



FDA/PhUSE 
Computational Science Symposium

March 18th – 19th 2013

Silver Spring Civic Center | Silver Spring, Maryland, USA



Semantic Technologies

Frederik Malfait, IMOS Consulting

Information Architect

Data Standards, PD Biometrics, Hoffmann-La Roche

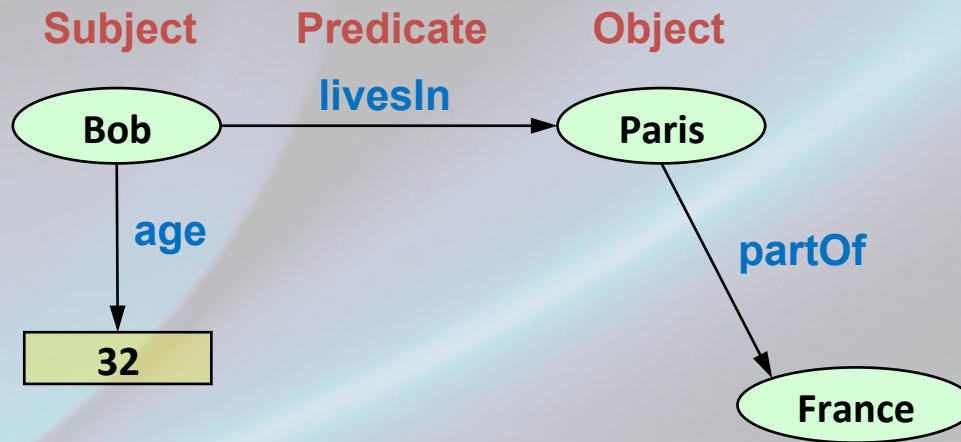


Part I

Introduction to Semantic Technology

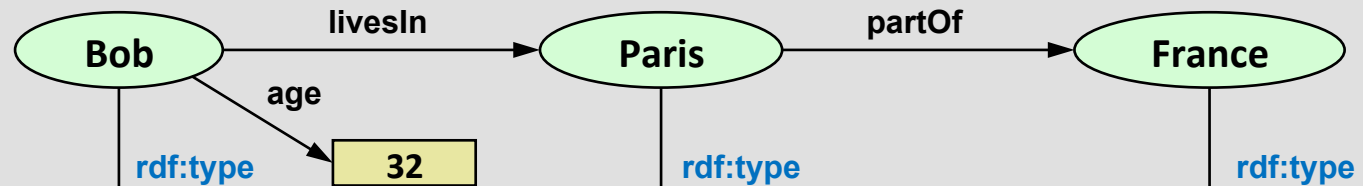
Resource Description Framework RDF

Directed graph of <Subject Predicate Object> triples

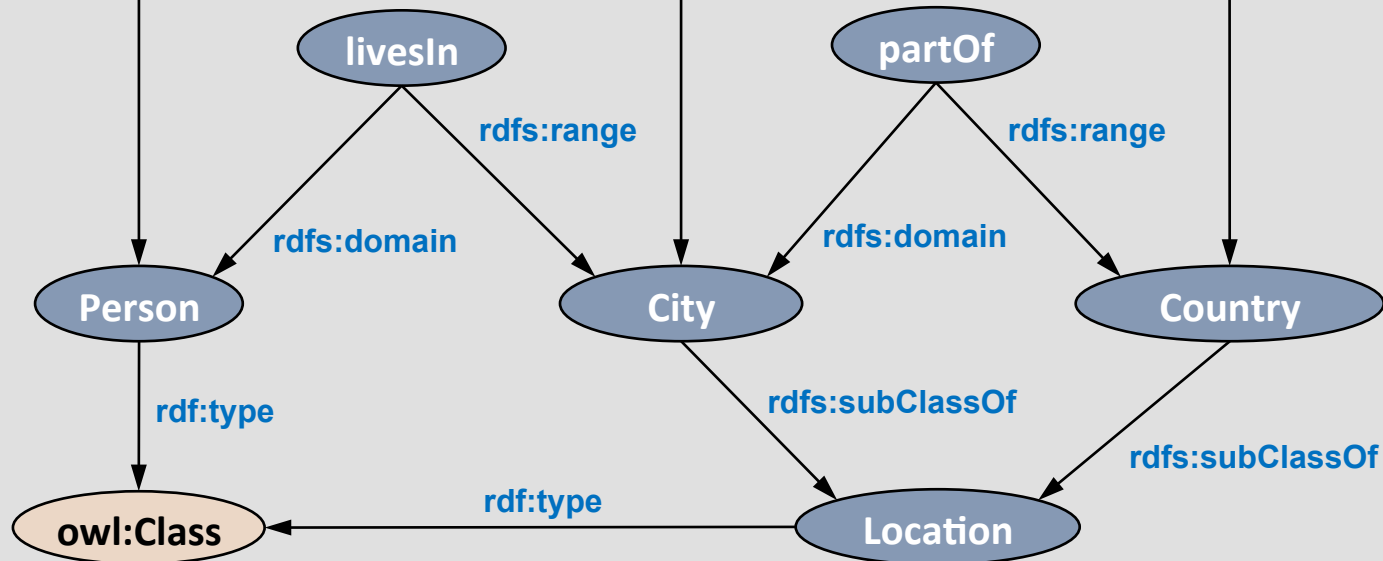


Ontologies

Fact Base



Ontology



Everything is a Triple



Uniform Resource Identifier URI

- In RDF everything is a triple built from resources and literals
 - <Subject Predicate Object>
 - <Subject Predicate Literal>
- Every RDF resource has a Uniform Resource Identifier (URI)
 - NCI Thesaurus, term Mild Adverse Event
<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C41338>
- Namespaces provide a convenient way to group related resources together
 - Define the namespace prefix nci
[xmlns:nci="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#"](http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#)
 - To obtain the Qualified Name (qname)
[nci:C41338](#)

