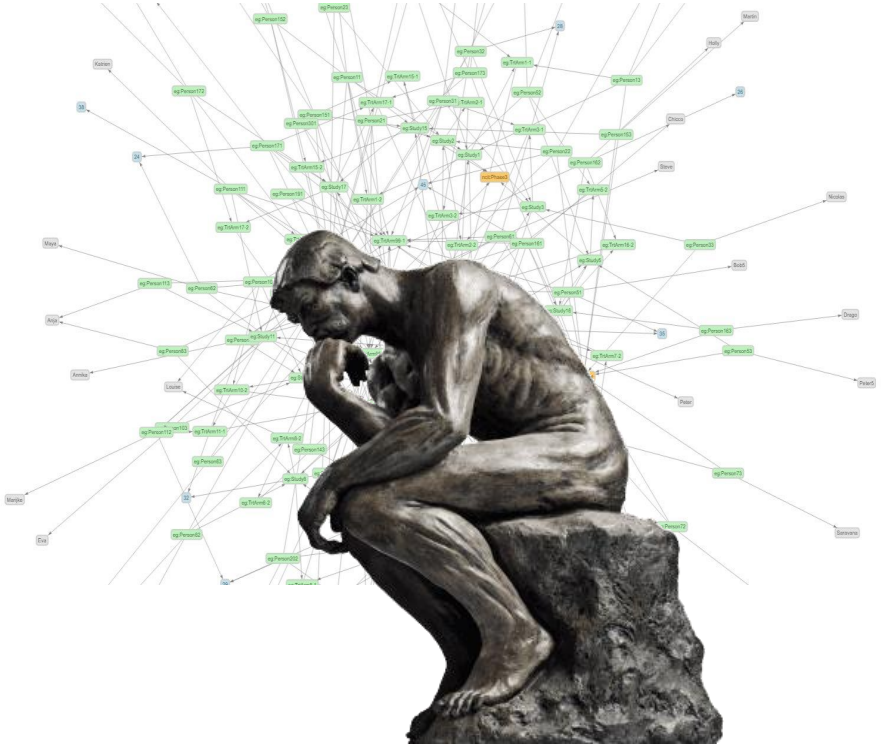
Changing How We Think About Data:

Data has Shape

Changing How We Think About Data



Last Name	First Name	Jan Board	Feb Board	Mar Board	Apr Board	May Board	June Board	July Board	Aug Board	Sept Board
Brown	Adalynne	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
Carr	Sylvia	12/27/16	1/20/16	2/28/16	3/30/16	4/30/16	6/1/16	6/30/16	8/1/16	8/31/16
Durr	Kerri	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
Franks	Becki	12/25/16	1/25/16	2/25/16	3/25/16	4/25/16	5/25/16	6/25/16	7/25/16	8/25/16
Gowens	April		2/16/16	3/16/16	4/16/16	5/31/16	6/16/16	7/16/16	8/16/16	9/16/16

