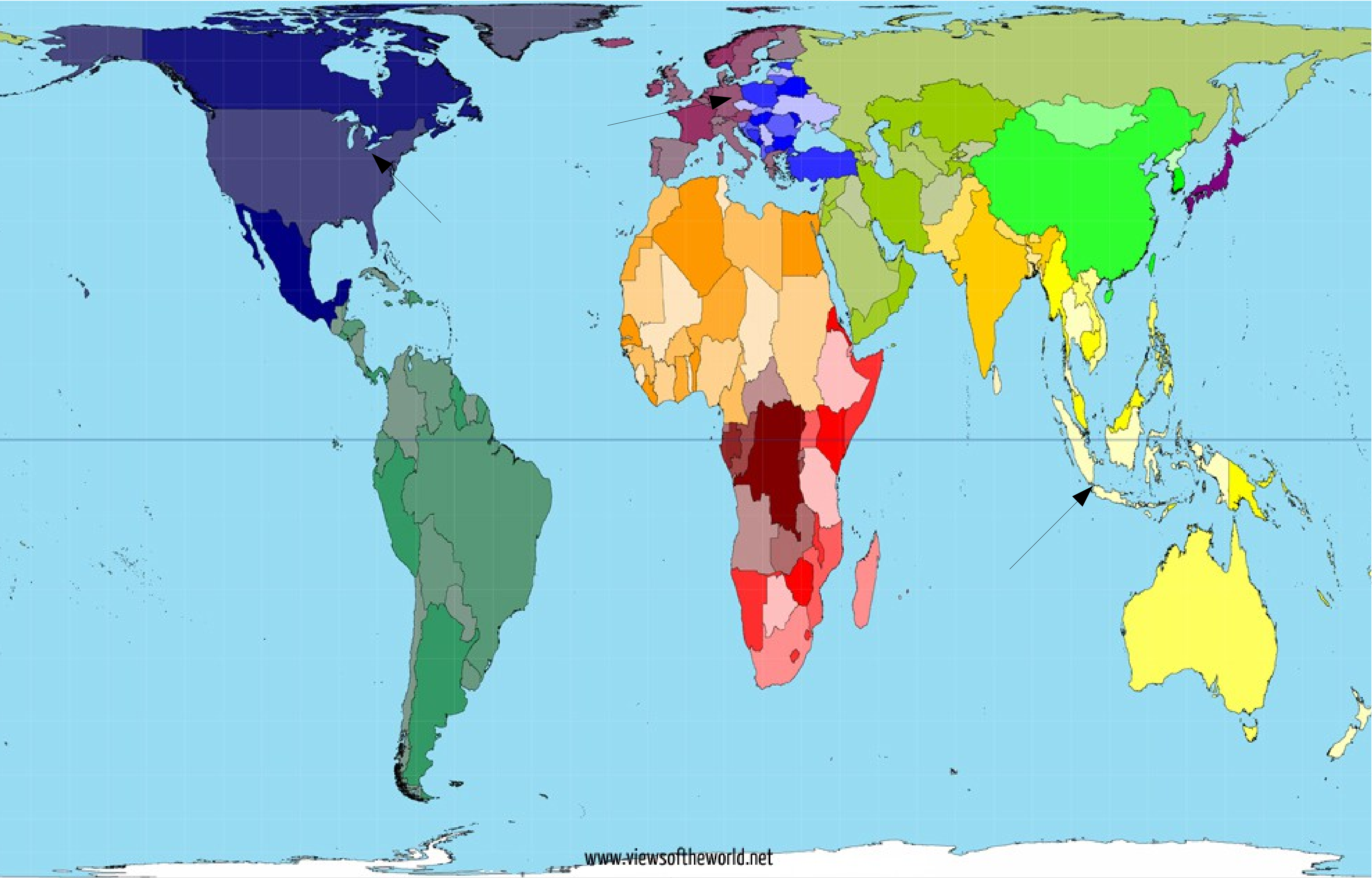
Cross-Repository Data Integration using Ontology Design Patterns

Adila Alfa Krisnadhi
DaSe Lab for Data Semantics
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

