PODD - Towards An Extensible, Domain-agnostic Scientific Data Management System

Yuan-Fang Li (liyf@itee.uq.edu.au)
eResearch Lab, School of ITEE
The University of Queensland, Australia

PODD - Towards An Extensible, Domain-agnostic Scientific Data Management System

An Ontology-Driven Approach In a Phenomics Setting

Yuan-Fang Li (liyf@itee.uq.edu.au)
eResearch Lab, School of ITEE
The University of Queensland, Australia

Motivation - The Data Deluge

Motivation - The Data Deluge

- Data heterogeneity
    - Images, spreadsheets, text files, publications, *etc.*

Motivation - The Data Deluge

- Data heterogeneity
  - Images, spreadsheets, text files, publications, *etc.*
- Data volume
  - High-throughput & high-resolution processes

Motivation - The Data Deluge

- Data heterogeneity
  - Images, spreadsheets, text files, publications, *etc.*
- Data volume
  - High-throughput & high-resolution processes
- Data evolution
  - Changes in model & data

Data Management Requirements

Data Management Requirements

- Collection
- Distribution & sharing
- Access control
- Archival & versioning
- Discovery & analysis
- Repurposing

PODD Goals

An extensible & domain-independent data management architecture

Related Models & Systems

Related Models & Systems

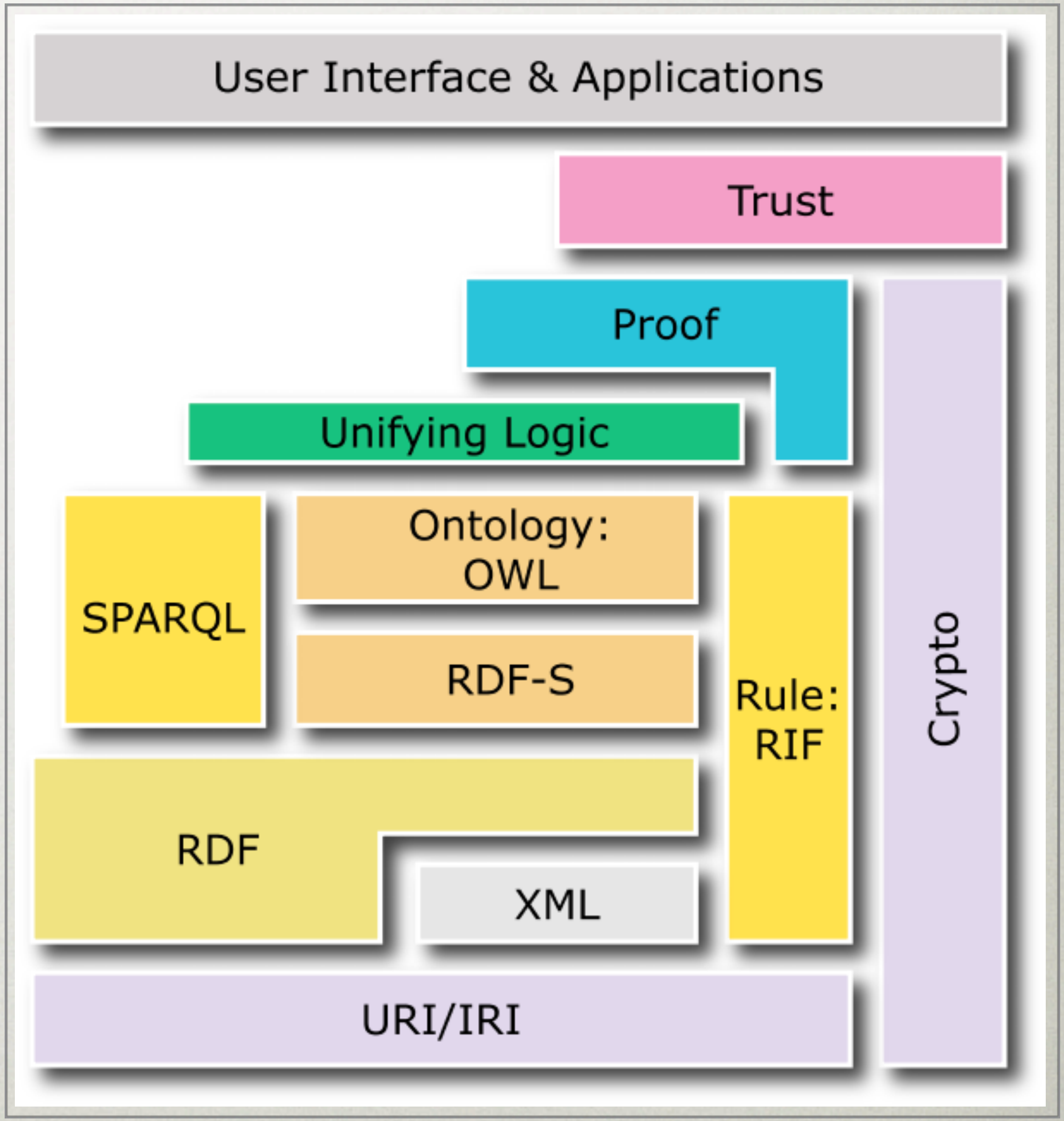
- Models
  - Functional Genomics Experiment Model (FuGe)
    - UML & database based
  - Ontology for Biomedical Investigations (OBI)
    - 26,000+ OWL classes & 10,000+ axioms