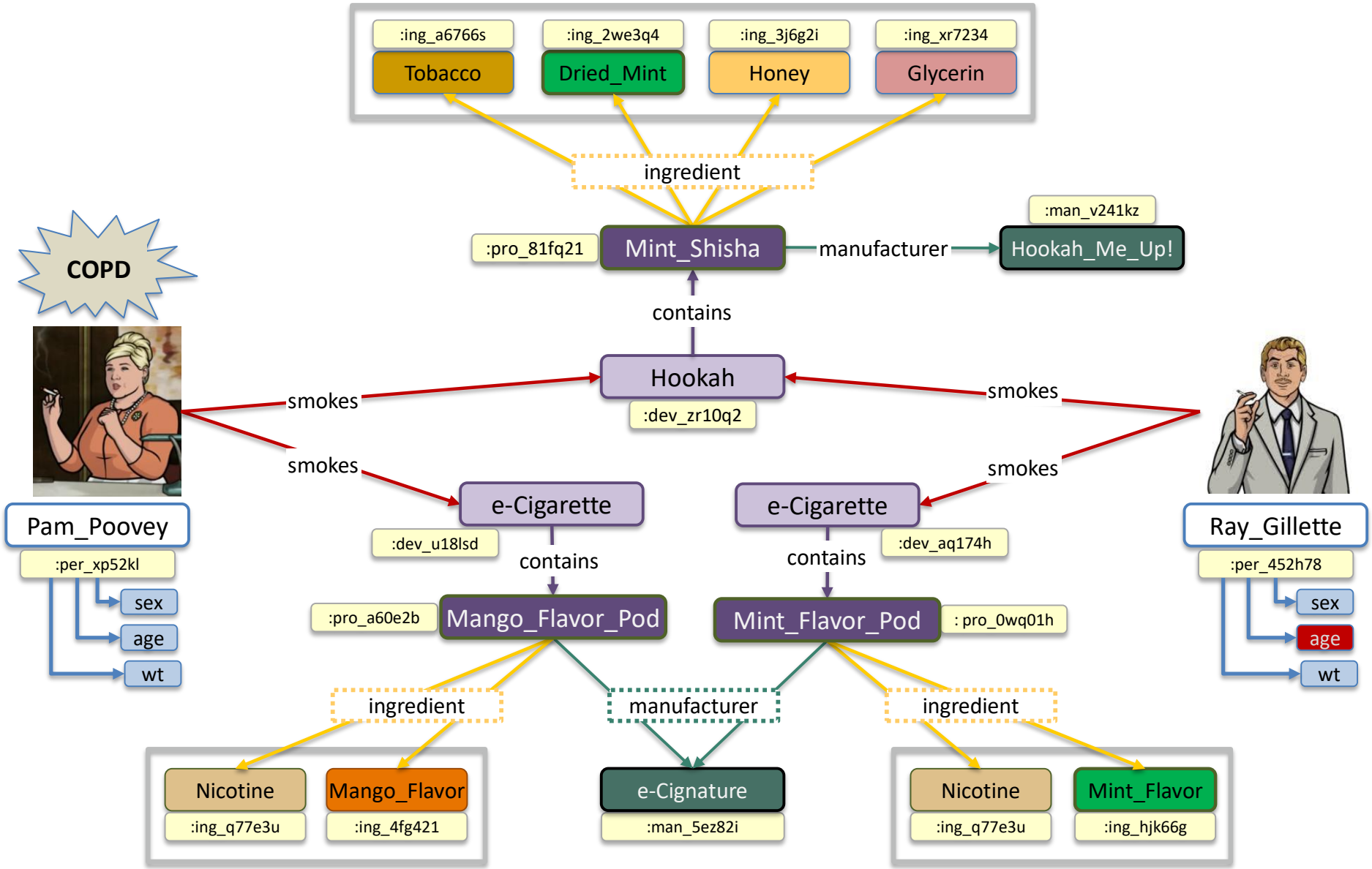


KG's: Easy Answers to Complex Questions

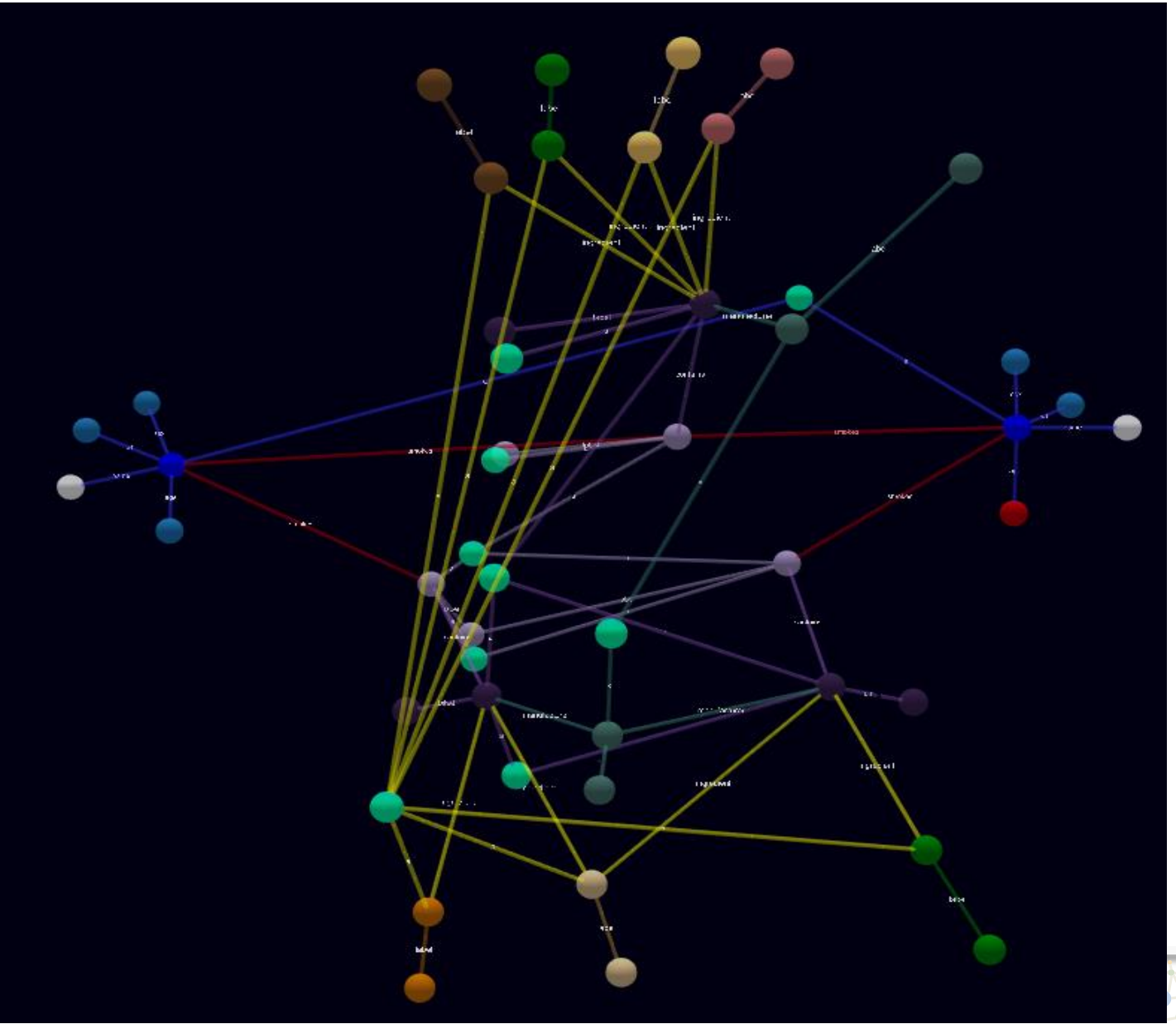
An example presented to FDA
*Division of Regulatory Science, Informatics (**DRSI**)*

- What *ingredient* may have contributed to Pam having COPD while Ray does not?

Pam's & Ray's Exposure



Exposure Data as a Graph



View the interactive visualization at:

https://bit.ly/PamAndRay

- Data error
- Demographics
- Device
- Entity Type
- Manufacturer
- Person
- Product

Ingredients

This seems complicated!