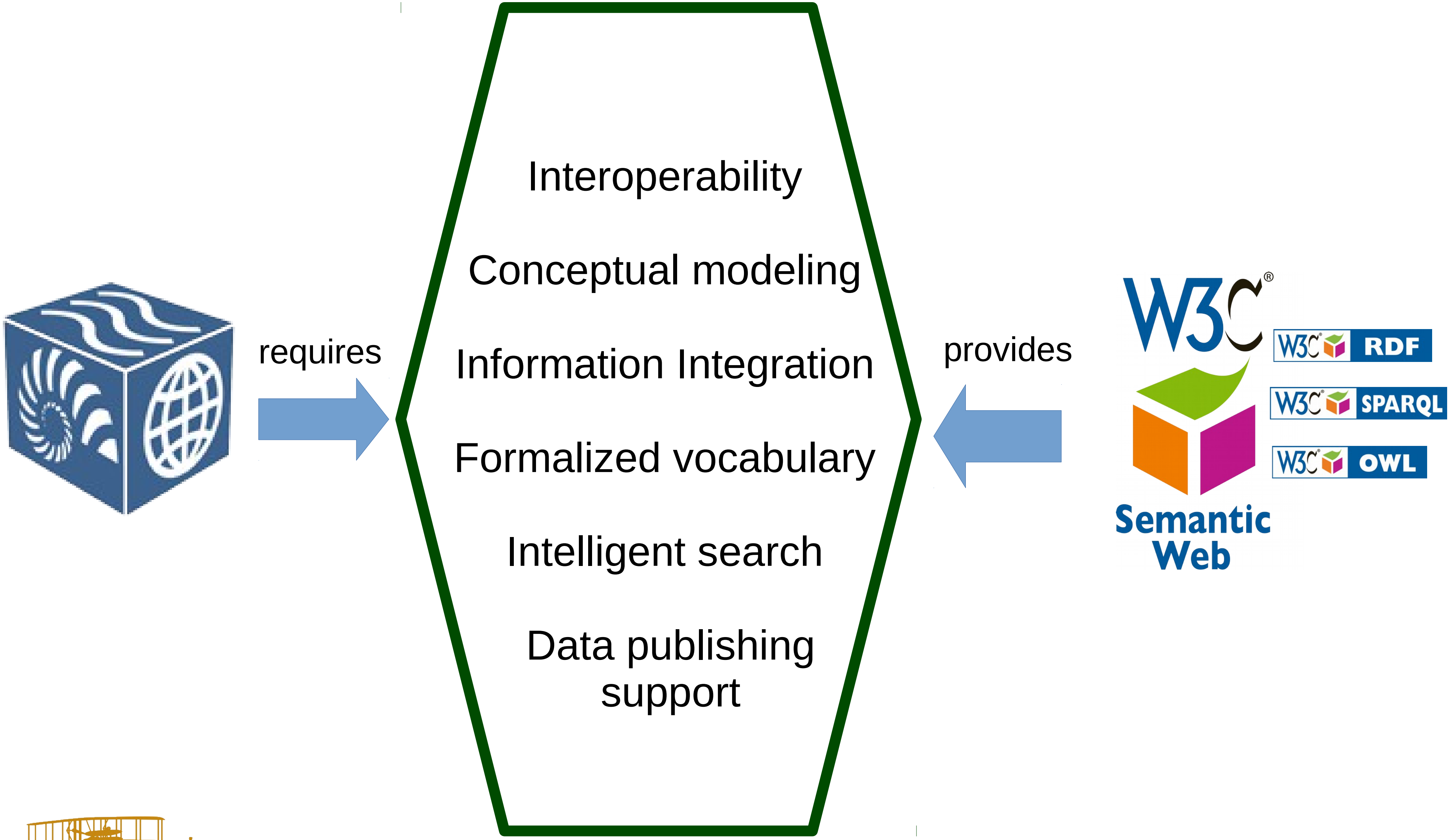
What is this about?

- 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.

EarthCube

- “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

- 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)

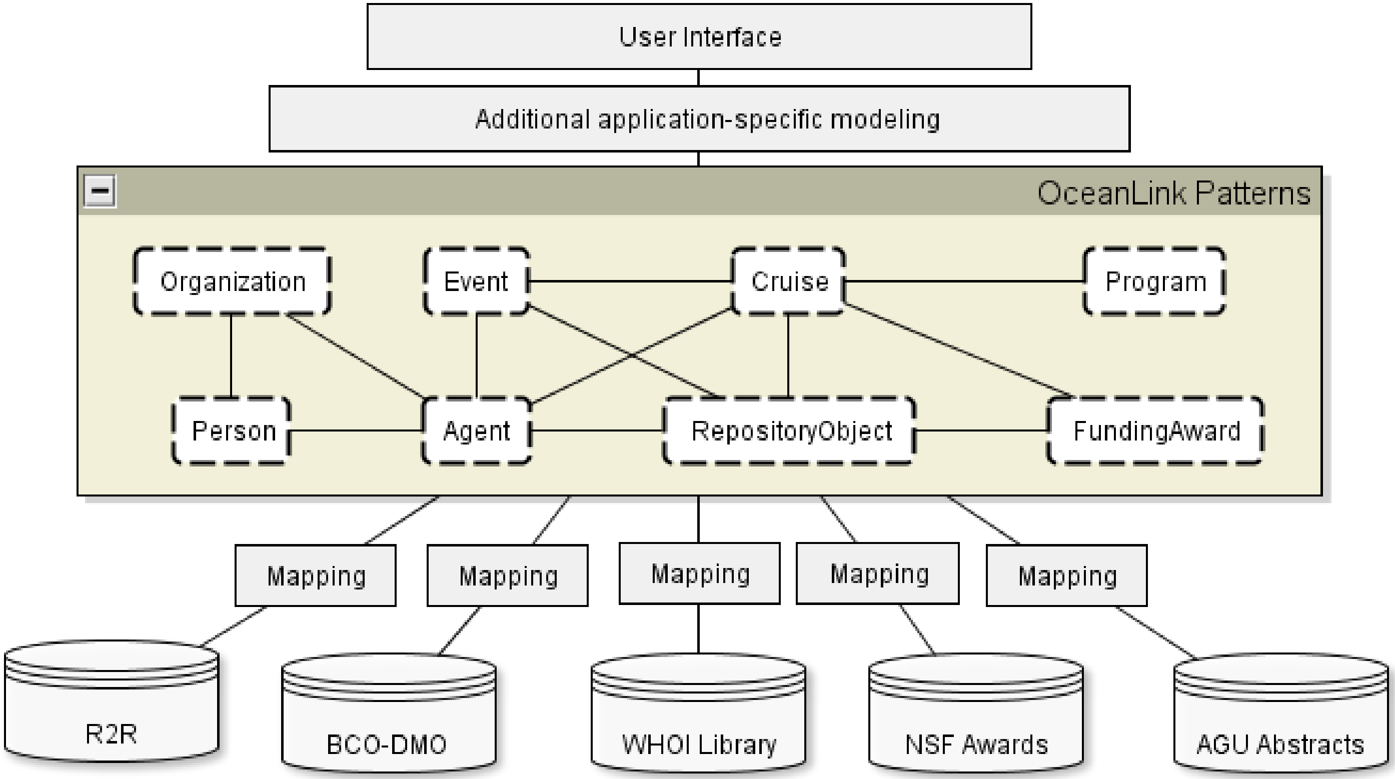
- Oceanographic data – BCO-DMO: >6000 datasets with supporting documents from 24 programs, 229 projects, 1673 deployments
- Field expeditions data – R2R: 400 expeditions per year; 3
- Conference and funded award abstracts – AGU: 30 mil. triples
- Theses, reports, journal articles – MBLWHOI Library: 5500 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