Related Models & Systems

- Models
  - Functional Genomics Experiment Model (FuGe)
    - UML & database based
  - Ontology for Biomedical Investigations (OBI)
    - 26,000+ OWL classes & 10,000+ axioms
- Systems
  - VIVO: ontology-based institutional research repository
  - PhonemicDB: a multi-organism phenotype-genotype database
  - Fedora Commons, Apache JackRabbit: digital content repository systems

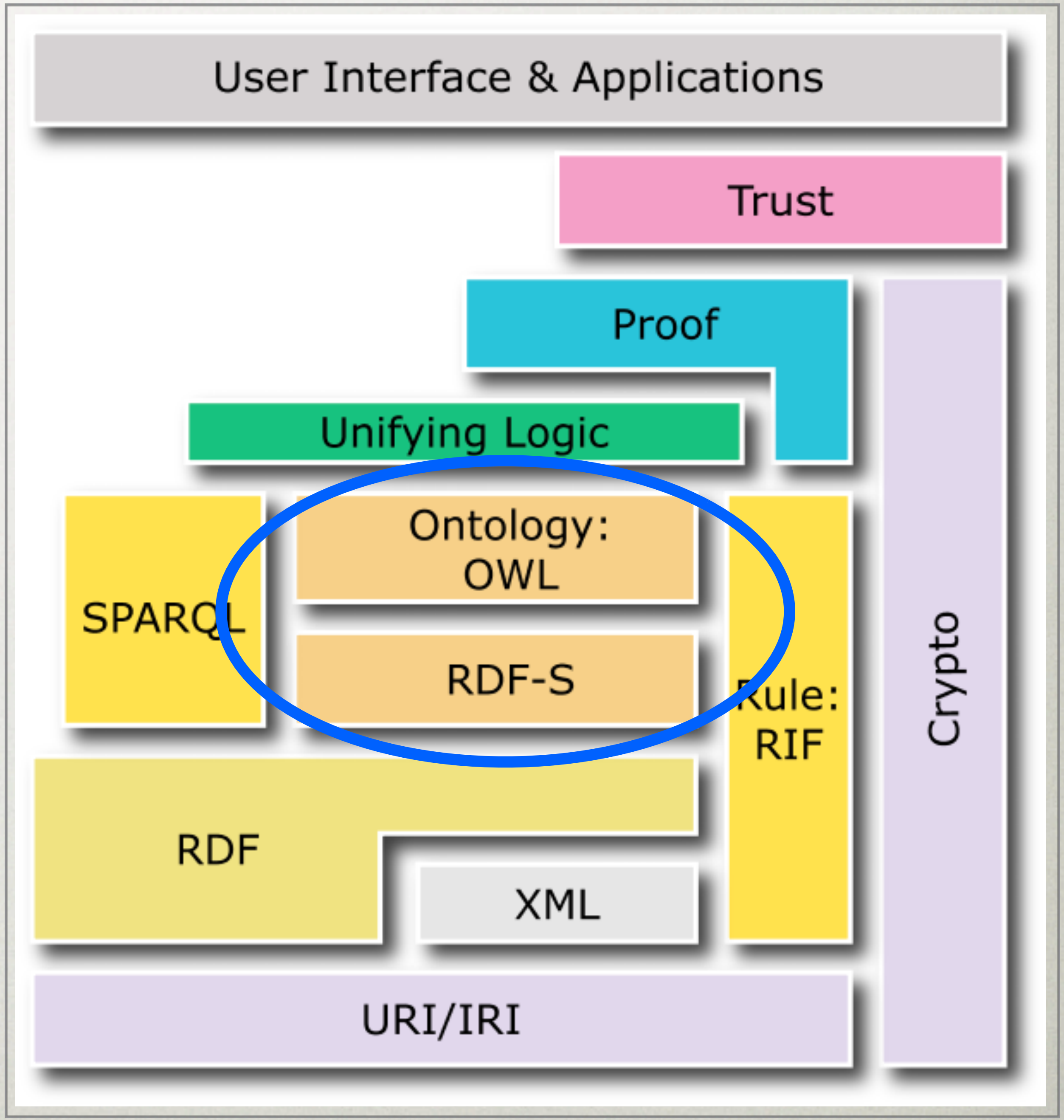
Ontologies

Ontologies



Source: Steve Bratt, <steve@w3.org>

