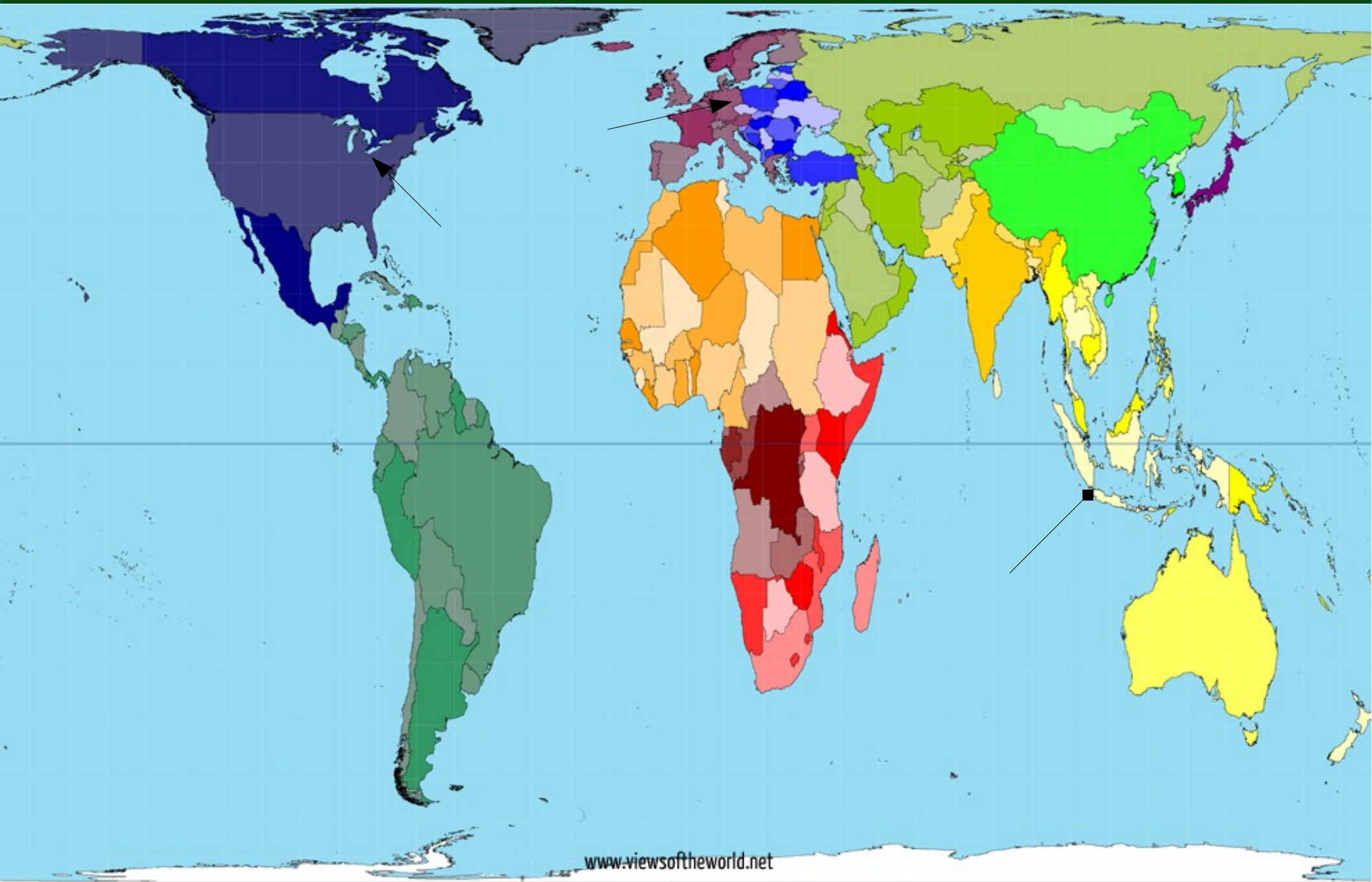
DaSe Lab

Cross-Repository Data Integration using Ontology Design Patterns

Adila Alfa Krisnadhi
DaSe Lab for Data Semantics
Wright State University

WRIGHT STATE UNIVERSITY


www.viewsoftheworld.net

About DaSe Lab

- 2 Faculty members: Dr. Pascal Hitzler & Dr. Michelle Cheatham
- 5 full time PhD students (+ a few master's and part-time PhD students)
- Topics:
  - Foundational research in
    - Formalisms for representation of information and knowledge
    - Algorithms for reasoning with data and knowledge
    - Algorithms for knowledge acquisition
  - Applied research in
    - Semantic Web
    - Data and knowledge integration
    - Linked and Big Data
    - Ontology-based systems
    - Ontology modeling and engineering

What is this about?

- Ontology-based data integration
- Domain: geoscience, starting with ocean science
- Modular ontology engineering approach using ontology patterns.
- Aiming for flexibility and extensibility.
- As respectful as possible to individual modeling choices.

Background

EarthCube

DaSe Lab

- “community-driven knowledge infrastructure for geosciences”
  - well-connected environment to share data and knowledge in an open, transparent, and inclusive manner, accelerating our ability to understand and predict the Earth system
- Consists of various projects (building blocks, RCNs, SIGs) to:
  - develop key technologies,
  - promote community building,
  - explore integrative systems, and
  - prototype a governance structure.

EarthCube needs, SW provides

requires

Interoperability

Conceptual modeling

Information Integration

Formalized vocabulary

Intelligent search

Data publishing support

provides

OceanLink

DaSe Lab

- An EarthCube building block
- Applying semantic technologies for integration of existing ocean science data repositories
- Flexible, extendible, modular, respecting heterogeneity

NSF award 1354778 "EAGER: Collaborative Research: EarthCube Building Blocks, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery."

Geosciences Data Repositories
(a very small snapshot)

DaSe Lab

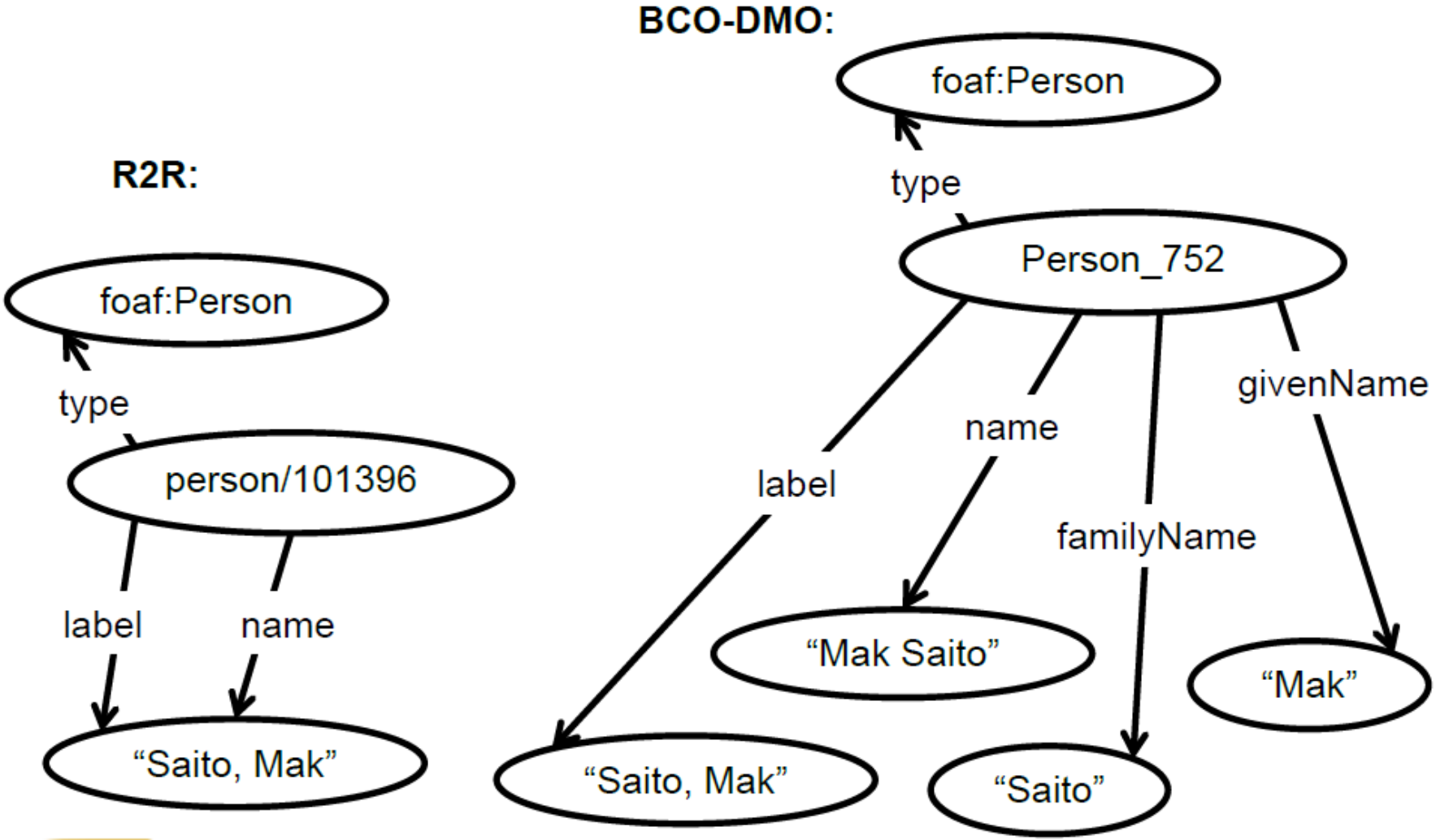
- Oceanographic data – BCO-DMO: >6000 datasets with supporting documents from 24 programs, 229 projects, 1673 deployments
- Field expeditions data – R2R: 400 expeditions per year; 3
- Conference and funded award abstracts – AGU: 30 mil. triples
- Theses, reports, journal articles – MBLWHOI Library: 5500 text documents
- Solid earth data – IEDA: hi-res bathymetry and samples from >730 cruises
- Marine geological data – IMLGS
- Ecological data – LTER
- Antarctic data – AMD
- Ocean drilling data – IODP
- Physiographic gazetteers – MRD
- ….