> "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."
>
> - attributed to Dan Brickley and Libby Miller

Data

```
:per_xp52kl :name          "Pam_Poovey"^^xsd:string ;
            a              :Person ;
            :age           "38"^^xsd:integer ;
            :wt            "205"^^xsd:integer ;
            :sex           :F ;
            :smokes        :dev_ul8lsd, :dev_zrl0q2  .

:per_452h78 :name          "Ray_Gillette"^^xsd:string ;
            a              :Person ;
            :age           "403"^^xsd:integer ;
            :wt            "165"^^xsd:integer ;
            :sex           :M ;
            :smokes        :dev_aql74h, :dev_zrl0q2 .

:dev_ul8lsd skos:prefLabel "e-Cigarette"^^xsd:string ;
            a              :eCigarette;
            a              :Device ;
            :contains      :pro_a60e2b .

:dev_zrl0q2 skos:prefLabel "Hookah"^^xsd:string ;
            a              :HookahPipe ;
            a              :Device ;
            :contains      :pro_81fq21 .

:dev_aql74h skos:prefLabel "e-Cigarette"^^xsd:string ;
            a              :eCigarette ;
            a              :Device ;
            :contains      :pro_0wq01h .

:pro_81fq21 skos:prefLabel "Mint_Shisha"^^xsd:string ;
            a              :Product ;
            a              :TobaccoMix ;
            :manufacturer  :man_v241kz ;
            :ingredient    :ing_a6766s, :ing_2we3q4, :ing_3j6g2i, :ing_xr7234 .

:pro_a60e2b skos:prefLabel "Mango_Flavor_Pod"^^xsd:string ;
            a              :Product ;
            a              :FlavorPod ;
```

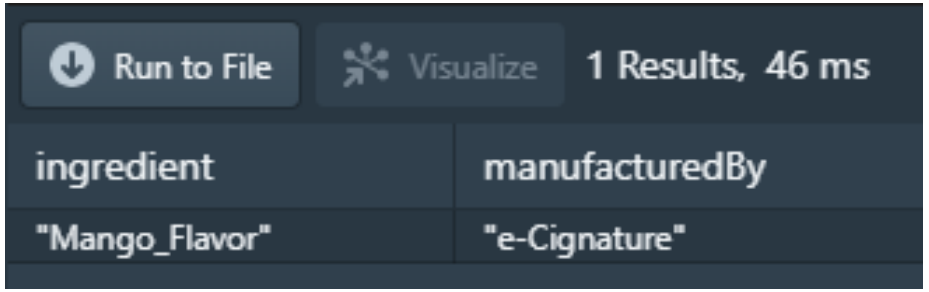
View the data file at:

https://bit.ly/PamAndRayTTL

Pam's Unique Exposure Ingredient?

```
PREFIX  :     <http://example.org/Eg#>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>

SELECT ?ingredient ?manufacturedBy
WHERE{
  # Pam : Person 1 Exposure
  ?person1     :smokes        ?smokeDevice ;
               :name          ?personName .
  ?smokeDevice :contains      ?mixture .
  ?mixture     :ingredient    ?ingred ;
               :manufacturer  ?man .
  ?ingred      skos:prefLabel ?ingredient .
  ?man         skos:prefLabel ?manufacturedBy .

  FILTER( regex(?personName , "Pam"))

  # Ray : Person 2 Exposure
  OPTIONAL{
    ?person2      :smokes        ?smokeDevice2 ;
                  :name          ?personName2 .
    ?smokeDevice2 :contains      ?mixture2 .
    ?mixture2     :ingredient    ?ingred2 .
    ?ingred2      skos:prefLabel ?ingredient.

    FILTER( regex(?personName2 , "Ray"))
    # Ingredients that are common to both will be BOUND to ingredient2
    FILTER (?ingred = ?ingred2)
  }
    # Keep only those that are not in the ingredient 2 set, i.e. not bound
    #   as an ingredient for person2
    FILTER(! BOUND(?ingred2))
}
```

PamUniqueIngred.rq



Ontology

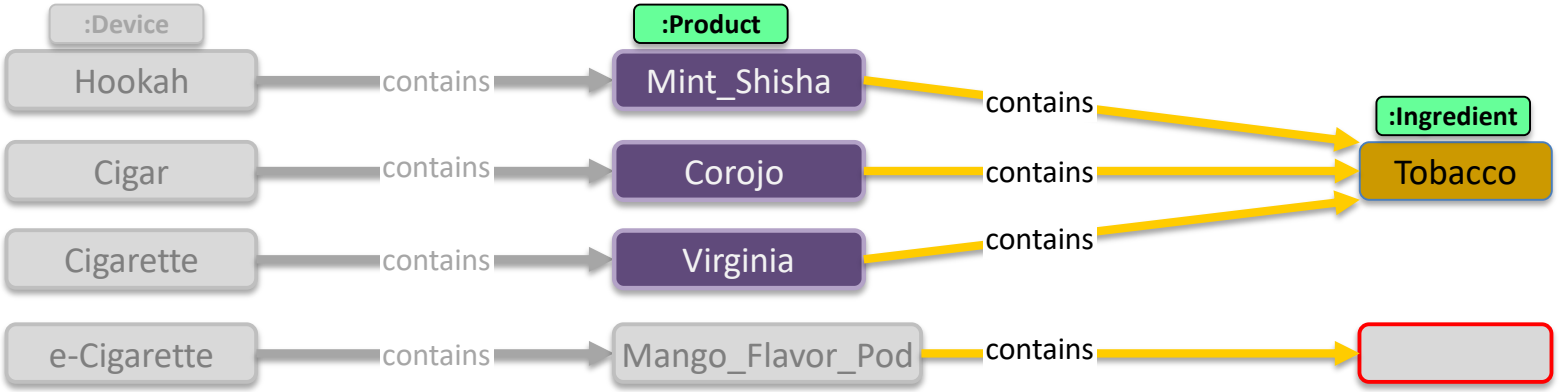
- **What** is a **TobaccoProduct**?