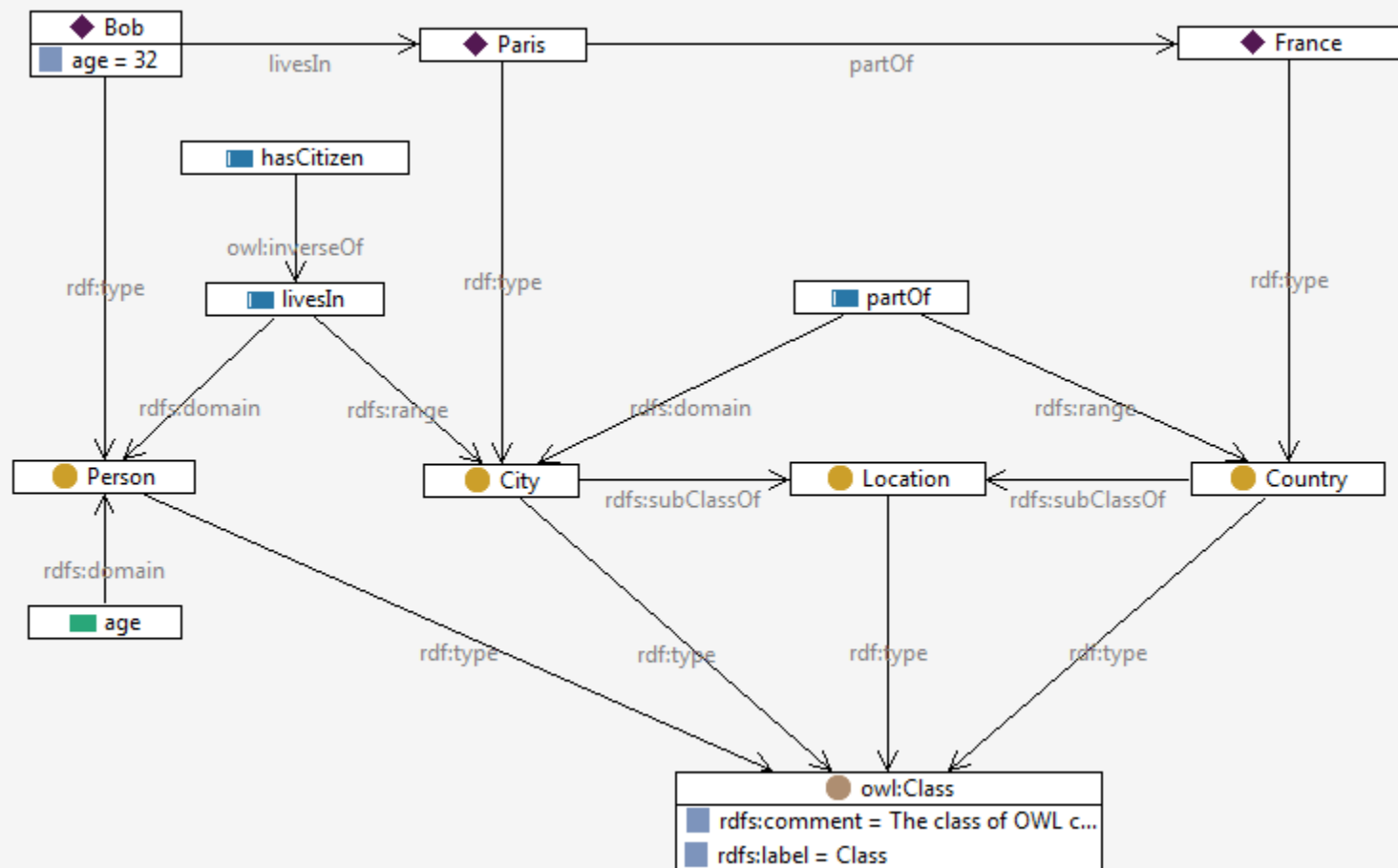


Inference

- RDFS and OWL define an extensive vocabulary for building ontologies
 - Example: `<hasCitizen owl:inverseOf livesIn>`
- Meaning is defined by the way new triples are inferred from existing triples
 - Stated Triples
 - `<Bob livesIn Paris>`
 - `<hasCitizen owl:inverseOf livesIn>`
 - Inferred Triple
 - `<Paris hasCitizen Bob>`
- Inference engines provide reasoning capabilities for formal ontologies
- Formal ontologies have a precise mathematical foundation based on first-order logic, set theory, and model theory

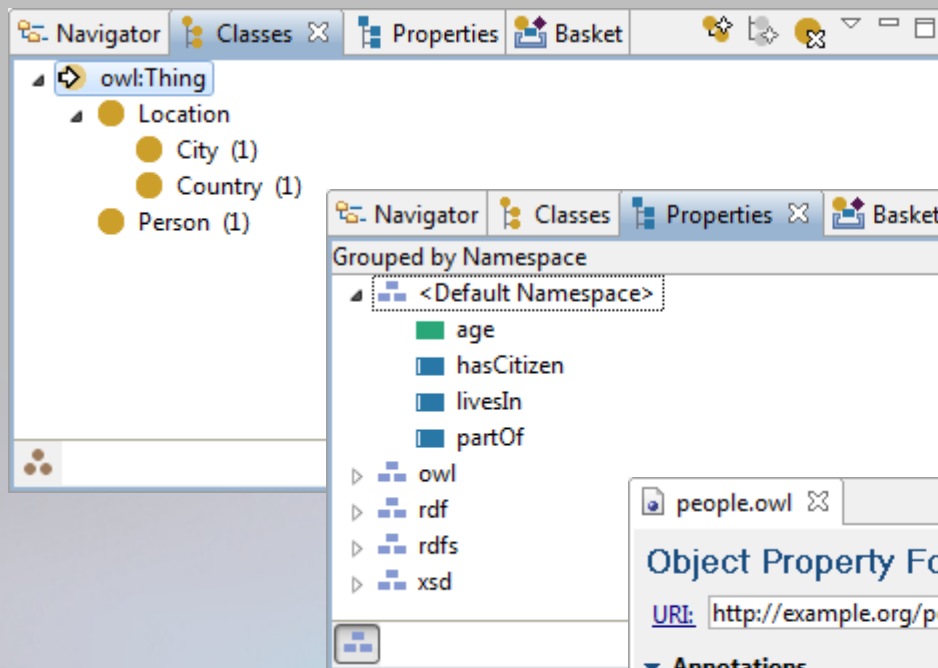
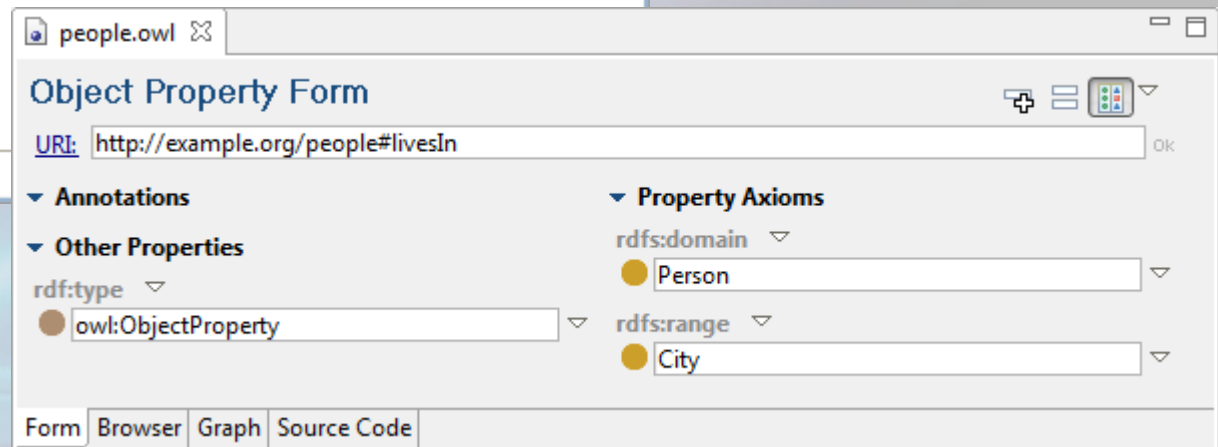
RDF Representations

Graph Based



RDF Representations

Ontology Editors





RDF Representations Serializations



RDF/XML

```
<rdf:Description rdf:about="#Bob">
  <rdf:type rdf:resource="#Person"/>
  <age rdf:datatype="http://www.w3.org/2001/XMLSchema#positiveInteger">32</age>
  <livesIn rdf:resource="#Paris"/>
</rdf:Description>
```