Technical AND Social Challenges

DaSe Lab

- **Technical challenge**:
  - Lack of interoperability in terms of formats, etc.
  - Semantic and content heterogeneity
- **Social challenge**: Data owners/providers are reluctant/unwilling to participate in sharing and integration if:
  - conceptual changes have to be made to their data repositories
  - their usual business process have to be reworked, or even worse, completely discarded (note: each data repository usually represents its own research sub-community);
  - the global schema is too difficult to comprehend and manage (because the data owners are also the data consumers);
  - retrieving their own data becomes more complicated using the integrated system.

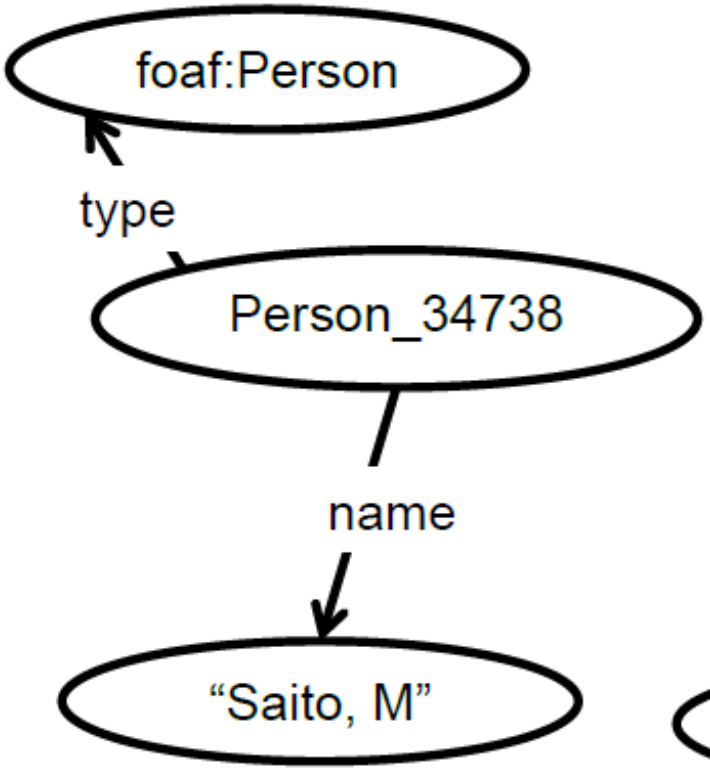
Different ways of modeling Person



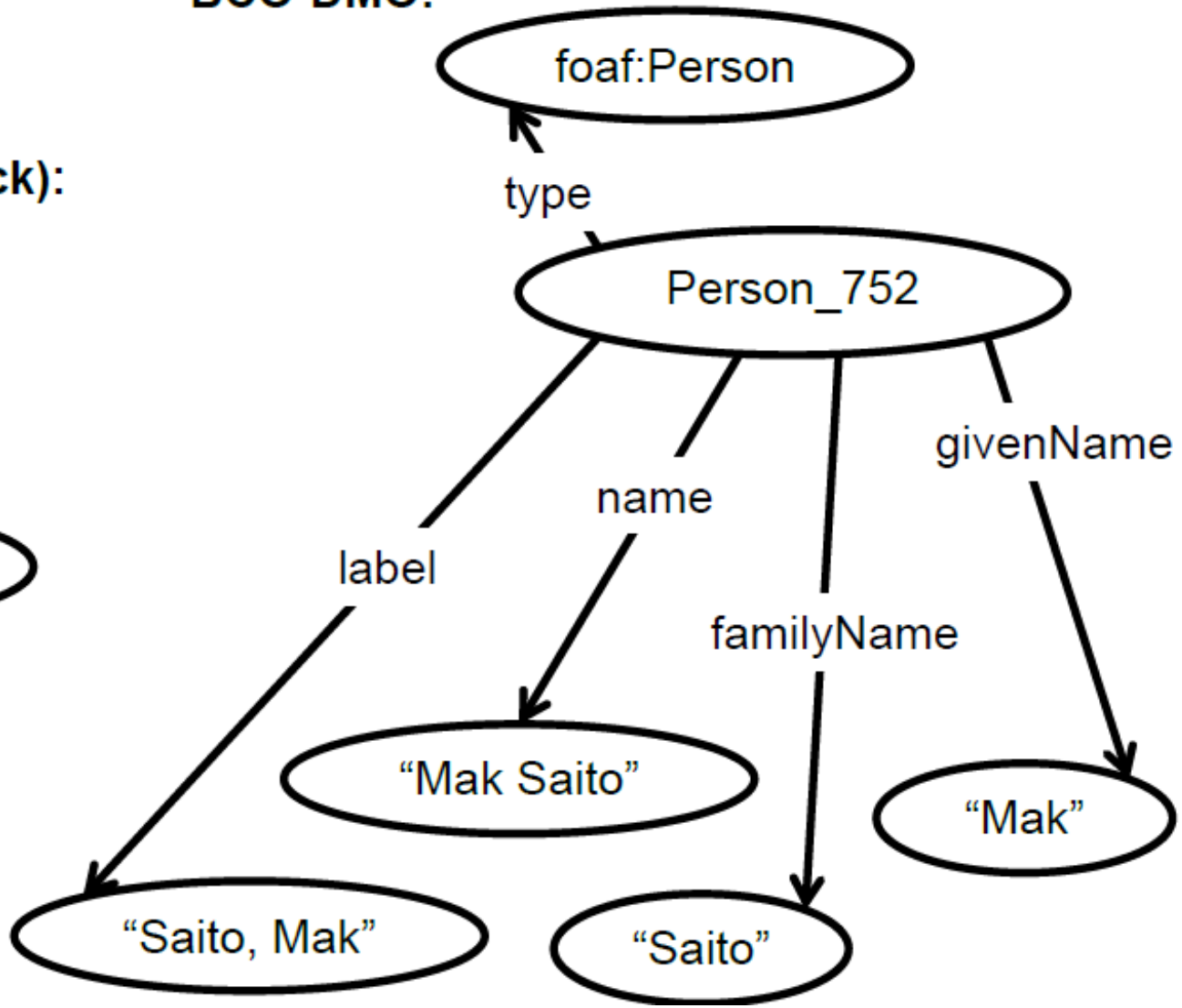
Different ways of modeling Person

DaSe Lab

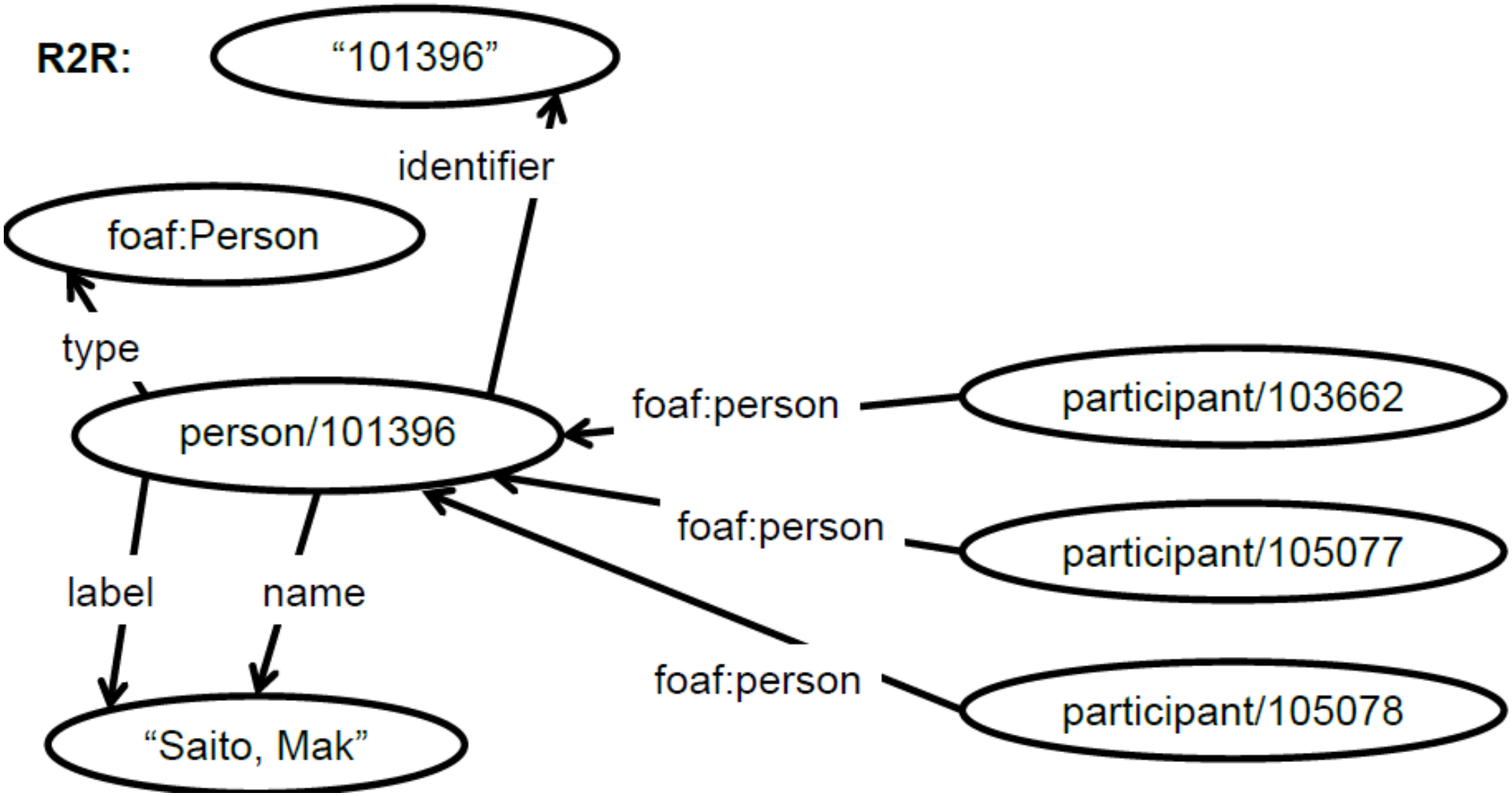
AGU Abstracts (Tom Narock):



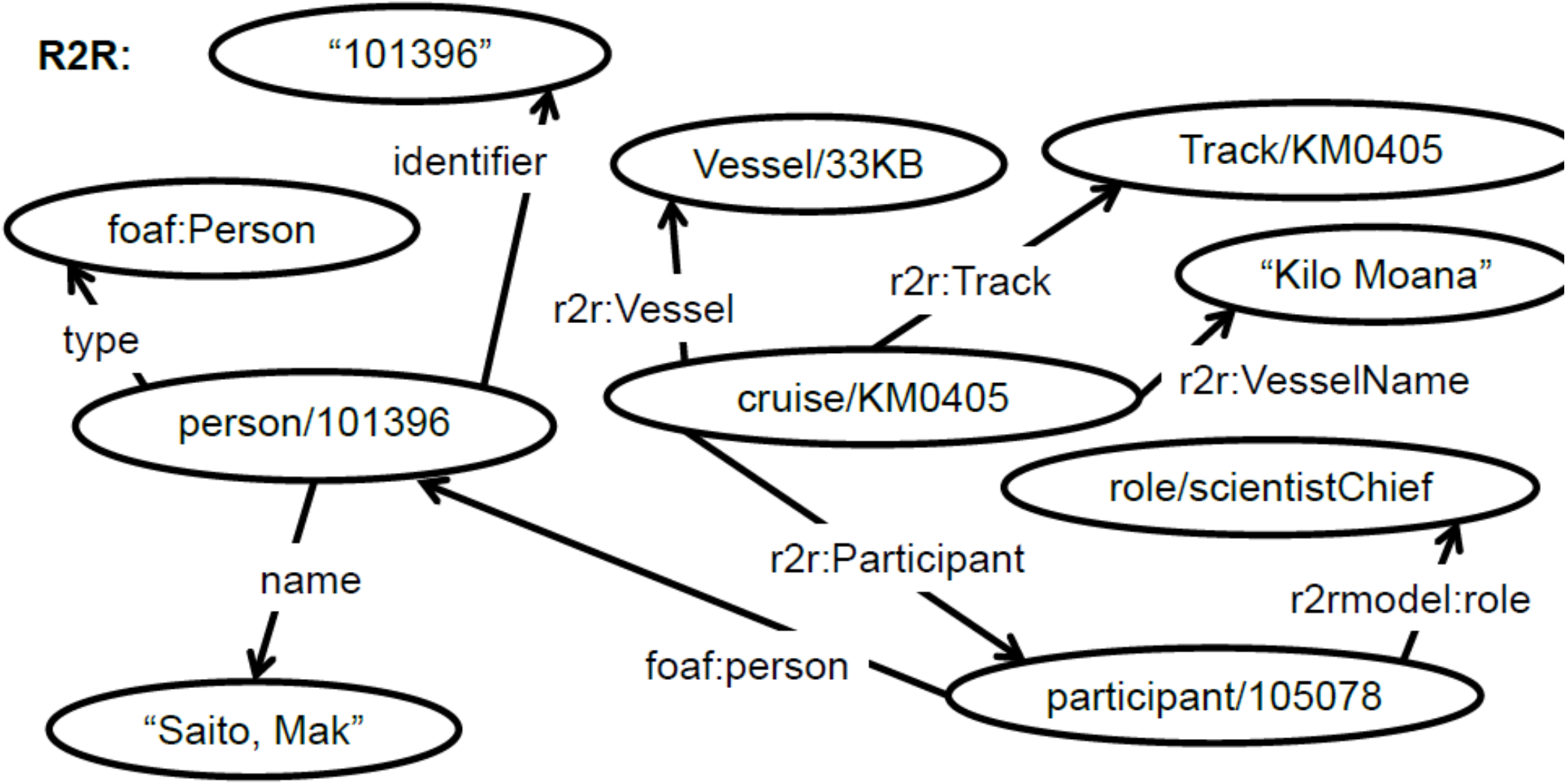
BCO-DMO:



Different ways of modeling Person



Different ways of modeling Person



Ontology (Design Pattern) Based Data Integration

Data Integration Problem

DaSe Lab

- Data Integration Problem:
  - The problem of providing a user with a unified view over data residing at different, autonomous, and possibly heterogeneous data sources.
  - The unified view is called the *global schema*.
  - When a user issues a query over the global schema, the system should translate it into queries over suitable data sources and assemble the results into the final answer to be presented to the user.

Ontology-based Data Integration

DaSe Lab

- Interoperability through RDF standard:
  - Web standards on data model (RDF: a set of triples), querying language and protocol (SPARQL; REST)
  - Each data repository only needs to make their data available for SPARQL querying.
- Vocabulary of global schema is defined using ontology.
  - Allows more expressive conceptualization using a language like OWL – ontology given as a set of logical axioms.
  - Bridging semantic heterogeneity
  - *But, using overrarching monolithic ontology is too difficult, complex, unwieldy, not flexible, too constraining, not easily extendible, etc.*

Modular Ontology Engineering (Iterative)

DaSe Lab

- Model one key notion at a time
- Keep ontological commitments minimum (avoid too constraining axiomatic statements)
- Gathered constraints & requirements are formalized (e.g., with OWL) outside the modeling sessions
- Document the translation and communicate it with the domain people
  - Useful if domain experts can test the resulting patterns against real data