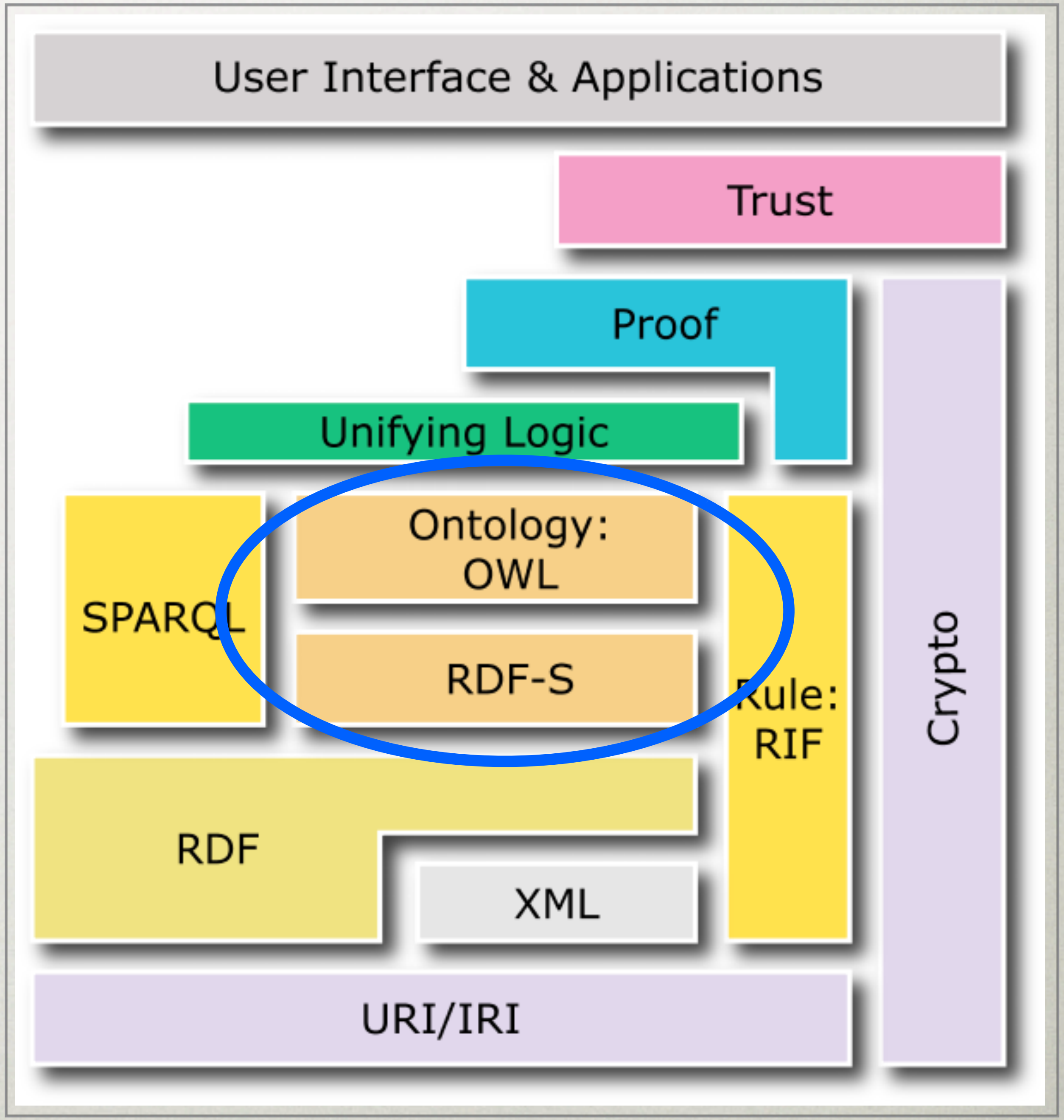
Ontologies



Source: Steve Bratt, <steve@w3.org>

Ontologies

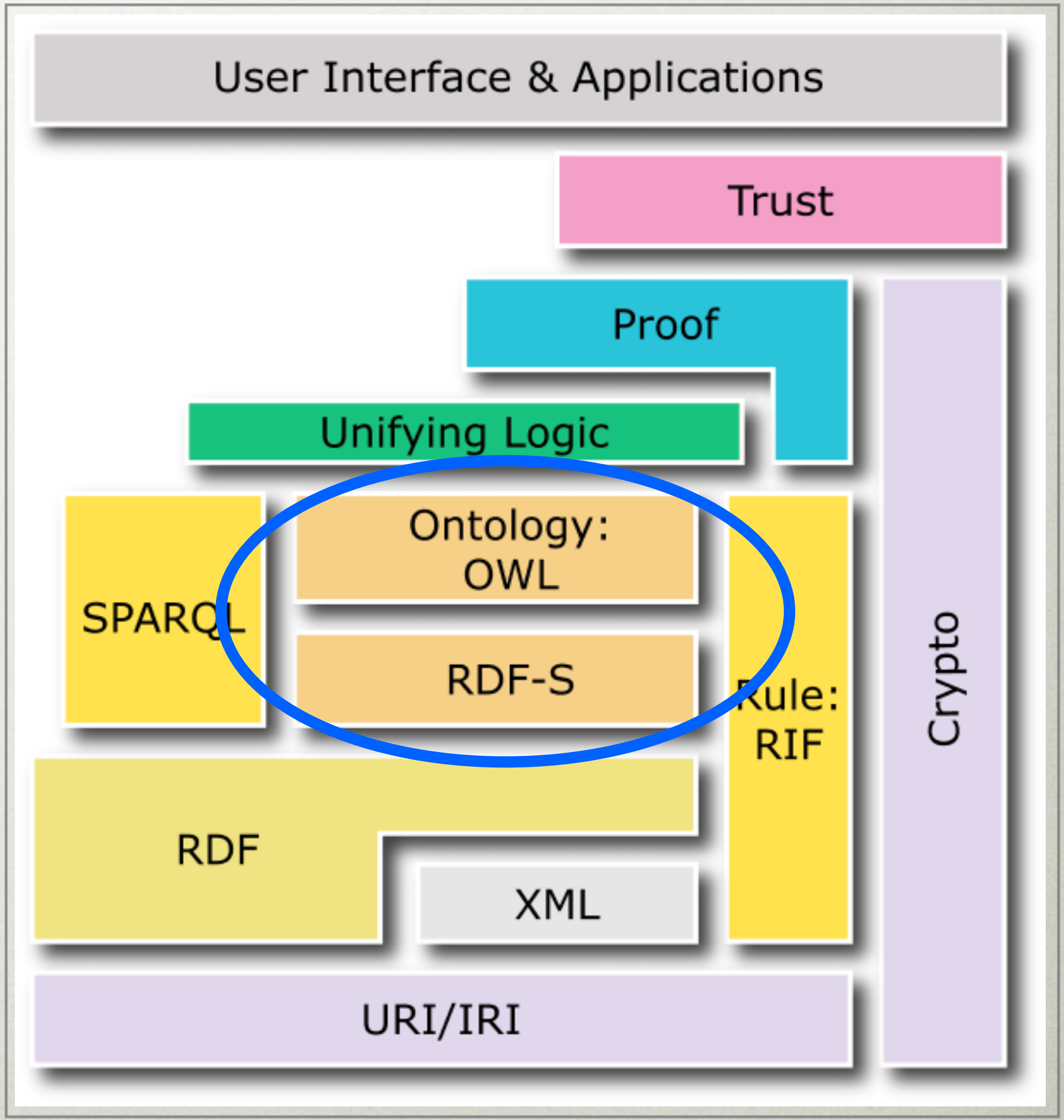
- Expressed in OWL (& RDF Schema)
  - Provides syntax & **semantics** - enables reasoning
  - Expressivity vs decidability