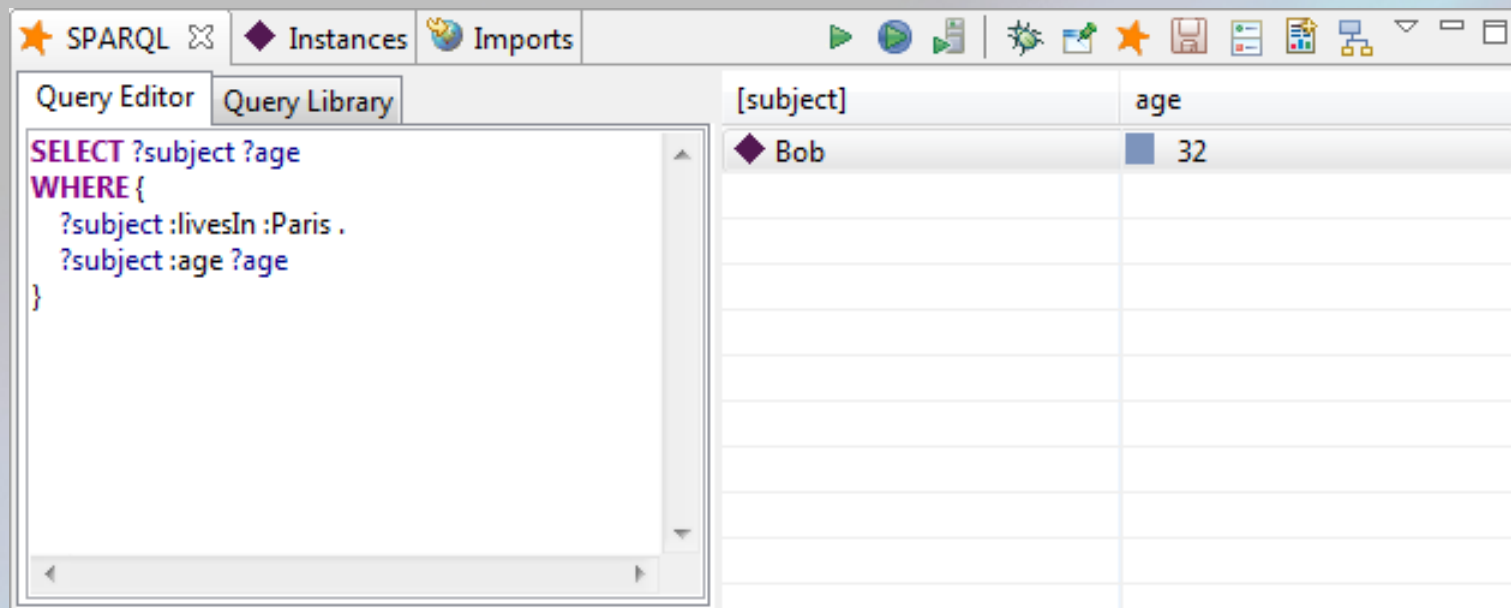
Turtle

```
:Bob a      :Person ;
      :age   "32"^^xsd:positiveInteger ;
      :livesIn :Paris .
```

N-Triple

```
<http://gdsr.roche.com/people#Bob> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://gdsr.roche.com/people#Person> .
<http://gdsr.roche.com/people#Bob> <http://gdsr.roche.com/people#age> "32"^^<http://www.w3.org/2001/XMLSchema#positiveInteger> .
<http://gdsr.roche.com/people#Bob> <http://gdsr.roche.com/people#livesIn> <http://gdsr.roche.com/people#Paris> .
```

RDF Query Language SPARQL



The screenshot displays a SPARQL query editor interface. The top toolbar includes icons for SPARQL, Instances, Imports, and various query execution and management functions. The interface is divided into two main sections: a Query Editor on the left and a Results table on the right.

Query Editor: The query text is as follows:

```
SELECT ?subject ?age
WHERE {
  ?subject :livesIn :Paris .
  ?subject :age ?age
}
```

Results Table: The table has two columns: [subject] and age. It contains one row of results:

[subject]	age
Bob	32



W3C Semantic Standards

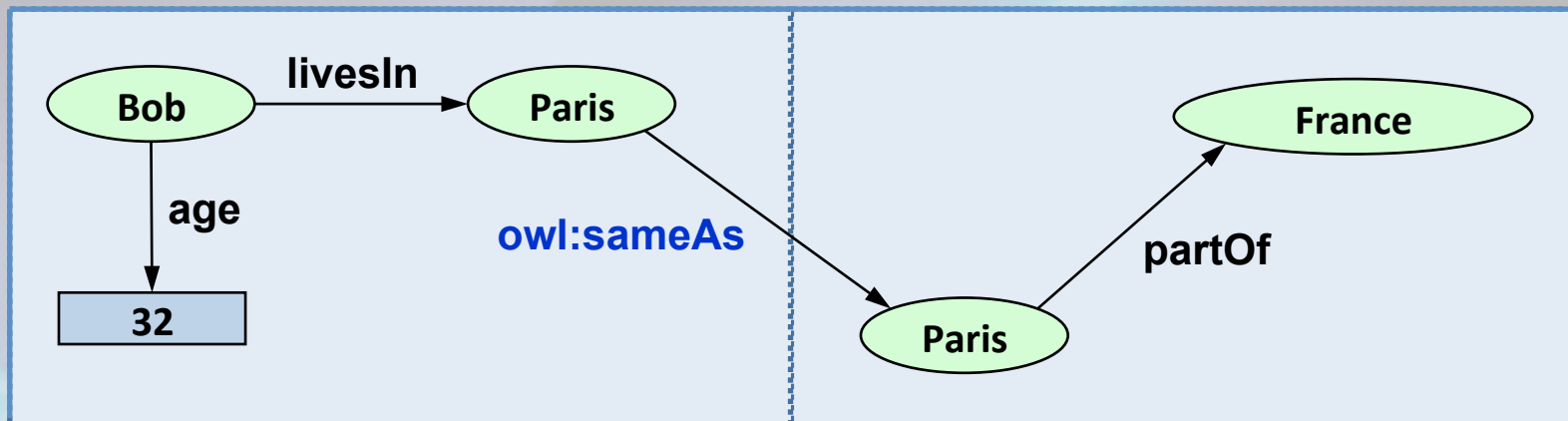
- Resource Description Framework (RDF)
 - RDF defines how to express information as a directed graph of resources
 - Every resource has a URI and is part of a namespace
- RDF Schema (RDFS) and Web Ontology Language (OWL)
 - A set of standard predicates to build ontologies (vocabularies, schemas)
 - Inference capabilities
- SPARQL Protocol and RDF Query Language (SPARQL)
 - Language to query a RDF data sets
- Simple Knowledge Organization System (SKOS)
 - Small RDF vocabulary for building concept models (thesauri, etc.)