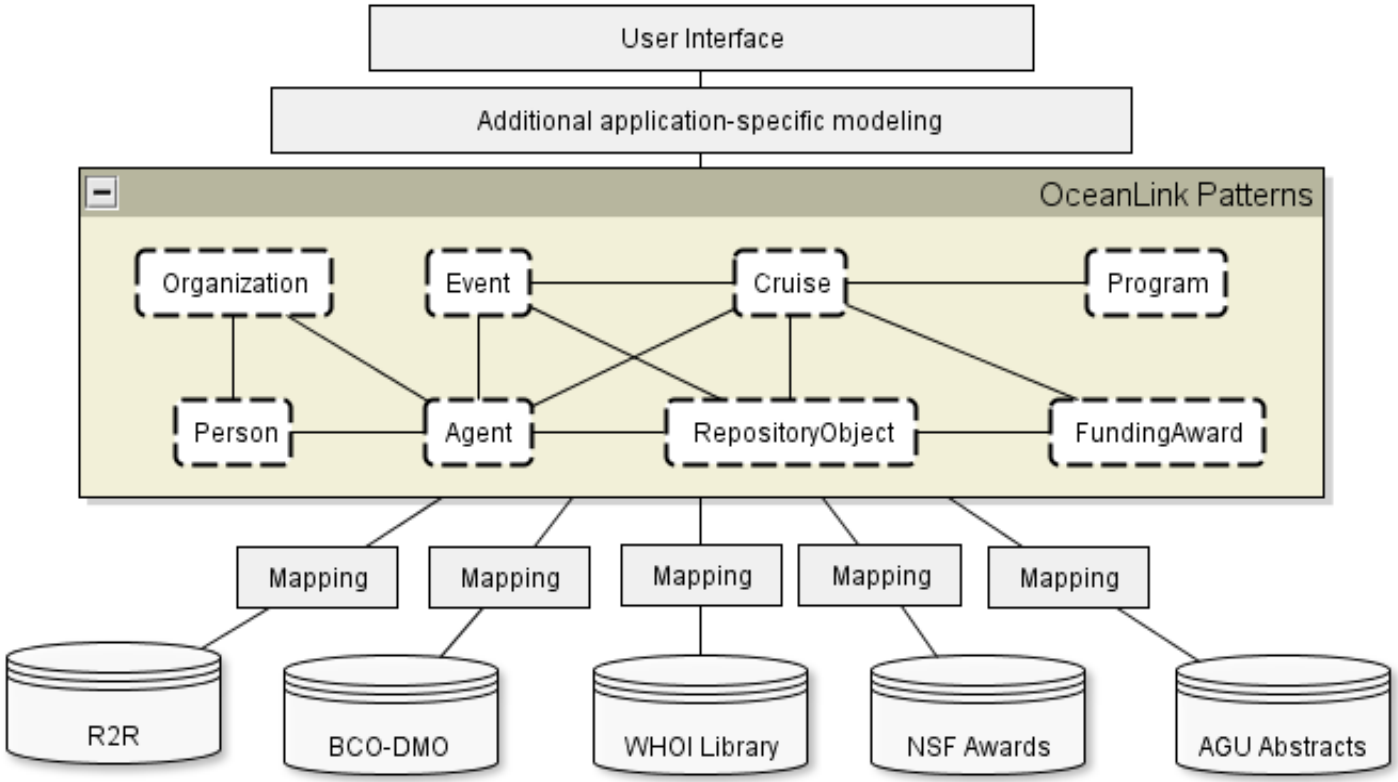
Ontology Design Patterns

DaSe Lab

- Reusable solution to some frequently occurring ontological modeling problem emerging in different domains
- **Content pattern**: encapsulates one key notion in a particular domain, providing modular, reusable, replaceable pieces.
- By reusing generic patterns (but leaving the relationships between patterns to a specific assembly for a specific purpose), we can have a reuse while respecting heterogeneity.
- Patterns “follow” data, rather than data “follow” the patterns.

OceanLink Architecture

- Data providers are actively involved (have a say) in the creation of the global schema.
- Definition of mappings is essentially in the hand of the data providers (knowledge engineers may help if needed, of course).



Identifying Key Notions

DaSe Lab

- BCO-DMO
  - Award, Cruise, Device, Fileset, Format, Geometry, Holding, Model, Organization, Participant, Person, Port, Product, Program, Report, Repository, Track, Vessel
- R2R
  - Program, Project, Deployment, Dataset, Instrument, Parameter, People, Affiliation, Funding Source, Award
- AGU Abstract
  - Meeting, Meeting Section, Meeting Session, Abstracts of Contribution (Poster, Talk, etc.), People, Organization
- MBLWHOI Library
  - Article, Dataset, Book, Report, Person, Organization
- NSF Funded Award Repository
  - Award, Person, Organization, Publication, Dataset

Patterns from Key Notions

- Cruise
- Vessel
- Trajectory
- Person
- Organization
- Roles of Agents
- Repository Object
- Dataset
- and a few other patterns (about 15 in total)

We are not starting from zero, of course.

Engineering an Ontology Design Pattern

Cruise Pattern: Competency Questions

DaSe Lab

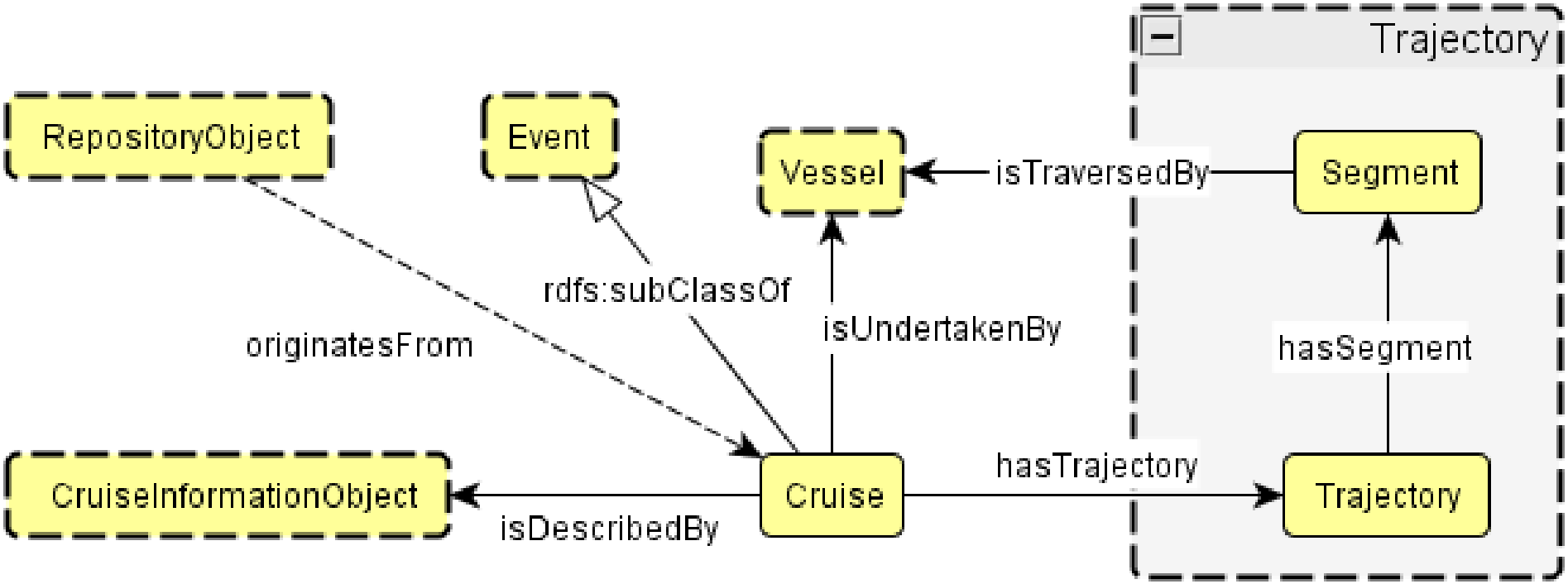
- Find all cruises passing through Gulf of Maine in August 2013.
- List all cruise vessels that departed from Woods Hole in 2012.
  - Cruise is conducted on some vessel, running on some track (trajectory), which is bounded by some spatio-temporal boundary.
- Find the chief scientists of any cruise that collected samples of carbon-isotope data in Lake Superior.
  - Some people/organization performs certain roles in a cruise.
- What datasets were produced by the cruise AE0901?
  - Cruise has some (unique?) identifier, name, etc
  - Cruise (activities) may result in some reports, datasets, etc.
- Which cruises are funded by the NSF award DBI-0424599?
- List all cruises under the Ocean Flux Program.
  - Cruise is funded by some funding award or part of some program.

Can we reuse existing patterns?