Source: Steve Bratt, <steve@w3.org>

Ontologies

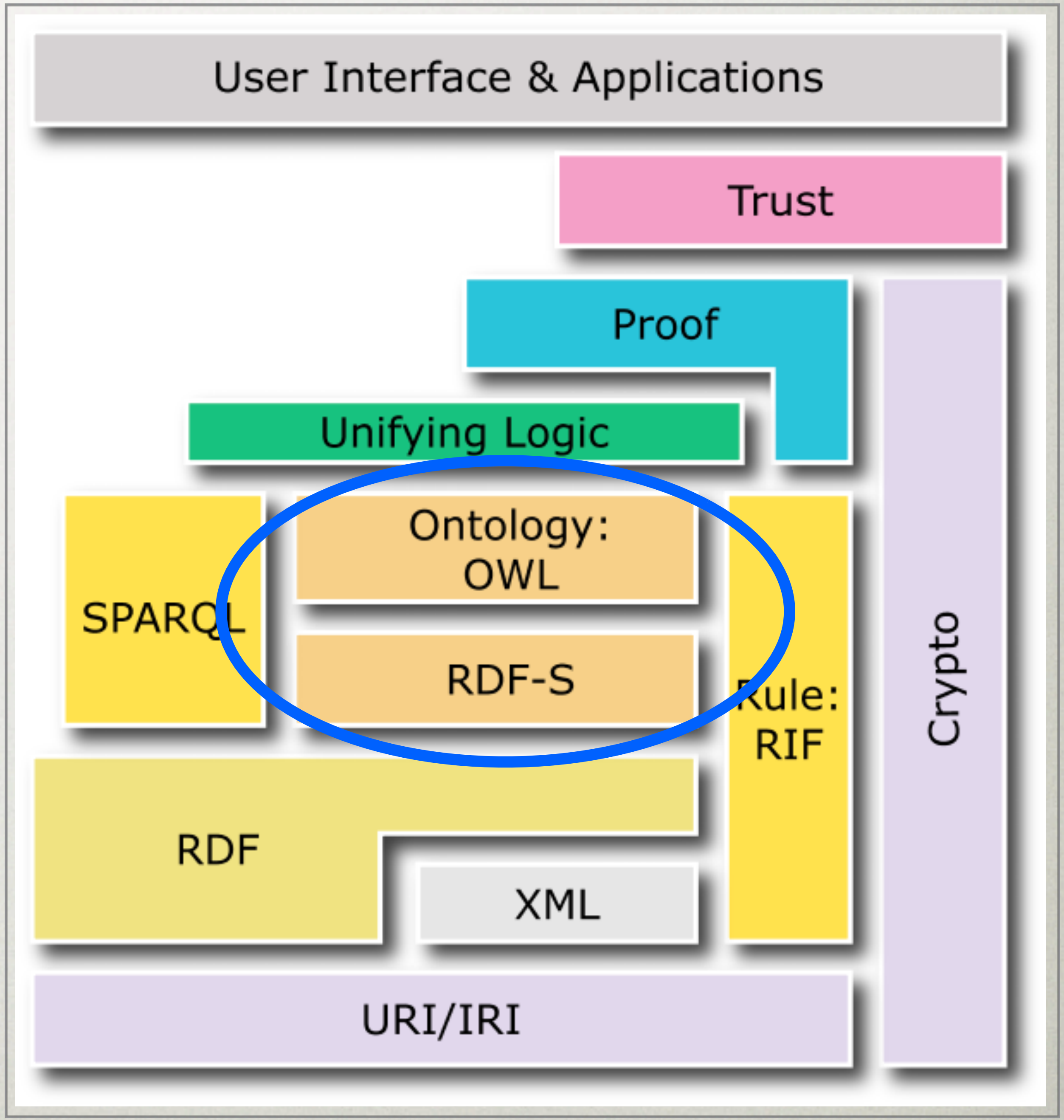
- Expressed in OWL (& RDF Schema)
  - Provides syntax & **semantics** - enables reasoning
  - Expressivity vs decidability
- Designed to be open & interoperable
  - Facilitates sharing, reuse & integration



Source: Steve Bratt, <steve@w3.org>

Ontologies

- Expressed in OWL (& RDF Schema)
  - Provides syntax & **semantics** - enables reasoning
  - Expressivity vs decidability
- Designed to be open & interoperable
  - Facilitates sharing, reuse & integration
- Maturing technology stacks
  - APIs, reasoners, triple stores, query engines



Source: Steve Bratt, <steve@w3.org>

The Ontology-driven Approach

The Ontology-driven Approach

- Basics: ontologies as domain models for scientific experiments data
    - Domain-**independent** & domain-**specific** ontologies

The Ontology-driven Approach

- Basics: ontologies as domain models for scientific experiments data
    - Domain-**independent** & domain-**specific** ontologies
- Models concepts/objects as ontological entities
    - OWL classes, individuals, restrictions