Federation and Linked Data

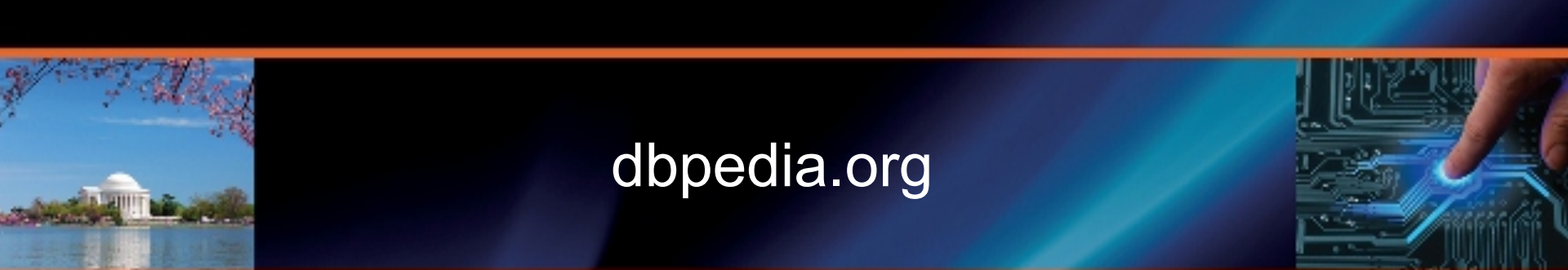
- Semantic models in RDF format are easy to federate
- Federation of Data = union of triples from both graphs
- Linked Data = referencing between graphs
- Use [owl:sameAs](#) to specify that two resources are equal

<http://example.org/people#>

<http://example.org/geo/#>

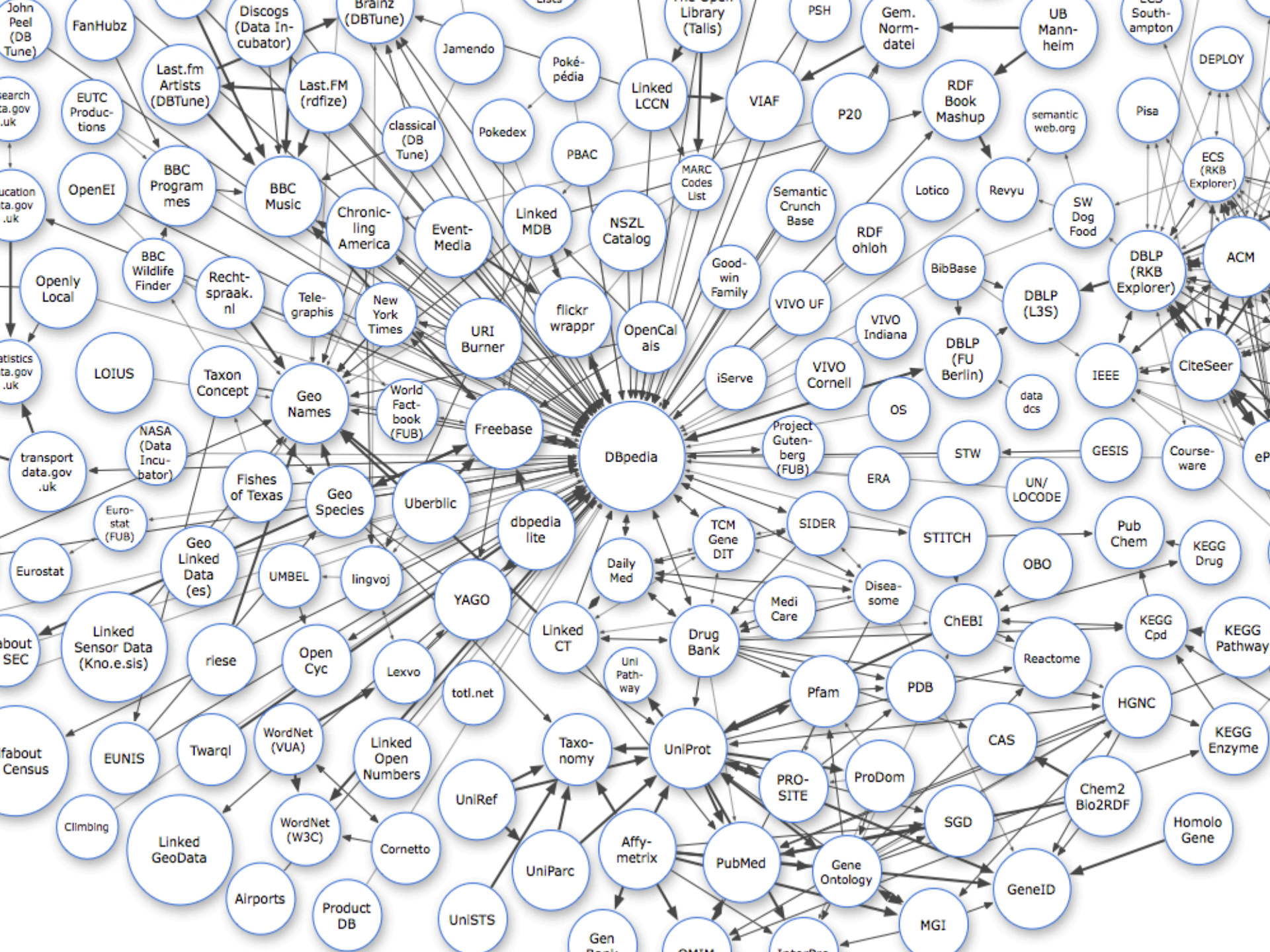


Federated Graph



dbpedia.org

In addition, we provide localized versions of DBpedia in 111 languages. All these versions together describe 20.8 million things, out of which 10.5 million overlap (are interlinked) with concepts from the English DBpedia. The full DBpedia data set features labels and abstracts for 10.3 million unique things in up to 111 different languages; 8.0 million links to images and 24.4 million HTML links to external web pages; 27.2 million data links into external RDF data sets, 55.8 million links to Wikipedia categories, and 8.2 million YAGO categories. The dataset consists of 1.89 billion pieces of information (RDF triples) out of which 400 million were extracted from the English edition of Wikipedia, 1.46 billion were extracted from other language editions, and about 27 million are data links to external RDF data sets. The [Datasets](#) page provides more information about the overall structure of the dataset. [Dataset Statistics](#) provides detailed statistics about 22 of the 111 localized versions.