Ontology Definition

A thing **?** that is a **:Product**

The **:Product** —contains→ an **:Ingredient**

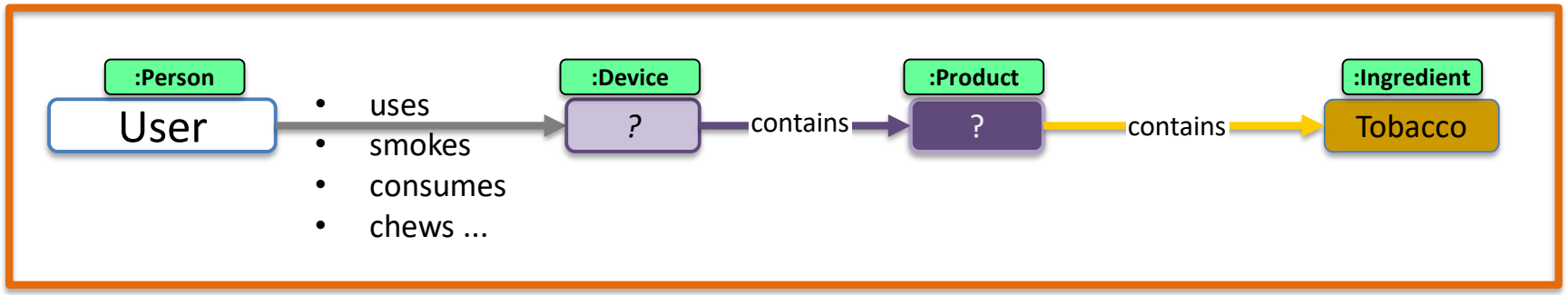
The **:Ingredient** is **Tobacco**



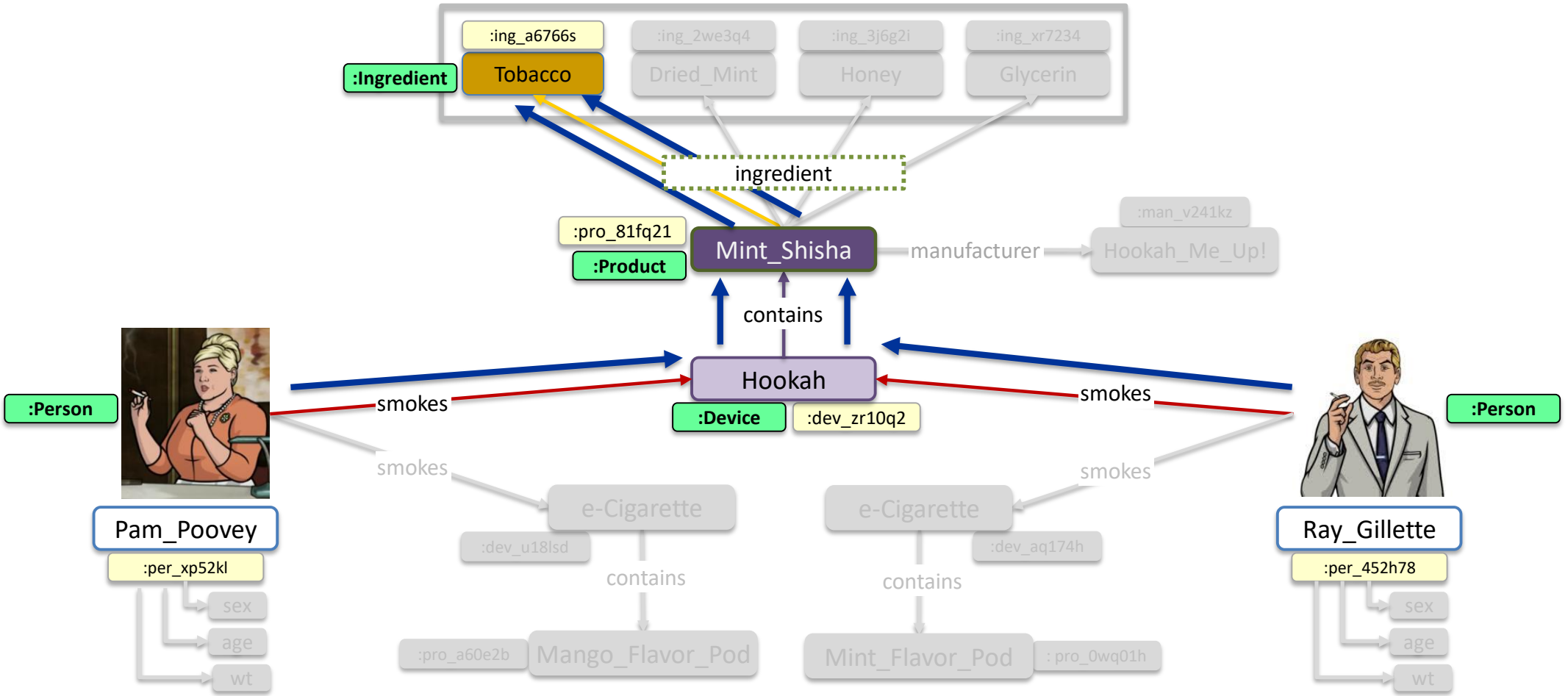
Ontology

- Who is a **TobaccoUser** ?