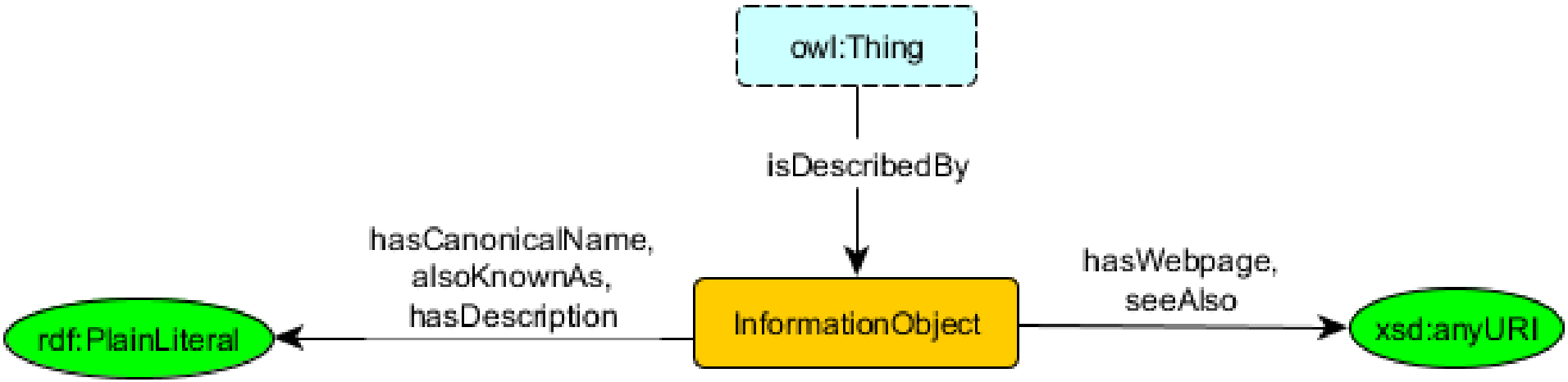
DaSe Lab

- Some of the use cases given by the CQs fit some existing patterns.
  - Simple Event Model [van Hage, et al., JWS 2011]; lack of formalization
  - Semantic Trajectory [Hu, et al., COSIT 2013]; formalization in OWL, but some details need to be modified
  - Information Object from DOLCE [Oberle, et al., JWS 2007]; pattern only used as an inspiration, formalization is redone for our purpose
- Reuse is not simply done via ontology import, but requires adjustment and massaging to make it appropriate for our needs.
- The challenge is finding appropriate existing patterns for reuse, and performing the suitable adjustments

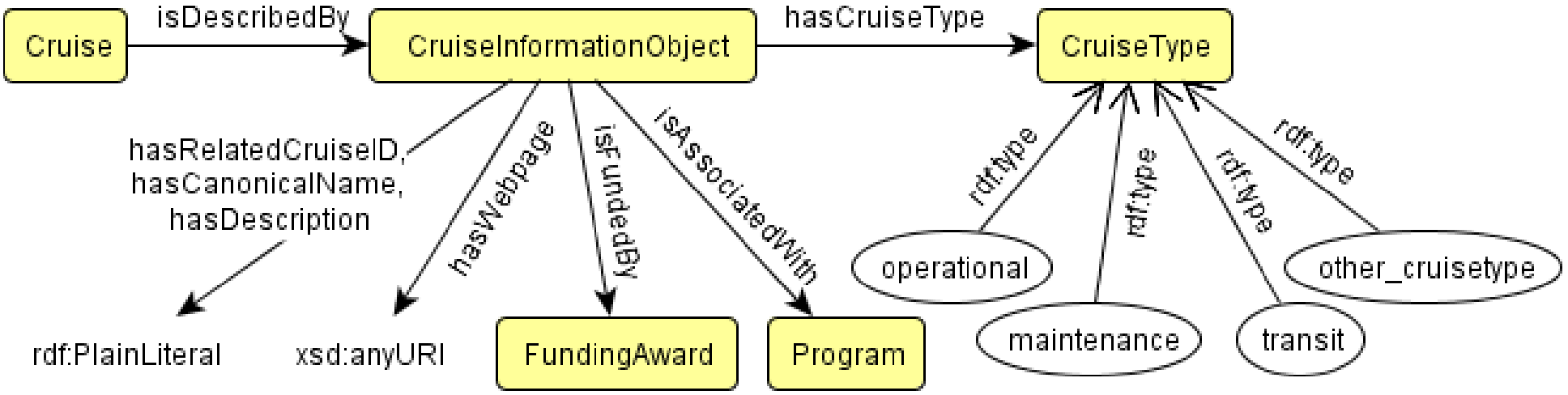
Cruise Pattern



InformationObject pattern



Cruise Information Object



Cruise Information Object (OWL)

DaSe Lab

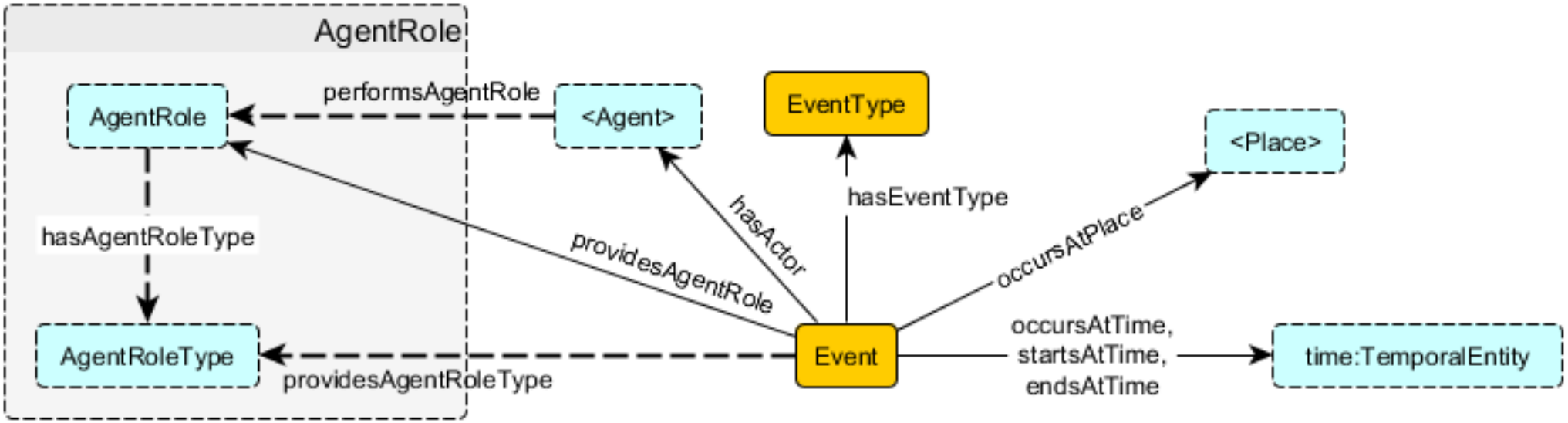
$$\mathsf{Cruise} \sqsubseteq (=1\ \mathsf{isDescribedBy}.\mathsf{CruiseInformationObject}) \quad (22)$$

$$\mathsf{CruiseInformationObject} \sqsubseteq (=1\ \mathsf{hasCruiseType}.\mathsf{CruiseType}) \quad (23)$$

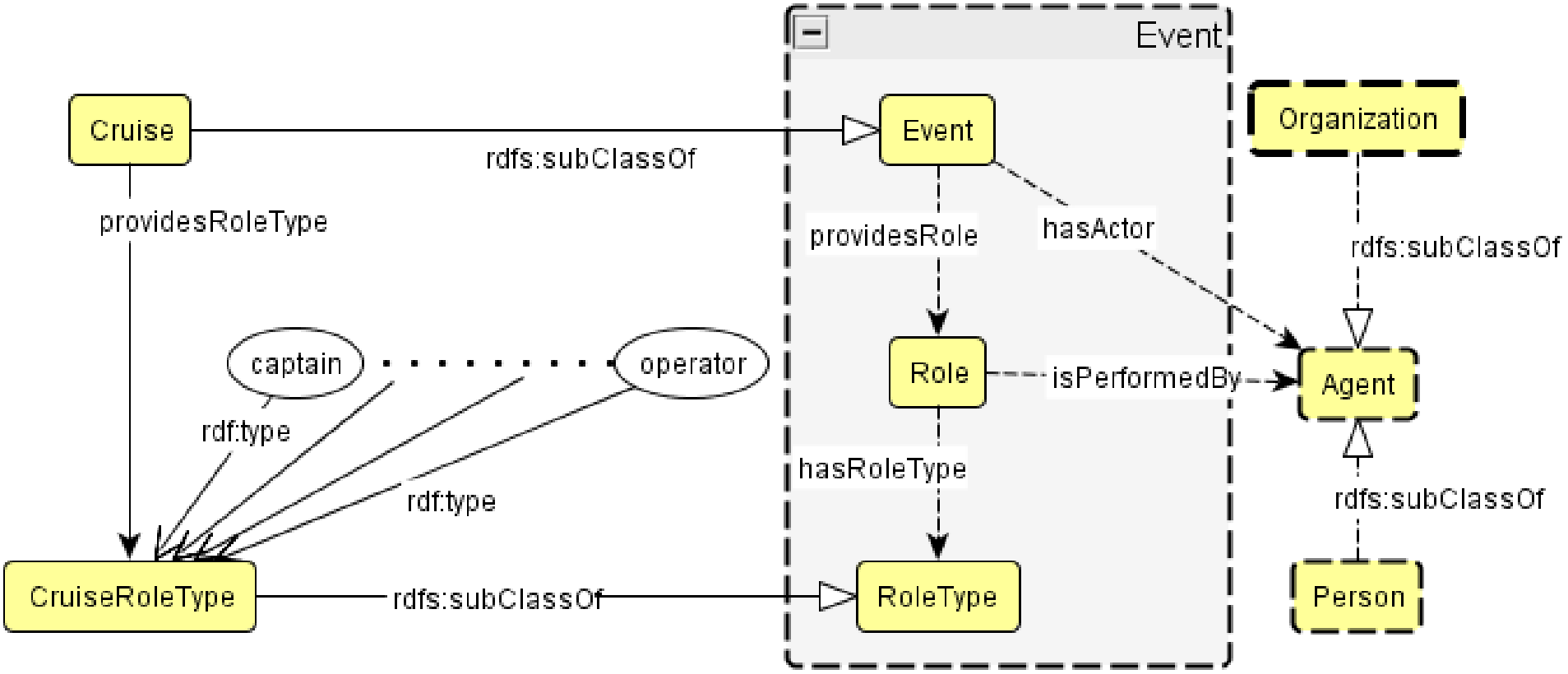
$$\mathsf{CruiseType} \equiv \{\texttt{operational}, \texttt{transit}, \texttt{maintenance}, \texttt{other\_cruisetype}\} \quad (24)$$