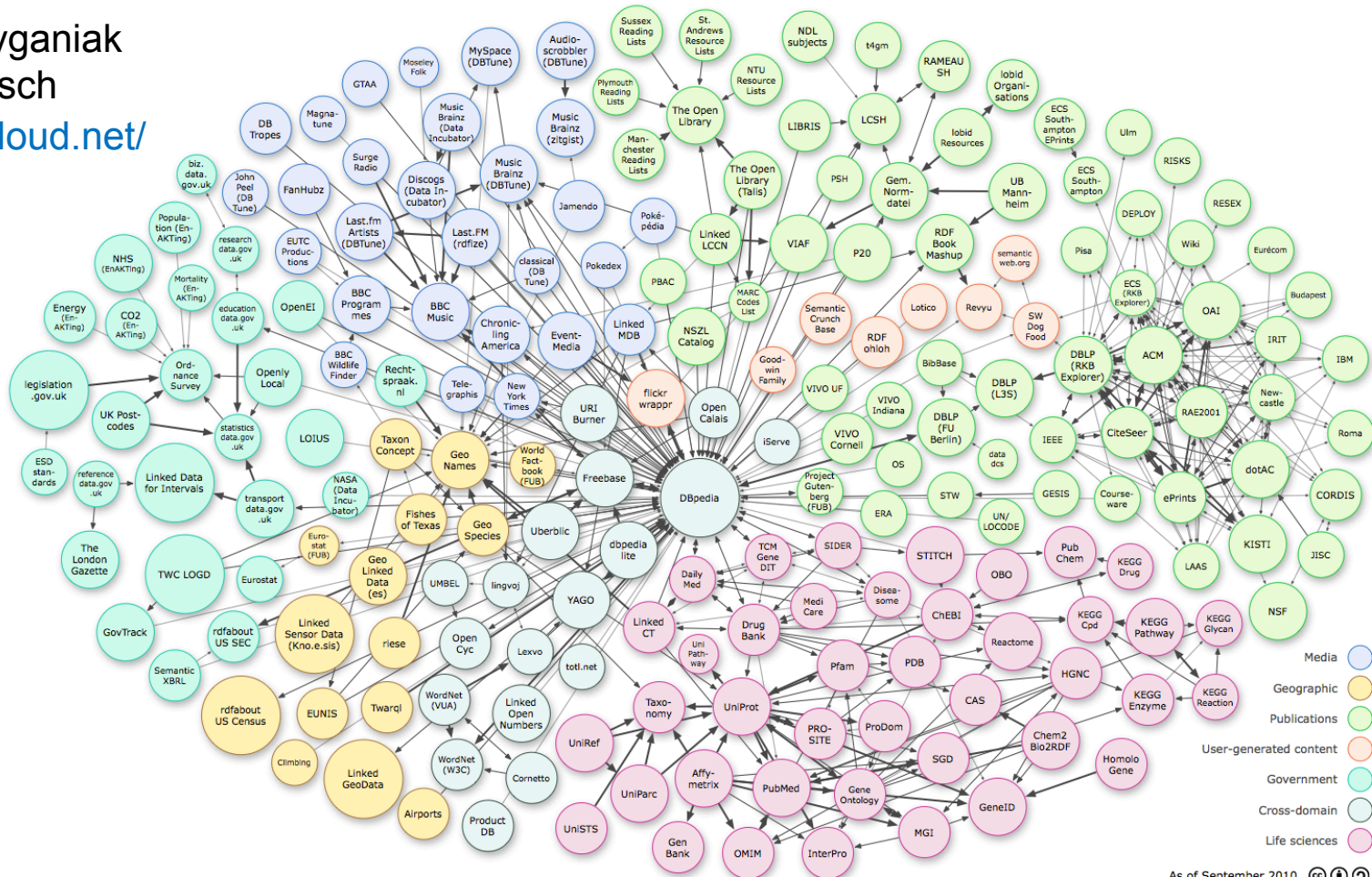


Linked Open Data (LOD) Cloud

Richard Cyganiak

Anja Jentzsch

<http://lod-cloud.net/>



As of September 2010



Semantic Technology Application Areas

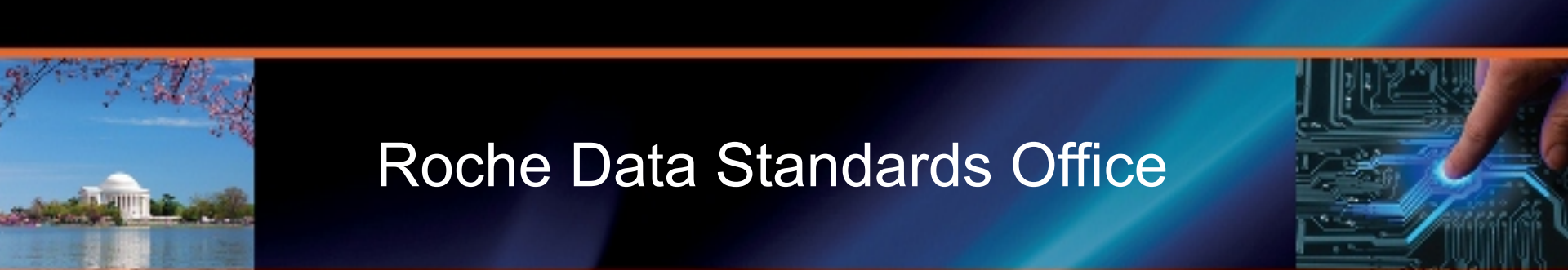


- Semantic Web
 - Move the web beyond a web of documents, add semantics
 - Large scale web data integration
- Knowledge Management
 - Vocabularies, taxonomies, thesauri, nomenclature, dictionaries e.g. SNOMED CT, NCI Thesaurus, HGNC, etc.
 - Domain ontologies, i.e. formal knowledge representation and reasoning capabilities for a certain domain of interest
- Roche Data Standards Office
 - Data standards and metadata driven workflow
 - Metadata registries, e.g. based on ISO 11179
 - Data integration and master data management



Part II

Semantic Models for CDISC Standards and Metadata Management at Hoffmann-La Roche



Roche Data Standards Office

- Design and manage end-to-end clinical trial data standards
 - from protocol to submission
 - with supporting business processes
- Drive adoption of current and future industry standards
- Provide lifecycle management and governance for all data standards
- Contribute to external standards bodies
 - CDISC
 - TransCelerate
 - Phuse/FDA



Global Data Standards Repository Objectives

- Single point of access for end-to-end data standards
 - Protocol and Study Design
 - Data Collection, Data Tabulation, Data Analysis
 - Regulatory Submission
- Enable consistent metadata management
 - Operational and conceptual metadata
 - Versioning and life cycle management
- Follow industry standards when available and fit for purpose
 - CDISC standards
 - ISO 11179 Metadata Registry (MDR) standards
 - W3C semantic standards (RDF, OWL, SKOS)

Global Data Standards Repository

Semantic Information Model

Production

Partial / Future

Sponsor Extensions

PRM

CDASH

SDTM

ADaM

define

Biomedical Concept Model [partial]

ISO 11179 MDR Schema [subset]