The Ontology-driven Approach

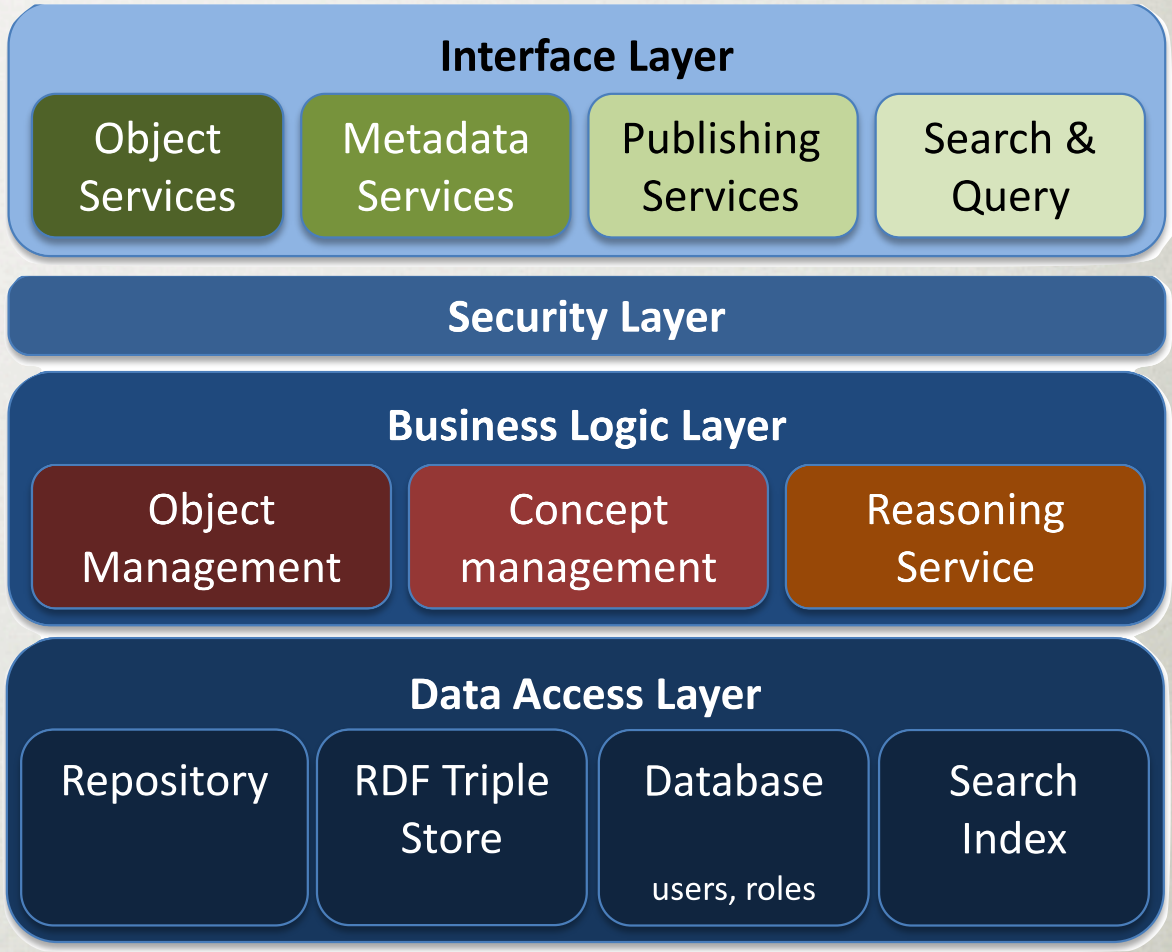
- Basics: ontologies as domain models for scientific experiments data
  - Domain-**independent** & domain-**specific** ontologies
- Models concepts/objects as ontological entities
  - OWL classes, individuals, restrictions
- Ontologies & RDF central to all operations in the data lifecycle

The Ontology-driven Approach

- Basics: ontologies as domain models for scientific experiments data
  - Domain-**independent** & domain-**specific** ontologies
- Models concepts/objects as ontological entities
  - OWL classes, individuals, restrictions
- Ontologies & RDF central to all operations in the data lifecycle
- Aims: improved ***extensibility*** & data ***integration***

The PODD System Architecture

- PODD: **P**henomics **O**ntology **D**riven **D**ata System
- Ontologies the core of the architecture
- Objects represented semantically
  - Semantics (metadata) captured in RDF
- Repository operations on RDF:
  - Ingestion, retrieval, update, query & search, export