$$\mathsf{Cruise} \sqcap \exists \mathsf{isDescribedBy}.\exists \mathsf{hasCruiseType}.\{\texttt{operational}\} \equiv \exists \mathsf{providesRole}.(\mathsf{Role} \sqcap \exists \mathsf{hasRoleType}.\{\texttt{chief\_scientist}\}) \sqcap \exists \mathsf{isFundedBy}.\mathsf{FundingAward} \quad (25)$$

Event pattern



Roles (Cruise as Event)



Cruise as Event (OWL)

DaSe Lab

$$\mathsf{Role} \sqcap \exists \mathsf{providesRole}^{-}.\mathsf{Event} \sqsubseteq (=1\ \mathsf{hasRoleType}.\mathsf{RoleType}) \sqcap \exists \mathsf{isPerformedBy}.\mathsf{Agent} \quad (14)$$

$$\mathsf{providesRole} \circ \mathsf{isPerformedBy} \sqsubseteq \mathsf{hasActor} \quad (15)$$

$$\mathsf{Cruise} \sqsubseteq \mathsf{Event} \quad (16)$$

$$\mathsf{CruiseRoleType} \sqsubseteq \mathsf{RoleType} \quad (17)$$

$$\mathsf{CruiseRoleType}(x) \text{ for every role type } x \text{ in } (*) \quad (18\text{a-t})$$

$$R_{\mathsf{Cruise}} \circ \mathsf{owl{:}topObjectProperty} \circ R_{\mathsf{CruiseRoleType}} \sqsubseteq \mathsf{providesRoleType} \quad (19)$$

$$\mathsf{Cruise} \equiv \exists R_{\mathsf{Cruise}}.\mathsf{Self}, \quad (20)$$

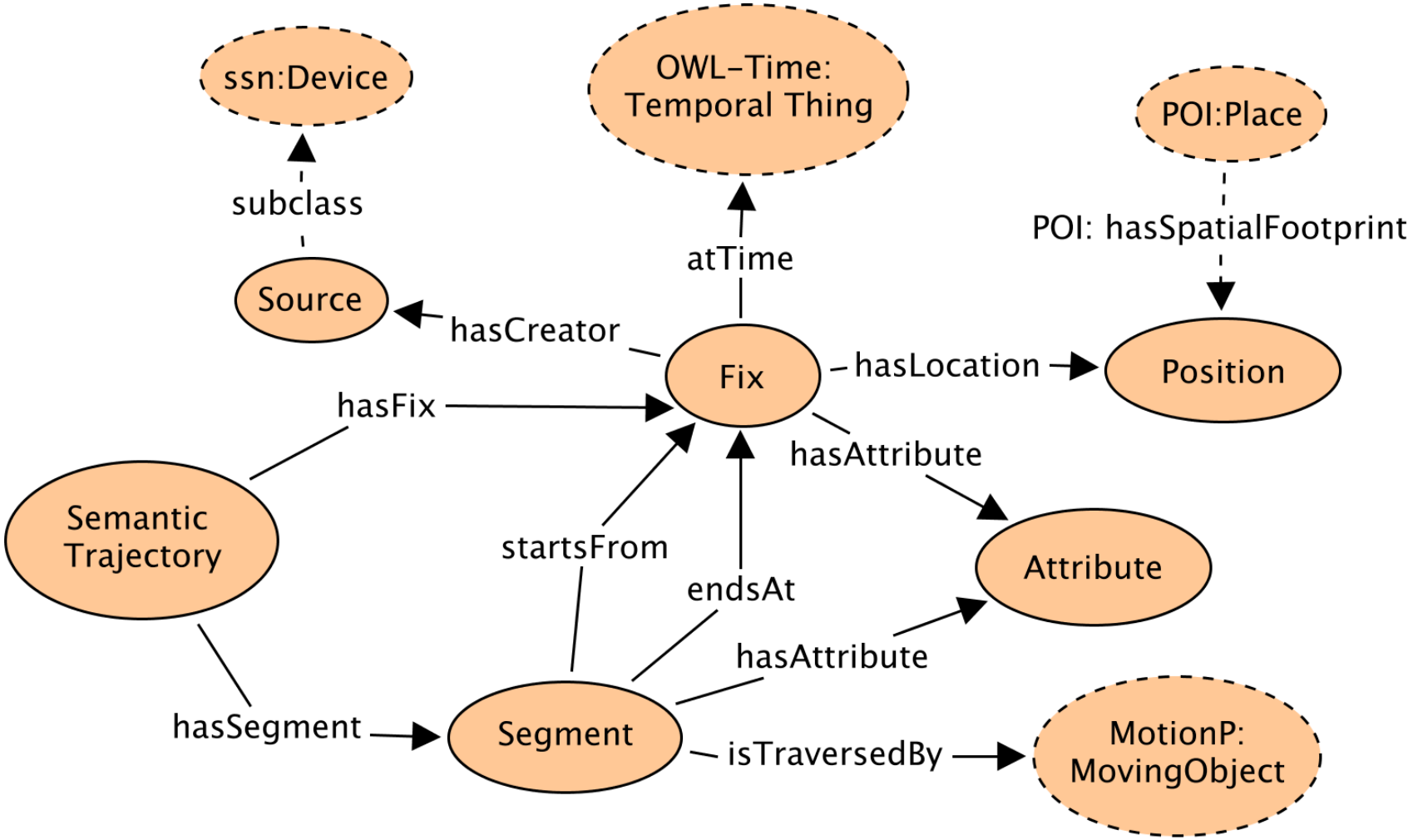
$$\mathsf{CruiseRoleType} \equiv \exists R_{\mathsf{CruiseRoleType}}.\mathsf{Self} \quad (21)$$

Cruise as Event (OWL)