BRIDG and ISO 21090

NCI Thesaurus

RDF

OWL

SKOS

**CDISC
Standards**

**Metadata
Management**

**Knowledge
Management**

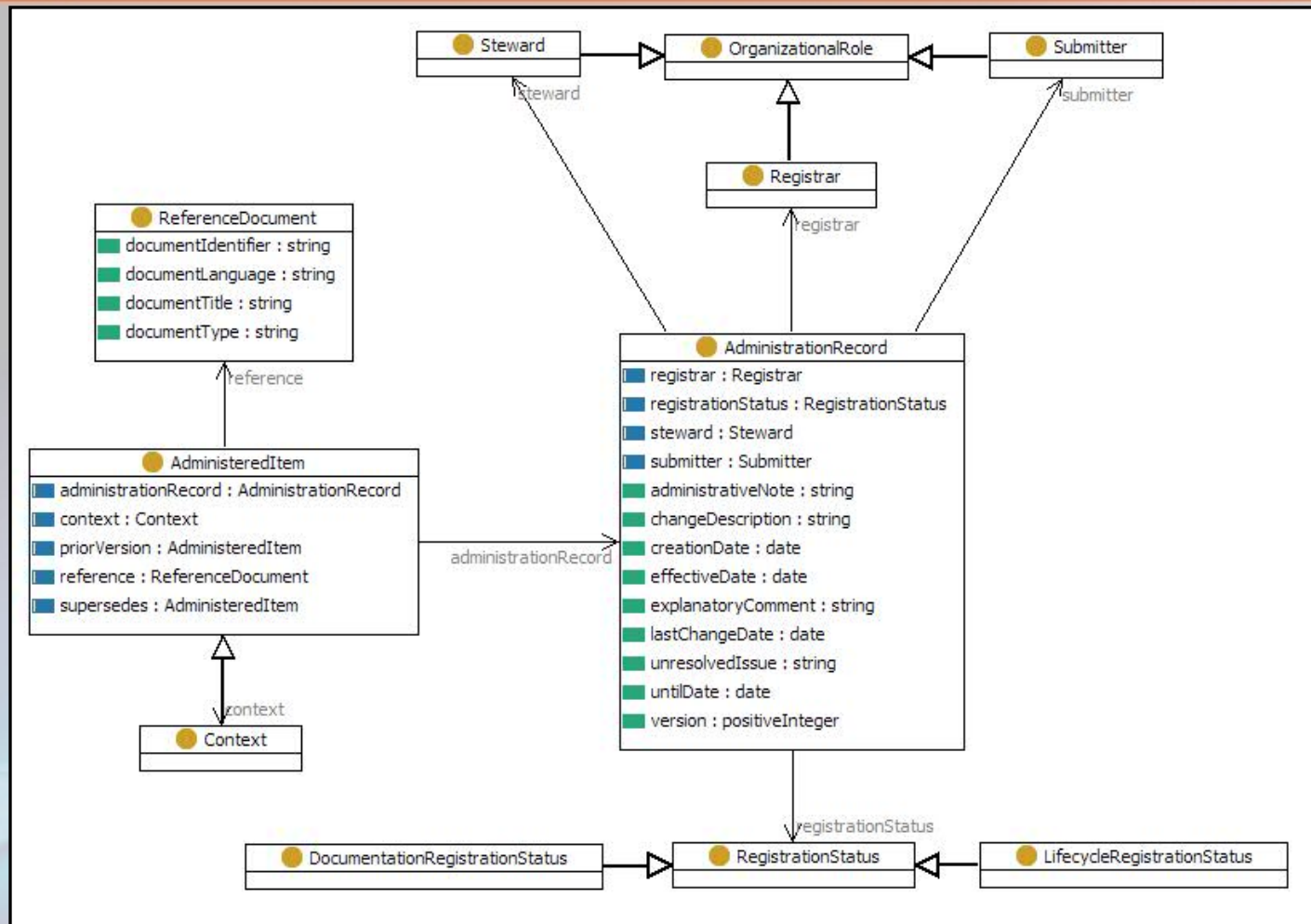


Global Data Standards Repository Release 3

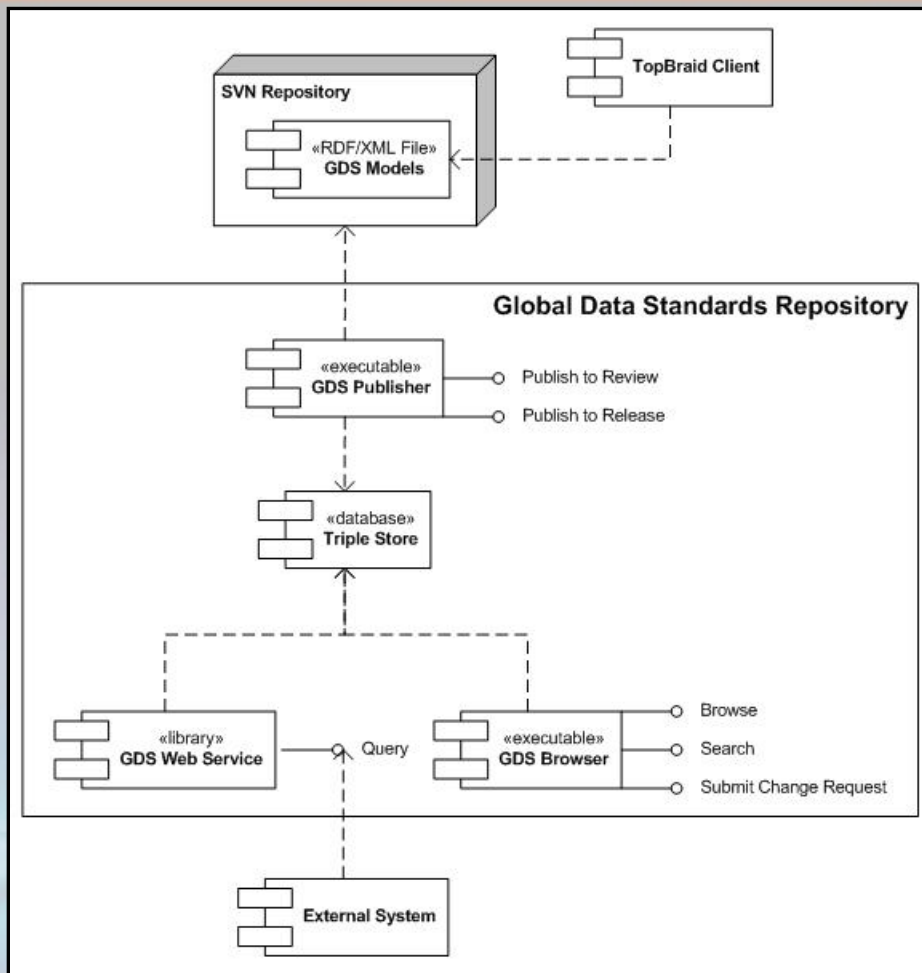
- External content
 - SDTM 1.2, SDTMIG 3.1.2, and amendments
 - CDISC Controlled Terminology
 - NCI Thesaurus
- Integrated data standards, Roche and Genentech
 - Data collection: Safety and every Roche TA (2000 data elements)
 - Data tabulation: SDTM Sponsor Extensions
- Value level metadata
 - Lab measurements (2000)
 - Preferred units and unit conversions
 - Questionnaires (150)

GDSR Metadata Registry Schema

ISO 11179 (partial)