PODD Ontologies

PODD Ontologies

- *Extensibility* through inheritance & versioning
- *Integration* through ontology alignment/ mapping

PODD Ontologies

- *Extensibility* through inheritance & versioning
- *Integration* through ontology alignment/mapping

Domain concepts	OWL classes
Attributes & relations	OWL restrictions
Domain objects	OWL individuals
Comments, descriptions	OWL/RDF annotations

PODD Ontologies

PODD Ontologies

- Models scientific experiments

PODD Ontologies

- Models scientific experiments
- Organizes data *logically*
  - Represented as metadata objects
  - Parent-child relationships
  - References relationships

PODD Ontologies

- Models scientific experiments
- Organizes data *logically*
  - Represented as metadata objects
  - Parent-child relationships
  - References relationships
- Base ontology: domain independent

PODD Ontologies

- Models scientific experiments
- Organizes data *logically*
  - Represented as metadata objects
  - Parent-child relationships
  - References relationships
- Base ontology: domain independent
- Phenomics ontology: domain specific

PODD Ontologies

Base

$PODDConcept \sqsubseteq \top$

$\top \sqsubseteq \forall\, contains.PODDConcept$

$isContainedBy \sqsubseteq (^{-}contains)$

$PODDConcept \sqsubseteq \leq 1\ isContainedBy$

$\top \sqsubseteq refersTo.PODDConcept$

$$\begin{aligned} Project \sqsubseteq \quad & = 1\ hasProjectPlan \sqcap \\ & \geq 1\ hasInvestigation \sqcap \\ & = 1 hasStartDate \sqcap \\ & \leq 1\ hasPublicationDate \sqcap \\ & \ldots \end{aligned}$$

PODD Ontologies

PODD Ontologies

Base

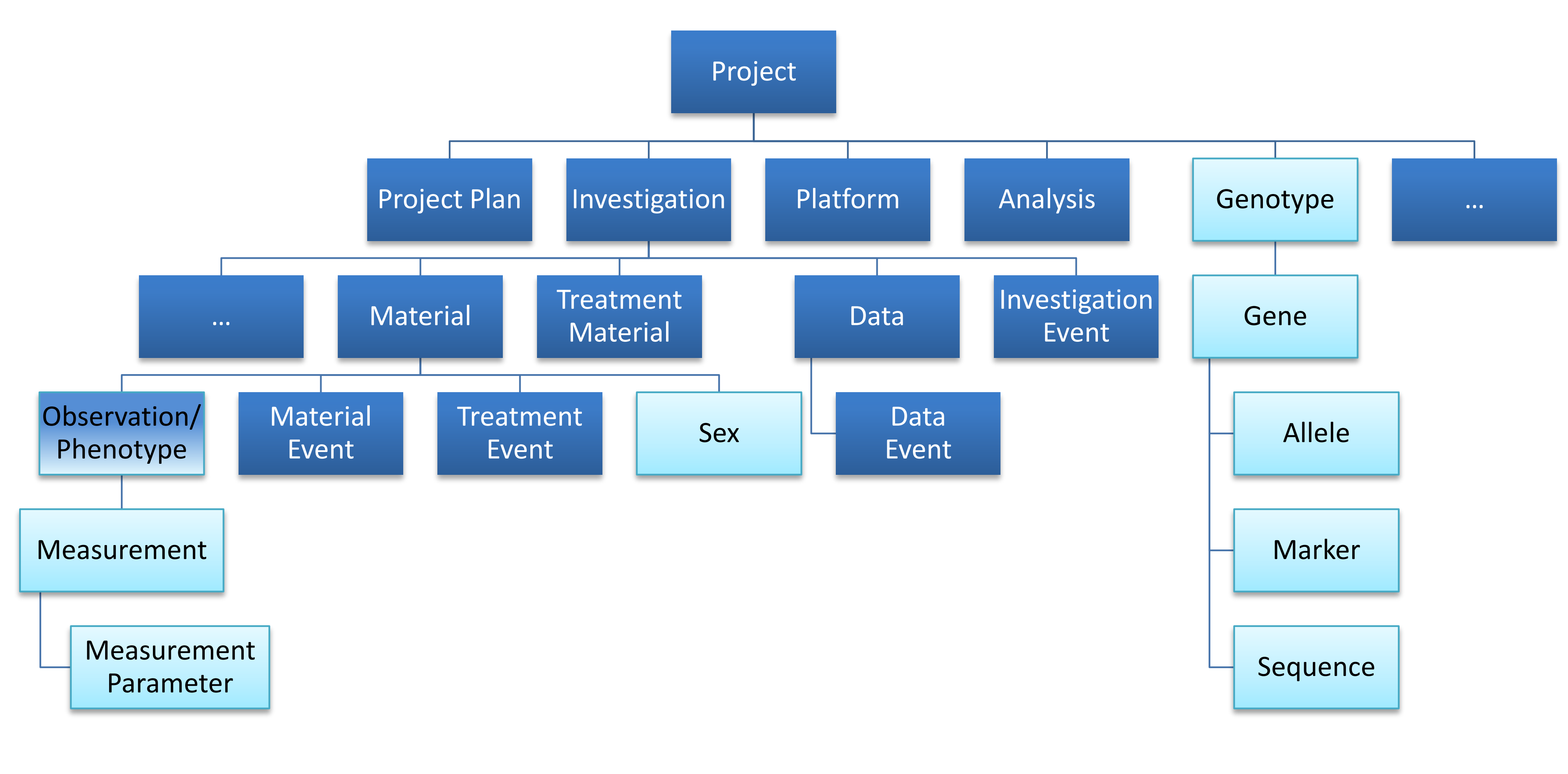
$$
\begin{aligned}
& PODDConcept \sqsubseteq \top \\
& \top \sqsubseteq \forall\, contains.PODDConcept \\
& isContainedBy \sqsubseteq (^{-} contains) \\
& PODDConcept \sqsubseteq \leq 1\ isContainedBy \\
& \top \sqsubseteq refersTo.PODDConcept
\end{aligned}
$$