Ontology Definition



Tobacco Smoker: The path between Person and Ingredient



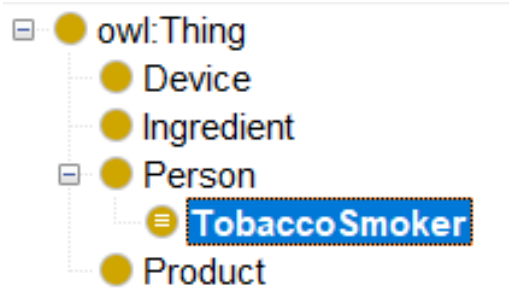
Query: Tobacco Smokers

TobaccoUsers.rq

```
PREFIX :        <http://www.example.org/eg#>
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>

SELECT ?name ?ingredient

WHERE{
 ?pers      :name          ?name ;
            :smokes        ?device .
  ?device   :contains      ?product .
  ?product  :ingredient    ?ingred .
  ?ingred   skos:prefLabel ?ingredient .
  FILTER (REGEX(?ingredient, "Tobacco"))
}
```

name	ingredient
"Pam_Poovey"	"Tobacco"
"Ray_Gillette"	"Tobacco"

Tobacco Smoker: Ontology Definition & Query




```
Description: TobaccoSmoker
Equivalent To
    Person
      and (smokes some
        (Device
         and (contains some
           (Product
            and (ingredient some
              (Ingredient
               and (skos:prefLabel value "Tobacco")))))))
```

TobaccoSmoker-Infer.rq

```
PREFIX :        <http://www.example.org/eg#>
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>

SELECT ?tobaccoSmoker
WHERE{
    ?person a       :TobaccoSmoker ;
            :name ?tobaccoSmoker .
}
```

tobaccoSmoker
"Ray_Gillette"
"Pam_Poovey"

When Data has Shape,

Validation has Shape

SHApes Constraint Language (SHACL)

“Person Shape”
(Validation Constraints)

Person Data

Validating Ray's Demographics

Ray Gillette

Demographics

sex: M

age: 43

wt: 165

Data Entry

M

403

165

SHACL

Person Shape

M/F/U ✓

SEX

> 0
< 120

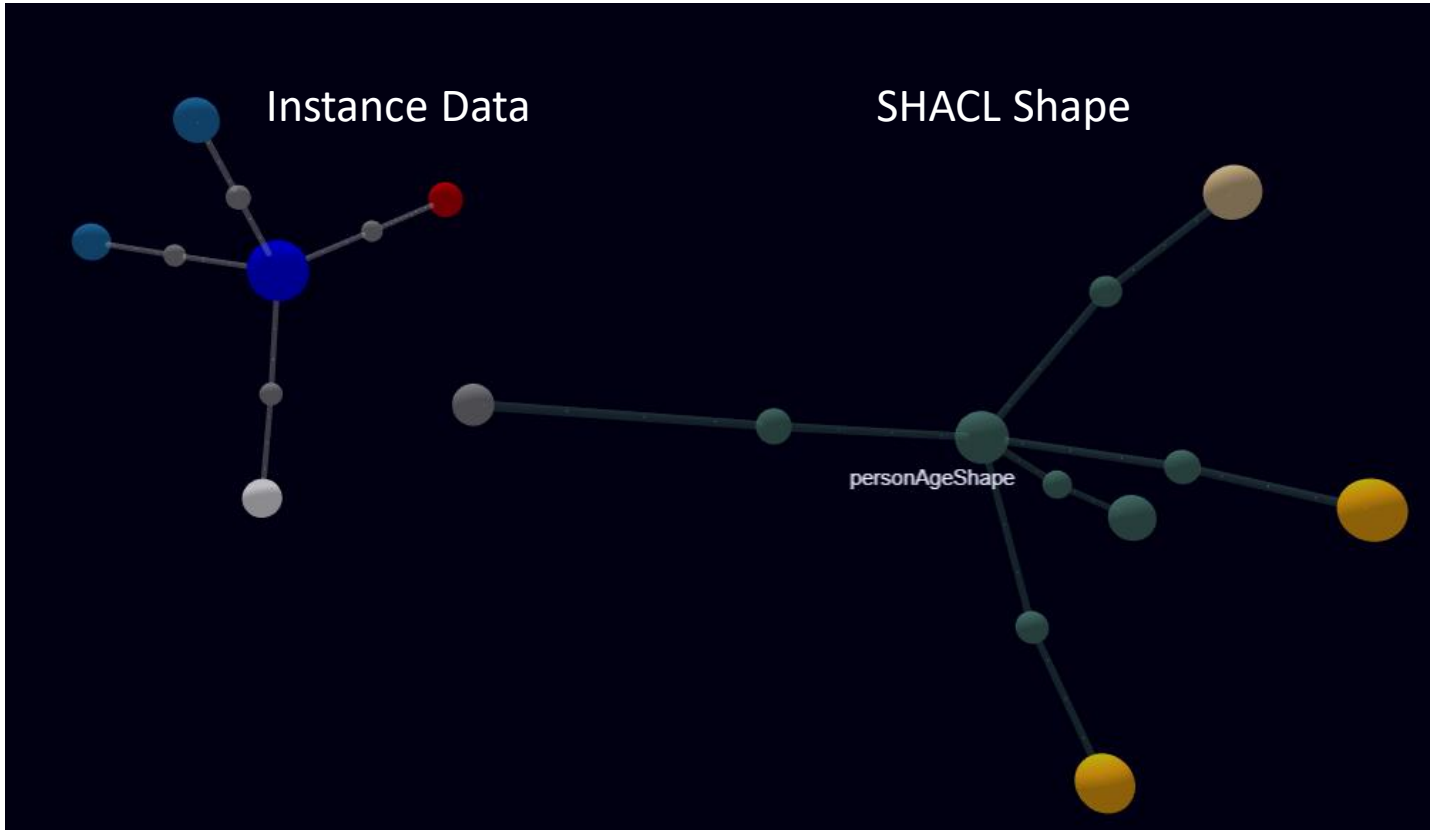
AGE

Violation!