Global Data Standards Repository System Components



Content Management

Content Publishing

Metadata Repository

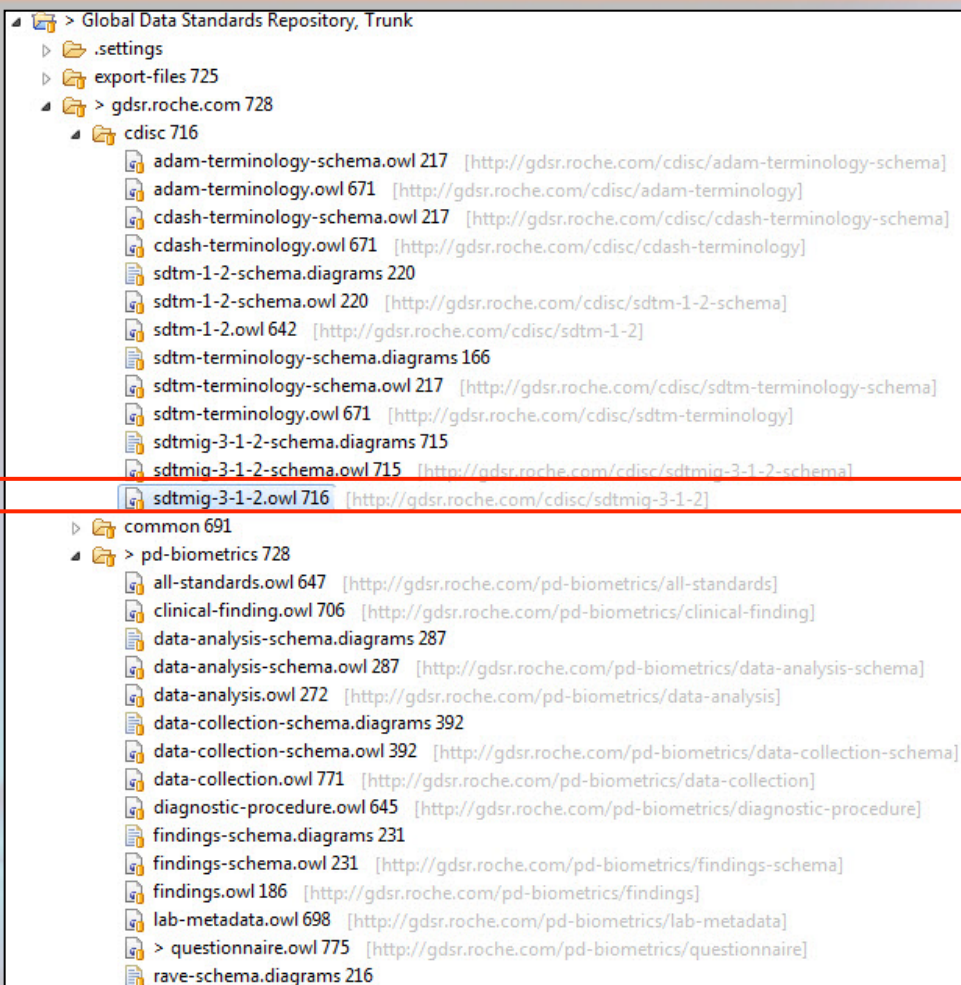
Single Point of Access



Part III

Semantic Modeling Workbench

SDTMIG 3.1.2 Walk-Through



Global Data Standards Repository, Trunk

- .settings
- export-files 725
- gdsr.roche.com 728
 - cdisc 716
 - adam-terminology-schema.owl 217 [<http://gdsr.roche.com/cdisc/adam-terminology-schema>]
 - adam-terminology.owl 671 [<http://gdsr.roche.com/cdisc/adam-terminology>]
 - cdash-terminology-schema.owl 217 [<http://gdsr.roche.com/cdisc/cdash-terminology-schema>]
 - cdash-terminology.owl 671 [<http://gdsr.roche.com/cdisc/cdash-terminology>]
 - sdtm-1-2-schema.diagrams 220
 - sdtm-1-2-schema.owl 220 [<http://gdsr.roche.com/cdisc/sdtm-1-2-schema>]
 - sdtm-1-2.owl 642 [<http://gdsr.roche.com/cdisc/sdtm-1-2>]
 - sdtm-terminology-schema.diagrams 166
 - sdtm-terminology-schema.owl 217 [<http://gdsr.roche.com/cdisc/sdtm-terminology-schema>]
 - sdtm-terminology.owl 671 [<http://gdsr.roche.com/cdisc/sdtm-terminology>]
 - sdtmig-3-1-2-schema.diagrams 715
 - sdtmig-3-1-2-schema.owl 715 [<http://gdsr.roche.com/cdisc/sdtmig-3-1-2-schema>]
 - sdtmig-3-1-2.owl 716** [<http://gdsr.roche.com/cdisc/sdtmig-3-1-2>]
 - common 691
 - pd-biometrics 728
 - all-standards.owl 647 [<http://gdsr.roche.com/pd-biometrics/all-standards>]
 - clinical-finding.owl 706 [<http://gdsr.roche.com/pd-biometrics/clinical-finding>]
 - data-analysis-schema.diagrams 287
 - data-analysis-schema.owl 287 [<http://gdsr.roche.com/pd-biometrics/data-analysis-schema>]
 - data-analysis.owl 272 [<http://gdsr.roche.com/pd-biometrics/data-analysis>]
 - data-collection-schema.diagrams 392
 - data-collection-schema.owl 392 [<http://gdsr.roche.com/pd-biometrics/data-collection-schema>]
 - data-collection.owl 771 [<http://gdsr.roche.com/pd-biometrics/data-collection>]
 - diagnostic-procedure.owl 645 [<http://gdsr.roche.com/pd-biometrics/diagnostic-procedure>]
 - findings-schema.diagrams 231
 - findings-schema.owl 231 [<http://gdsr.roche.com/pd-biometrics/findings-schema>]
 - findings.owl 186 [<http://gdsr.roche.com/pd-biometrics/findings>]
 - lab-metadata.owl 698 [<http://gdsr.roche.com/pd-biometrics/lab-metadata>]
 - > questionnaire.owl 775 [<http://gdsr.roche.com/pd-biometrics/questionnaire>]
 - rave-schema.diagrams 216

SDTMIG 3.1.2 Walk-Through

Class Form

URI: <http://gdsr.roche.com/cdisc/sdtmig-3-1-2-schema#EventsObservationClass>

Annotations

rdfs:isDefinedBy

<http://gdsr.roche.com/cdisc/sdtmig-3-1-2-schema>

rdfs:label

Events Observation Class

skos:definition

The Events class captures planned protocol milestones such as randomization and study completion, and occurrences, conditions, or incidents independent of planned study evaluations occurring during the trial (e.g., adverse events) or prior to the trial (e.g., medical history).

Other Properties

rdf:type

owl:Class

Class Axioms

rdfs:subClassOf

sdtmigs:GeneralObservationClass

Form Diagram Graph Form Layout Source Code

Instances

[Resource]

rdfs:label

rdfs:comment

◆ Table.AE

◆ Table.CE

◆ Table.DS

◆ Table.DV

◆ Table.MH

SDTMIG 3.1.2 Walk-Through

Resource Form

URI: <http://gdsr.roche.com/cdisc/sdtmig-3-1-2#Table.AE>

Annotations

Incoming References

← mms:dataElementDomain ▾

- ◆ Column.AE.AEACN ▾
- ◆ Column.AE.AEACNOTH ▾
- ◆ Column.AE.AEBODSYS ▾
- ◆ Column.AE.AECAT ▾
- ◆ Column.AE.AECONTRT ▾
- ◆ Column.AE.AEDECOD ▾
- ◆ Column.AE.AEDUR ▾
- ◆ Column.AE.AEENDTC ▾
- ◆ Column.AE.AEENDY ▾
- ◆ Column.AE.AEENRF ▾
- ◆ Column.AE.AEENRTPT ▾
- ◆ Column.AE.AEENTPT ▾
- ◆ Column.AE.AEGRPID ▾
- ◆ Column.AE.AELOC ▾
- ◆ Column.AE.AEMODIFY ▾
- ◆ Column.AE.AEOUT ▾
- ◆ Column.AE.AEPATT ▾
- ◆ Column.AE.AEPRESP ▾
- ◆ Column.AE.AEREFID ▾
- ◆ Column.AE.AEREL ▾

Other Properties

sdtmigs:domainCode ▾

S AE ▾

sdtmigs:domainStructure ▾

S One record per adverse event per subject ▾

mms:domainLabel ▾

Adverse Events ▾

mms:domainName ▾

S AE ▾

mms:domainOrdinal ▾

8 ▾

mdr:administrationRecord ▾

◆ mdr:AR.SDTMIG-3-1-2 ▾

mdr:context ▾

◆ mms:Model.SDTMIG-3-1-2 ▾

rdf:type ▾

sdmigs:EventsObservationClass ▾

SDTMIG 3.1.2 Walk-Through

Resource Form

URI: <http://gdsr.roche.com/cdisc/sdtmig-3-1-2#Column.AE.AEOUT>

Annotations

Incoming References

Other Properties

sdtms:dataElementCompliance ▾

◆ sdtms:Classifier.PermissibleVariable ▾

sdtms:dataElementRole ▾

◆ sdtms:Classifier.RecordQualifier ▾

sdtms:dataElementType ▾

◆ sdtms:Classifier.Char ▾

sdtmigs:controlledTermsOrFormat ▾

S (OUT) ▾

sdtmigs:dataElementCodelist ▾

◆ sdtmt:C66768 ▾

sdtmigs:references ▾

S SDTM 2.2.2 ▾

mms:dataElement ▾

◆ sdtm:DE.Event.--OUT ▾

mms:dataElementDescription ▾

S Description of the outcome of an event. ▾

mms:dataElementDomain ▾

◆ Table.AE ▾

mms:dataElementLabel ▾

S Outcome of Adverse Event ▾

mms:dataElementName ▾

S AEOUT ▾

SDTMIG 3.1.2 Walk-Through

Resource Form

URI: <http://gdsr.roche.com/cdisc/sdtm-terminology#C66768>

Annotations

Incoming References

← sdtmts:parentCodelist ▾

🔗 sdtmt:C66768.C17998 ▾

🔗 sdtmt:C66768.C48275 ▾

🔗 sdtmt:C66768.C49494 ▾

🔗 sdtmt:C66768.C49495 ▾

🔗 sdtmt:C66768.C49496 ▾

🔗 sdtmt:C66768.C49498 ▾

← sdtmigs:dataElementCodelist ▾

◆ Column.AE.AEOUT ▾

Other Properties

sdtmts:cdiscDefinition ▾

🔗 A condition or event that is attributed to the adverse event and is the result or conclusion of the adverse event. (NCI) ▾