$$
\begin{aligned}
Project \sqsubseteq \quad & = 1\ hasProjectPlan\ \sqcap \\
& \geq 1\ hasInvestigation\ \sqcap \\
& = 1 hasStartDate\ \sqcap \\
& \leq 1\ hasPublicationDate\ \sqcap \\
& \ldots
\end{aligned}
$$

Phenomics

$$
\begin{aligned}
Genotype \sqsubseteq \quad & PODDConcept\ \sqcap \\
& \forall\, hasGene.Gene\ \sqcap \\
& \leq 1\ hasEcotype\ \sqcap \\
& \leq 1\ hasSubspecies\ \sqcap
\end{aligned}
$$

$$
\begin{aligned}
Project \sqsubseteq \quad & \forall\, hasGenotype.Genotype\ \sqcap \\
Material \sqsubseteq \quad & \forall\, hasPhenotype.Phenotype\ \sqcap \\
& \forall\, refersToGenotype.Genotype
\end{aligned}
$$



PODD Ontologies

The PODD System

The PODD System

- Making use of mature technologies
  - OWLAPI, Pellet, Fedora Commons, Sesame, Lucene & Solr, *etc.*

The PODD System

- Making use of mature technologies
  - OWLAPI, Pellet, Fedora Commons, Sesame, Lucene & Solr, *etc.*
- Facilitates extensibility & evolution
  - Ontology reasoning instead of DB integrity constraint checking
  - Data & metadata are all versioned

The PODD System

- Making use of mature technologies
    - OWLAPI, Pellet, Fedora Commons, Sesame, Lucene & Solr, *etc.*
- Facilitates extensibility & evolution
    - Ontology reasoning instead of DB integrity constraint checking
    - Data & metadata are all versioned
- System exploration
    - Search, browsing, SPARQL querying, *etc.*

Conclusion

Conclusion

- ✓ What we have done
  - ✓ An ontology-driven architecture for improving extensibility
  - ✓ A set of ontologies as domain models
  - ✓ A system for phenomics data management

Conclusion

- ✓ What we have done
  - ✓ An ontology-driven architecture for improving extensibility
  - ✓ A set of ontologies as domain models
  - ✓ A system for phenomics data management
- ? Future works
  - ? Ontology/vocabulary mapping
  - ? Annotation of domain objects
  - ? Workflow support

Acknowledgment

Acknowledgment

- Co-authors: *Gavin Kennedy, Faith Davies, Jane Hunter* (UQ)

Acknowledgment

- Co-authors: *Gavin Kennedy, Faith Davies, Jane Hunter* (UQ)

- Colleagues: Xavier Sirault, Kai Xu, Philip Wu

Acknowledgment

- Co-authors: *Gavin Kennedy, Faith Davies, Jane Hunter* (UQ)

- Colleagues: Xavier Sirault, Kai Xu, Philip Wu

- Supported by **Australian National Data Service** (ANDS) & **Australian Research Collaboration Service** (ARCS)