DaSe Lab

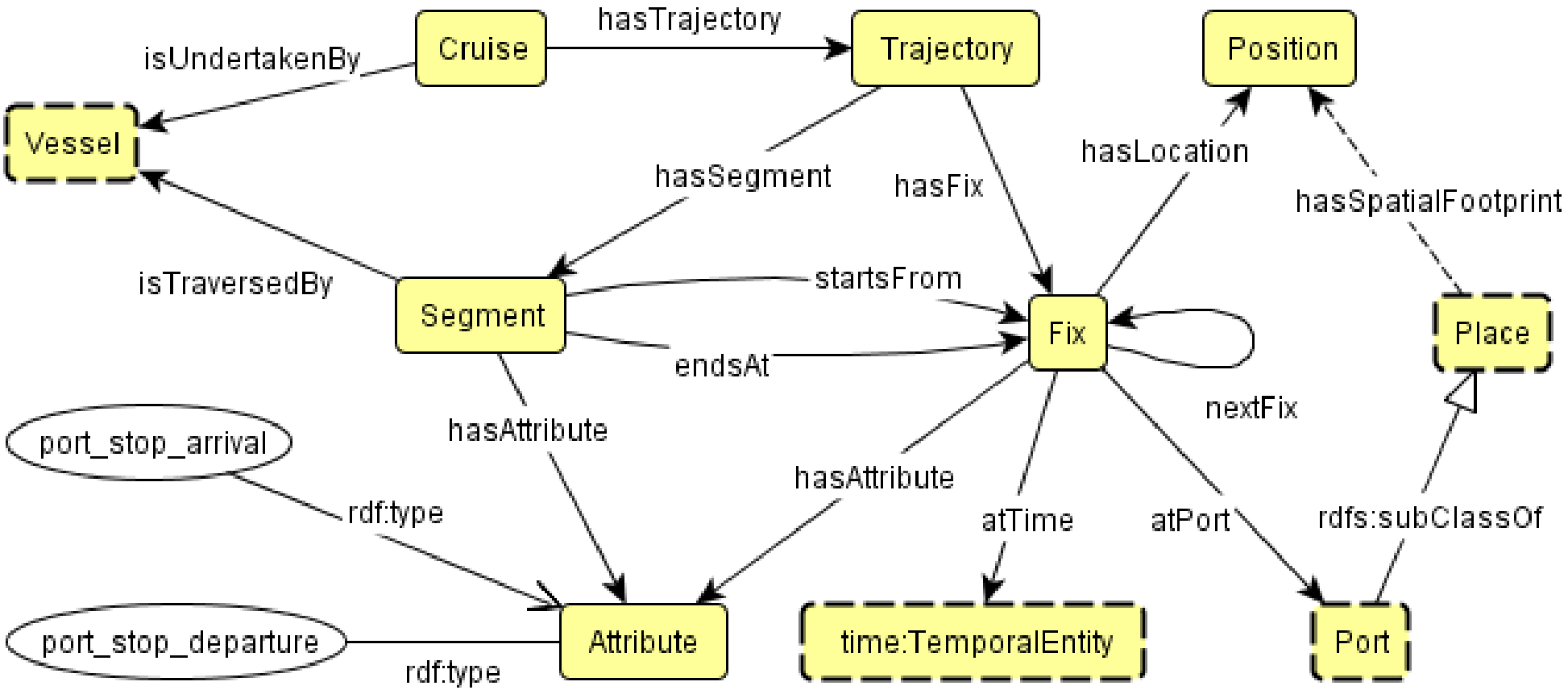
- Cruise role types:
  - captain,
  - chief engineer,
  - scientist,
  - chief scientist,
  - cochief scientist,
  - postdoc scientist,
  - student,
  - graduate student,
  - undergraduate student,
  - k12 student,
  - higher ed educator,
  - k12 educator,
  - technician,
  - marine technician,
  - lead marine technician,
  - inspector,
  - observer,
  - foreign observer,
  - other observer,
  - scheduler,
  - operator

Semantic Trajectory Pattern



Hu, Janowicz, Carral, Scheider, Kuhn, Berg-Cross, Hitzler, Dean, Kolas. COSIT 2013

Cruise Trajectory



Cruise Trajectory (OWL)

DaSe Lab

$$\mathsf{Cruise} \sqsubseteq (=1\ \mathsf{hasTrajectory}.\mathsf{Trajectory}) \quad (1)$$

$$\mathsf{Cruise} \sqsubseteq (=1\ \mathsf{isUndertakenBy}.\mathsf{Vessel}) \quad (2)$$

$$\mathsf{Fix} \sqsubseteq \exists \mathsf{atTime}.\mathsf{time{:}TemporalEntity} \sqcap \exists \mathsf{hasLocation}.\mathsf{Position} \sqcap (=1\ \mathsf{hasFix}^{-}.\mathsf{Trajectory}) \sqcap (\leqslant 1\ \mathsf{nextFix}.\mathsf{Fix}) \quad (3)$$

$$\mathsf{Segment} \sqsubseteq (=1\ \mathsf{startsFrom}.\mathsf{Fix}) \sqcap (=1\ \mathsf{endsAt}.\mathsf{Fix}) \sqcap \exists \mathsf{hasSegment}^{-}.\mathsf{Trajectory} \quad (4)$$

$$\exists \mathsf{nextFix}.\top \sqsubseteq (=1\ \mathsf{startsFrom}^{-}.\mathsf{Segment}) \quad (5)$$

$$\exists \mathsf{nextFix}^{-}.\top \sqsubseteq (=1\ \mathsf{endsAt}^{-}.\mathsf{Segment}) \quad (6)$$

$$\mathsf{startsFrom} \circ \mathsf{nextFix} \sqsubseteq \mathsf{endsAt} \quad (7)$$

Cruise Trajectory (OWL)

DaSe Lab

$$\mathsf{Port} \sqsubseteq \mathsf{Place} \tag{8}$$

$$\mathsf{Attribute}(\texttt{port\_stop\_arrival})$$

$$\mathsf{Attribute}(\texttt{port\_stop\_departure}) \tag{9a,b}$$

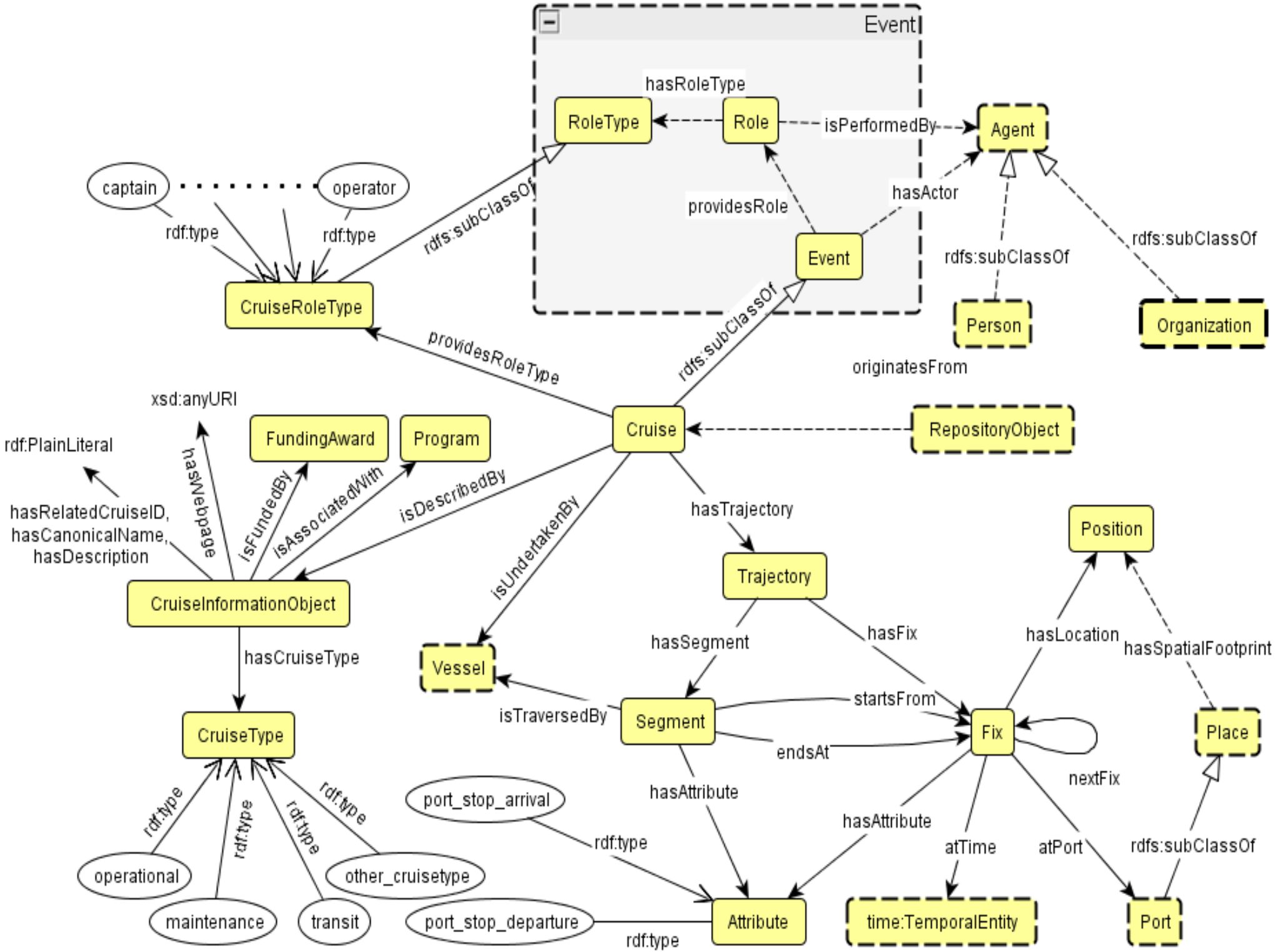
$$\mathsf{PortFix} \sqsubseteq \mathsf{Fix} \sqcap \exists \mathsf{atPort}.\mathsf{Port}$$

$$\exists \mathsf{hasAttribute}.\{\texttt{port\_stop\_arrival}\} \sqsubseteq \mathsf{PortFix} \tag{10}$$

$$\exists \mathsf{hasAttribute}.\{\texttt{port\_stop\_departure}\} \sqsubseteq \mathsf{PortFix} \tag{11}$$

$$\mathsf{atPort} \circ \mathsf{hasSpatialFootprint} \sqsubseteq \mathsf{hasLocation} \tag{12}$$

$$\mathsf{hasTrajectory} \circ \mathsf{hasSegment} \circ \mathsf{isTraversedBy} \sqsubseteq \mathsf{isUndertakenBy} \tag{13}$$