> 0
< 800 ✓

WEIGHT

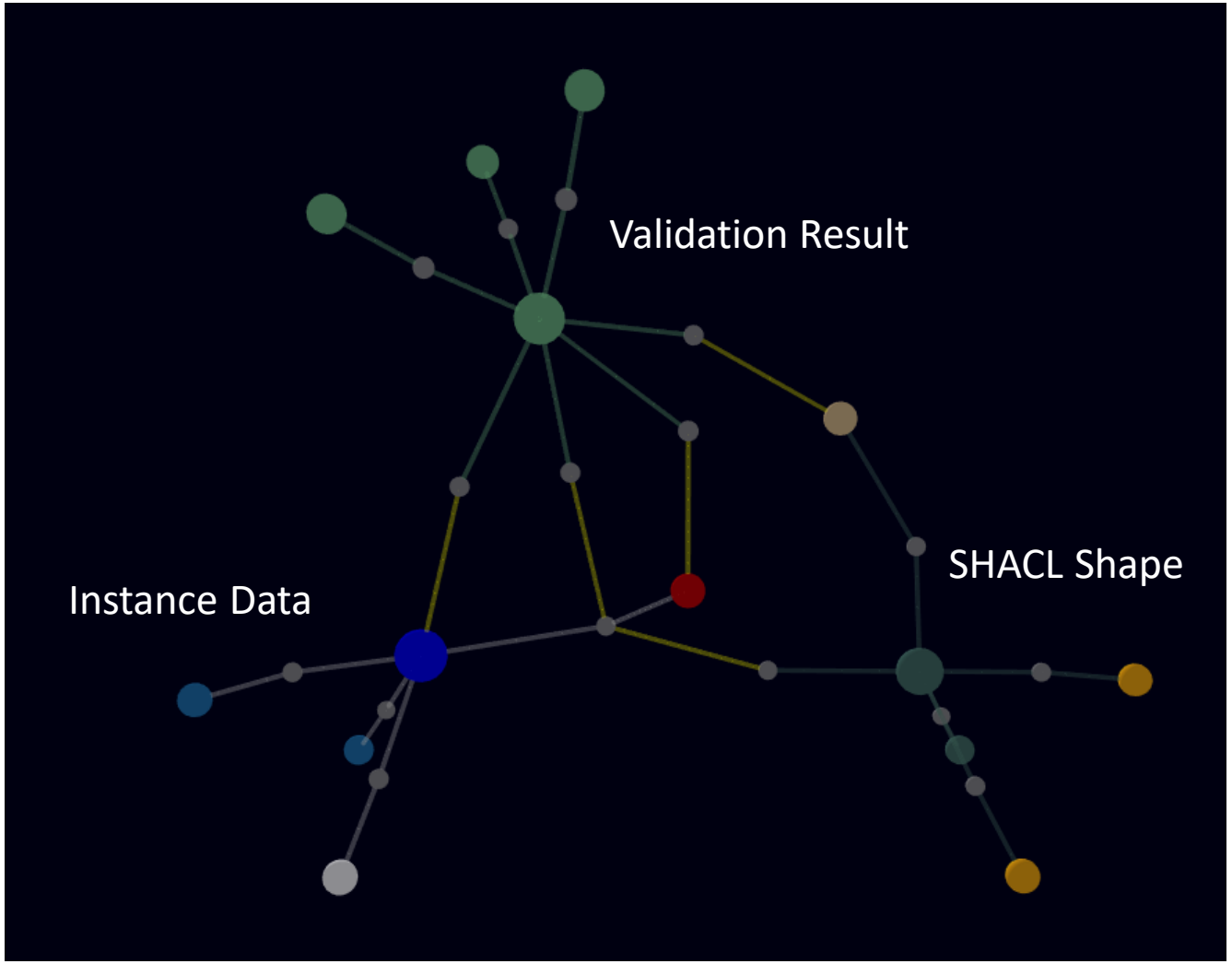
Validating Age with SHACL



View the interactive visualization at:

https://bit.ly/RayDemogAndSHACL

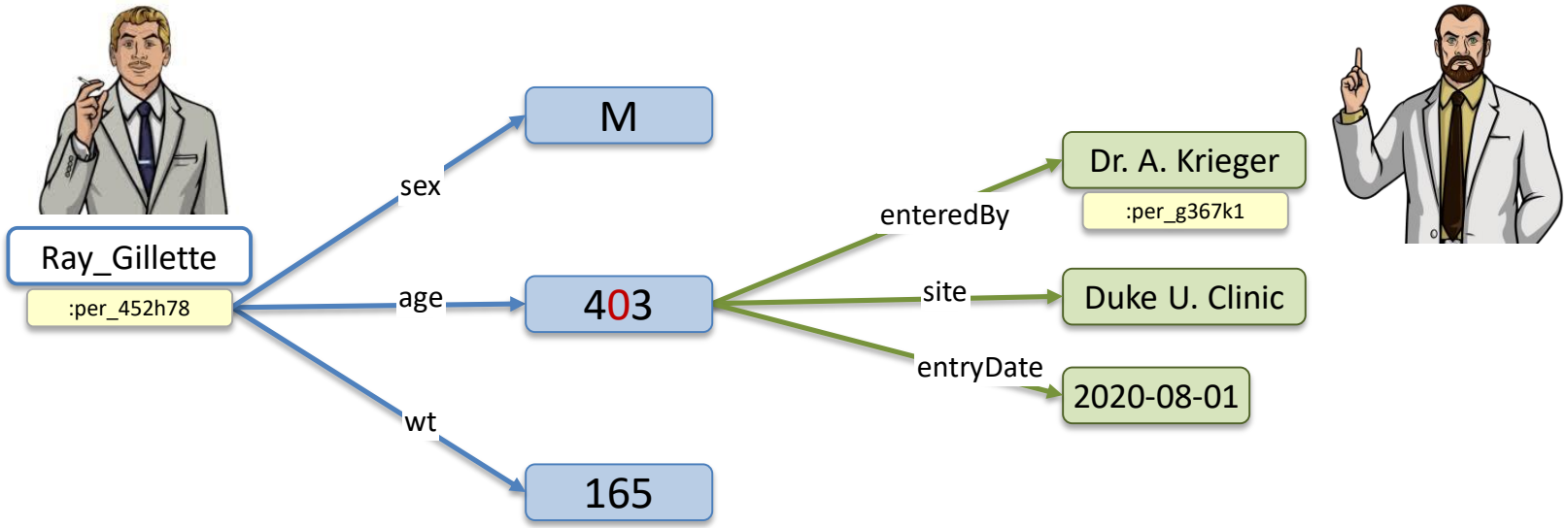
Demographics, Constraints, Report



View the interactive visualization at:

https://bit.ly/RaySHACLResult

Metadata is also Part of the Graph



Strategies for Implementation

The Roofshot / Moonshot Manifesto

Moonshot

Invent and apply state-of-the-art

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

Roofshot

Incremental impacts in production

Examples

1. Unique Identifiers
2. Validation Rules in SHACL
3. Open Ontology Development for your domain

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

The Stairway to the Stars Manifesto

Solve a Simple, “Every-day Problem” with a Knowledge Graph

What are the interrelationships in code, inputs, outputs in my project?

- Dynamic R Markdown documentation based on query of a Project Documentation Knowledge Graph.
- Interactive visualization and query of dependencies.
- Process management, automation, & efficiency

Data Conversion to RDF

Folder: `C:\_sandbox\OAA\Autotrial\scripts\r\rdfConversion`

Description

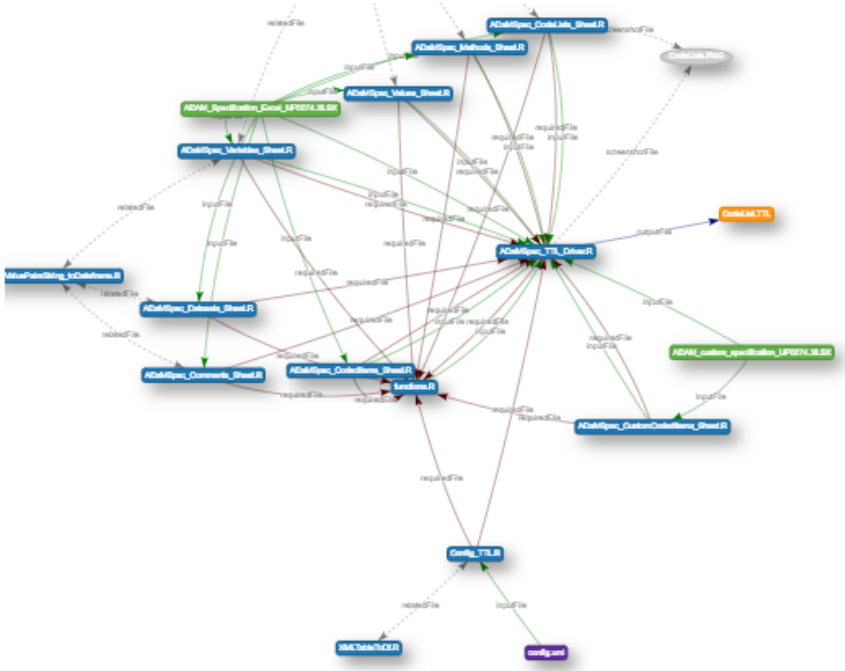
Create RDF from the various files needed to create the AutoTrial Knowledge Graph. The resulting TTL files in the /data/rdf folder are uploaded to the AutoTrial graph using a batch load process.

▸ *Show Files*

▾ *Show Relationships*

File Relationships

Select by id



EXAMPLE

Strategy

- Target known problems in your organization develop scalable projects with **demonstrable R**eturn **O**n **I**nvestment.
- **Dive in**: Model what you need and reuse existing ontologies and terms, but don't get stuck on this point.
- Add **instance data** early. **Query**. **Adapt** models as you go.
- Write **validation** early. Use SHACL (or SHEX).
- Be **consistent** across projects and revise across projects as you go.
- **Plan for growth**.

Credit: Adapted from Kurt Cagle

Who will Lead the Transformation?