sdtmts:cdiscSubmissionValue ▾

🔗 OUT ▾

sdtmts:cdiscSynonyms ▾

🔗 Outcome of Event ▾

sdtmts:codelistName ▾

🔗 Outcome of Event ▾

sdtmts:isExtensibleCodelist ▾

🔗 false ▾

sdtmts:nciCode ▾

🔗 C66768 ▾

sdtmts:nciPreferredTerm ▾

🔗 CDISC SDTM Adverse Event Outcome Terminology ▾

rdf:type ▾

🔗 sdtmts:Codelist ▾

SDTMIG 3.1.2 Walk-Through

Resource Form

URI: <http://gdsr.roche.com/cdisc/sdtm-terminology#C66768.C49494>

Annotations

Incoming References

Other Properties

sdtmts:cdiscDefinition

One of the possible results of an adverse event outcome that indicates that the event has not improved or recuperated. (NCI)

sdtmts:cdiscSubmissionValue

NOT RECOVERED/NOT RESOLVED

sdtmts:nciCode

C49494

sdtmts:nciPreferredTerm

Not Recovered or Not Resolved

sdtmts:parentCodelist

sdtmt:C66768

rdf:type

sdtmts:CodelistElement

SDTMIG 3.1.2 Walk-Through

Resource Form

URI: <http://gdsr.roche.com/cdisc/sdtmig-3-1-2#Column.AE.AEOUT>

Annotations

Incoming References

Other Properties

sdtms:dataElementCompliance

◆ sdtms:Classifier.PermissibleVariable

sdtms:dataElementRole

◆ sdtms:Classifier.RecordQualifier

sdtms:dataElementType

◆ sdtms:Classifier.Char

sdtmigs:controlledTermsOrFormat

S (OUT)

sdtmigs:dataElementCodelist

◆ sdtmt:C66768

sdtmigs:references

S SDTM 2.2.2

mms:dataElement

◆ sdtm:DE.Event.--OUT

mms:dataElementDescription

S Description of the outcome of an event.

mms:dataElementDomain

◆ Table.AE

mms:dataElementLabel

S Outcome of Adverse Event

mms:dataElementName

S AEOUT

SDTMIG 3.1.2 Walk-Through

Resource Form

URI: <http://gdsr.roche.com/cdisc/sdtm-1-2-schema#Classifier.PermissibleVariable>

Annotations

rdfs:isDefinedBy

<http://gdsr.roche.com/cdisc/sdtm-1-2-schema>

rdfs:label

Permissible Variable

skos:definition

A Permissible variable should be used in a domain as appropriate when collected or derived. Except where restricted by specific domain assumptions, any SDTM Timing and Identifier variables, and any Qualifier variables from the same general observation class are permissible for use in a domain based on that general observation class. The Sponsor can decide whether a Permissible variable should be included as a column when all values for that variable are null. The sponsor does not have the discretion to not submit permissible variables when they contain data.

Other Properties

rdf:type

sdtms:DataElementCompliance

Incoming References

← sdtms:dataElementCompliance

Column.AE.AEACNOTH

Column.AE.AECAT

Column.AE.AECONTRT

Column.AE.AEDUR

Column.AE.AEENDY

Column.AE.AEENRF

Column.AE.AEENRTPT

Column.AE.AEENTPT



Part IV

Readiness Assessment



Unique Value Proposition

- Convergence of standards and technology
 - Establishment of CDISC standards within the industry
 - MDR based standards as a foundation for metadata driven workflow
 - Upcoming role of semantic web standards and linked data principles
- Well positioned to address significant challenges
 - Manage data standards at a large pharma company
 - Capability to adopt new industry standards as they get published
 - Apply data standards at Roche to
 - Increase quality
 - Reduce cost
 - Improve time to market
 - Support a significantly growing pipeline



Collaborations



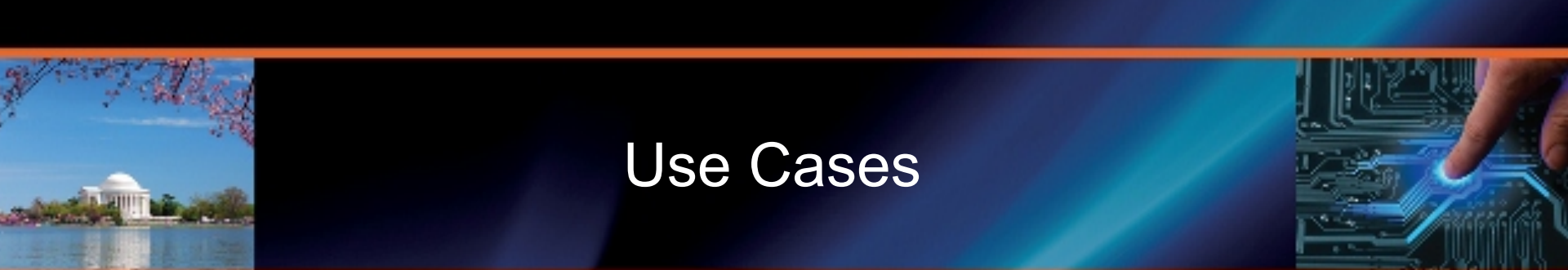
- CDISC and TransCelerate
 - Contribute to the SHARE modeling effort
- cdisc2rdf
 - Precompetitive collaboration to publish CDASH, SDTM, ADaM in RDF
 - First deliverable SDTM 1.2, SDTMIG 3.1.2, SDTM CT published
- National Cancer Institute
 - Publish CDISC Controlled Terminology in RDF by April or June 2013
- Phuse/CSS
 - Roche proposal to explore RDF in the context of emerging technologies



Semantic Technology Emerging?



- Emerging
 - Adoption rate has been slow
 - Adoption rate is growing, but fragmented
 - Toolsets have room to advance considerably
 - Model driven development
- Mature
 - W3C standards are established, mathematically precise
 - RDF is a universal language
 - Linked data principles provide a robust paradigm to integrate data
 - The Roche GDSR system is a production level validated application



Use Cases

- Represent existing CDISC standards in RDF
 - CDASH, SDTM, ADaM, Controlled Terminology
 - Conformance Checks
- Represent future CDISC standards in RDF
 - Protocol, Schedule of Assessments, Trial Design
 - Biomedical concepts (SHARE)
- Represent exchange standards in RDF
- Represent clinical trial data in RDF
- Explore linked data opportunities (BRIDG, NCI Thesaurus, HGNC, etc.)
- Seek collaboration with other RDF publishers (Open PHACTS, etc.)