Event
hasRoleType
RoleType
Role
isPerformedBy
Agent
captain
operator
rdf:type
rdf:type
rdfs:subClassOf
providesRole
hasActor
Event
rdfs:subClassOf
rdfs:subClassOf
CruiseRoleType
Person
Organization
providesRoleType
rdfs:subClassOf
originatesFrom
xsd:anyURI
rdf:PlainLiteral
FundingAward
Program
Cruise
RepositoryObject
hasWebpage
isFundedBy
isAssociatedWith
isDescribedBy
isUndertakenBy
hasTrajectory
hasRelatedCruiseID,
hasCanonicalName,
hasDescription
Position
Trajectory
CruiseInformationObject
hasSegment
hasFix
hasLocation
hasSpatialFootprint
hasCruiseType
Vessel
isTraversedBy
Segment
startsFrom
Fix
Place
endsAt
nextFix
CruiseType
hasAttribute
hasAttribute
rdf:type
rdf:type
rdf:type
rdf:type
port_stop_arrival
rdf:type
atTime
atPort
rdfs:subClassOf
operational
maintenance
transit
other_cruisetype
port_stop_departure
rdf:type
Attribute
time:TemporalEntity
Port

Disjointness, Domain & Range

DaSe Lab

- Class disjointness asserted to pairs of classes, unless they are a subclass-superclass pair.
- Domain & Range use a guarded version:

$$\exists \text{hasFix}.\text{Fix} \sqsubseteq \text{Trajectory}, \text{Trajectory} \sqsubseteq \forall \text{hasFix}.\text{Fix} \quad (27)$$

$$\exists \text{hasRelatedCruiseID.rdf:PlainLiteral} \sqsubseteq \text{CruiseInformationObject} \quad (28)$$

$$\text{CruiseInformationObject} \sqsubseteq \forall \text{hasRelatedCruiseID.rdf:PlainLiteral} \quad (29)$$

Queries, Query Shortcuts, and Mappings

Queries

- Find all ports at which the researcher “Mak Saito” stopped by in any of his expeditions.

```
DESCRIBE ?port WHERE {
  ?port a :Port.
  ?cruise :hasTrajectory ?t ;
          :hasActor ?x.
  ?t :hasFix ?f.
  ?f :atPort ?port.
  ?x rdf:type :Person; :hasLegalName "Mak Saito". }
```

Queries

DaSe Lab

- Find out who joined any cruise that went through “Gulf of Maine”, what their role was in the cruise, and what funding award supported their trip.

```
SELECT ?name ?role ?fund WHERE {
  ?cruise :isDescribedBy ?d; :providesRole ?r;
          :hasFix ?x.
  ?d :isFundedBy ?f.
  ?f :hasAwardID ?fund.
  ?r :hasRoleType ?role; :isPerformedBy ?p.
  ?p rdf:type :Person; :hasLegalName ?name.
  ?x :hasLocation ?pos.
  ?pl :hasSpatialFootprint ?pos; rdfs:label ?pln.
  FILTER regex(?pln, "Gulf of Maine", "i").
```

Views (shortcuts for common queries)

DaSe Lab

$$\mathsf{Cruise}(x) \wedge \mathsf{providesRole}(x, y) \wedge \ \mathsf{isPerformedBy}(y, z) \wedge \mathsf{Person}(z) \wedge \mathsf{hasRoleType}(y, \texttt{chief\_scientist}) \rightarrow \mathsf{hasChiefScientist}(x, z) \quad (30)$$

$$\mathsf{Fix} \sqcap \neg \exists \mathsf{endsAt}^{-}.\mathsf{Segment} \sqsubseteq \mathsf{StartingFix} \quad (31)$$

$$\mathsf{Cruise}(x) \wedge \mathsf{hasTrajectory}(x, y) \wedge \mathsf{hasFix}(y, z) \wedge \mathsf{StartingFix}(z) \wedge \mathsf{atPort}(z, p) \rightarrow \mathsf{hasStartingPort}(x, p) \quad (32)$$

Mapping Rules (Examples)

DaSe Lab

For R2R

```
CONSTRUCT ?x rdf:type :Cruise
WHERE { ?x rdf:type r2r:Cruise. }
```

For BCO-DMO

```
CONSTRUCT ?x rdf:type :Cruise
WHERE { ?x a bcodmo:Deployment;
        bcodmo:ofPlatform [a bcodmo:Vessel]. }
```

AGU

Dr Peter Wiebe **may have** authored:

Year	Meeting	Section	Session	Abstract
2002	Ocean Sciences	OS	OS21P	OS21P-10

BCO-DMO

Peter Wiebe was found to have the following roles

Vessel	Cruise	Program	Role
Albatross IV	AL9404	Hydrography	Chief Scientist
Albatross IV	AL9508	Hydrography	Chief Scientist
Albatross IV	AL9906	Hydrography	Chief Scientist
Atlantis II	AT85	Cold Core Rings	Chief Scientist
Endeavor	EN261	Hydrography	Chief Scientist
Nathaniel B. Palmer	NBP0103	U.S. GLOBEC Southern Ocean	Chief Scientist
Nathaniel B. Palmer	NBP0104	U.S. GLOBEC Southern Ocean	Chief Scientist
Nathaniel B. Palmer	NBP0202	U.S. GLOBEC Southern Ocean	Chief Scientist
Nathaniel B. Palmer	NBP0204	U.S. GLOBEC Southern Ocean	Chief Scientist
Oceanus	OC275	Hydrography	Chief

Next Steps/Challenges

DaSe Lab

- Full-fledged implementation (current prototype: www.oceanlink.org)
- Evaluation (w.r.t. technical and social challenges)
- Tools for assisting pattern developments
  - Ease in extending the pattern collection to cover other repositories.
  - Interesting theoretical aspect: studying various ways of ontology reuse.
- Mappings
  - Abstraction may sometimes be more complex than the modeling on the data level, so simple query unfolding may not work (need to introduce blank nodes)
- Reasoning
  - Entailment in queries
  - Co-reference resolution
  - Integrity checking on data (missing or errorneous data)

OceanLink Collaborators

DaSe Lab

- Robert Arko – Lamont-Doherty Earth Observatory, Columbia University
- Suzanne Carbotte – Lamont-Doherty Earth Observatory, Columbia University
- Cynthia Chandler – Woods Hole Oceanographic Institution
- Michelle Cheatham – Wright State University
- Timothy Finin – University of Maryland, Baltimore County
- Pascal Hitzler – Wright State University
- Krzysztof Janowicz – University of California, Santa Barbara
- Adila A. Krisnadhi – Wright State University
- Thomas Narock – Marymount University
- Lisa Raymond – Woods Hole Oceanographic Institution
- Adam Shepherd – Woods Hole Oceanographic Institution
- Peter Wiebe – Woods Hole Oceanographic Institution

Acknowledgements

- The presented work is part of the NSF OceanLink project: “EAGER: Collaborative Research: EarthCube Building Blocks, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery.”

References

- Aldo Gangemi. Ontology design patterns for semantic web content. ISWC 2005
- Yingjie Hu, Krzysztof Janowicz, David Carral, Simon Scheider, Werner Kuhn, Gary Berg-Cross, Pascal Hitzler, Mike Dean, and Dave Kolas. A geo-ontology design pattern for semantic trajectories. COSIT 2013.
- Willem Robert van Hage, Veronique Malaise, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the Simple Event Model (SEM). JWS 9(2): 2011
- Daniel Oberle, Anupriya Ankolekar, Pascal Hitzler, Philipp Cimiano, Michael Sintek, Malte Kiesel, Babak Mougouie, Stephan Baumann, Shankar Vembu, Massimo Romanelli, Paul Buitelaar, Ralf Engel, Daniel Sonntag, Norbert Reithinger, Berenike Loos, Hans-Peter Zorn, Vanessa Micelli, Robert Porzel, Christian Schmidt, Moritz Weiten, Felix Burkhardt, and Jianshen Zhou. DOLCE ergo SUMO: On Foundational and Domain Models in the SmartWeb Integrated Ontology (SWIntO). JWS 5(3): 2007

Thanks!