Favor *status quo*	Favor Change
Legacy Vendors	Graph Vendors
Traditional Consultants	Data-centric Consultants
Legacy Corporate IT	Enlightened IT
Standards Organizations	Standards Organizations (future)
Regulatory Agencies	Regulatory Agencies (future)

Research

- Drug Discovery
- Genomics
- Key Opinion Leader ID

Analytics

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

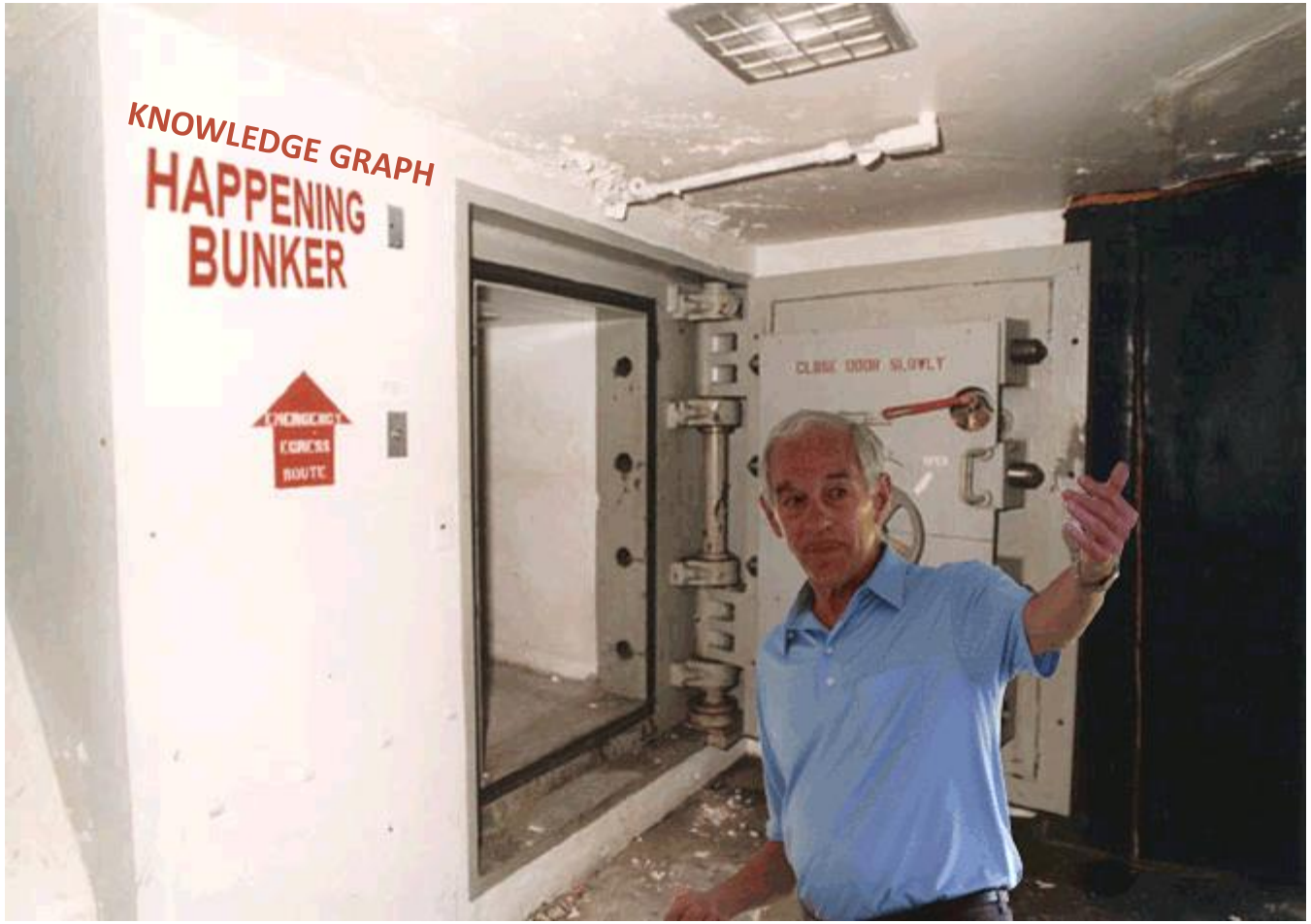
Challenges Remain

- Property Graph vs. RDF vs. Hybrid .. RDF* ?
- Ontologies
  - Often do not fit your view of the world
  - Can be poorly designed (inferencing problems)
- Inconsistent use of Standards.
  - Not everyone follows W3C standards
  - How do I form an identifier? (IRI/URI)
- Lack of INDUSTRY Expertise
  - Hire consultants with a view toward Knowledge Transfer to your staff.

Credit: Adapted from Kurt Cagle

Knowledge Graphs are the (Data) Shape of the Future!

It ***is*** happening...



...let's go!

Additional Reading

General / Introductory

- A Data Engineer's Guide to Semantic Modeling - Maresi . *Free e-book download.*
- The Data Centric Revolution - McComb

Technical

- Semantic Web for the Working Ontologist (3rd Ed, 2020) – Hendler, Gandon, Allemang
- Demystifying OWL for the Enterprise – Uschold, Ding, Groth
- Validating RDF Data – Gayo, Prud'hommeaux, Boneva . *Comparison of SHEX and SHACL.*
- Learning SPARQL - DuCharme . *Learn RDF by querying the data.*
- 3D-force-graph *Network graph visualizations in this presentation.*

https://bit.ly/PamAndRay
https://bit.ly/RayDemogAndSHACL
https://bit.ly/RaySHACLResult

Pam and Ray Exposure Data
https://bit.ly/PamAndRayTTL

Thank you!!

Contact:

LinkedDataTim@gmail.com

https://www.linkedin.com/in/**timpwilliams**

@NovasTaylor