- **Technical challenge**:
  - Lack of interoperability in terms of formats, etc.
  - Semantic 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.

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).



Modular Ontology Engineering

- Model one key notion at a time
- Keep ontological commitments minimum
- 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 people can test the resulting patterns against real data

Ontology Design Patterns

- 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 Patterns

DaSe Lab

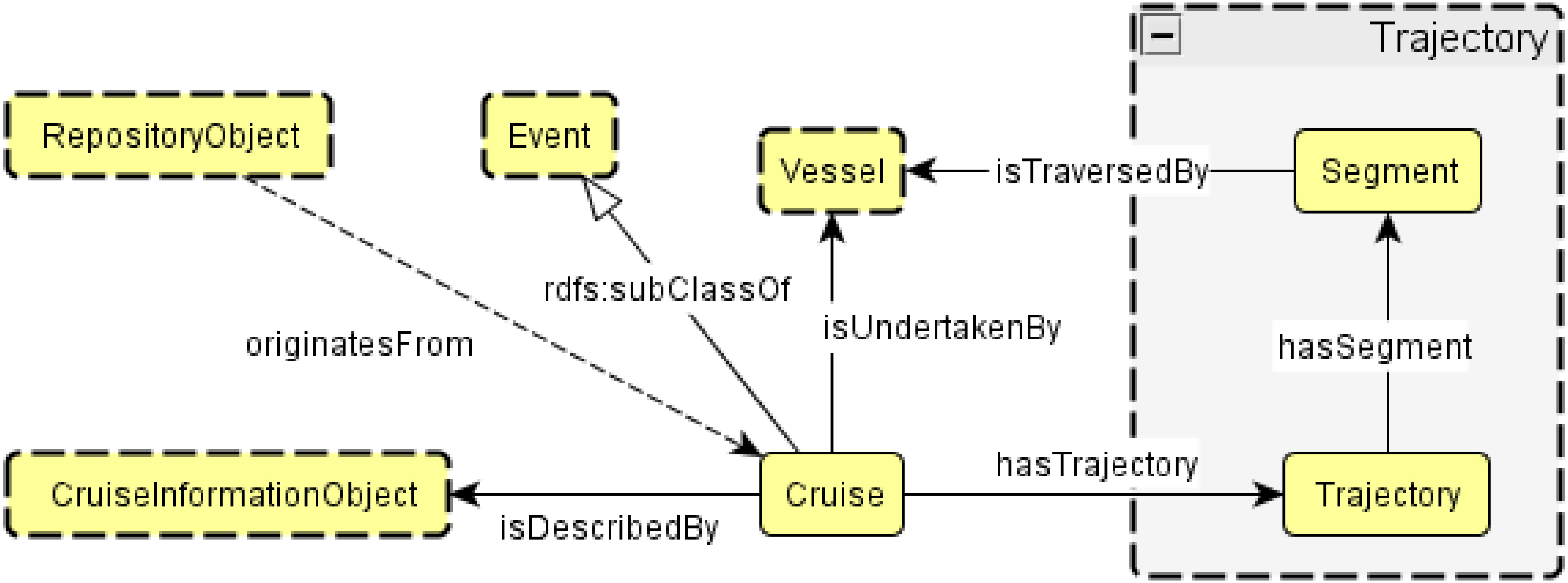
- Cruise
- Vessel
- Trajectory
- Person
- Organization
- Roles of Agents
- Repository Object
- Dataset
- and a few other patterns (about 15 in total)

We are not starting from zero, of course.

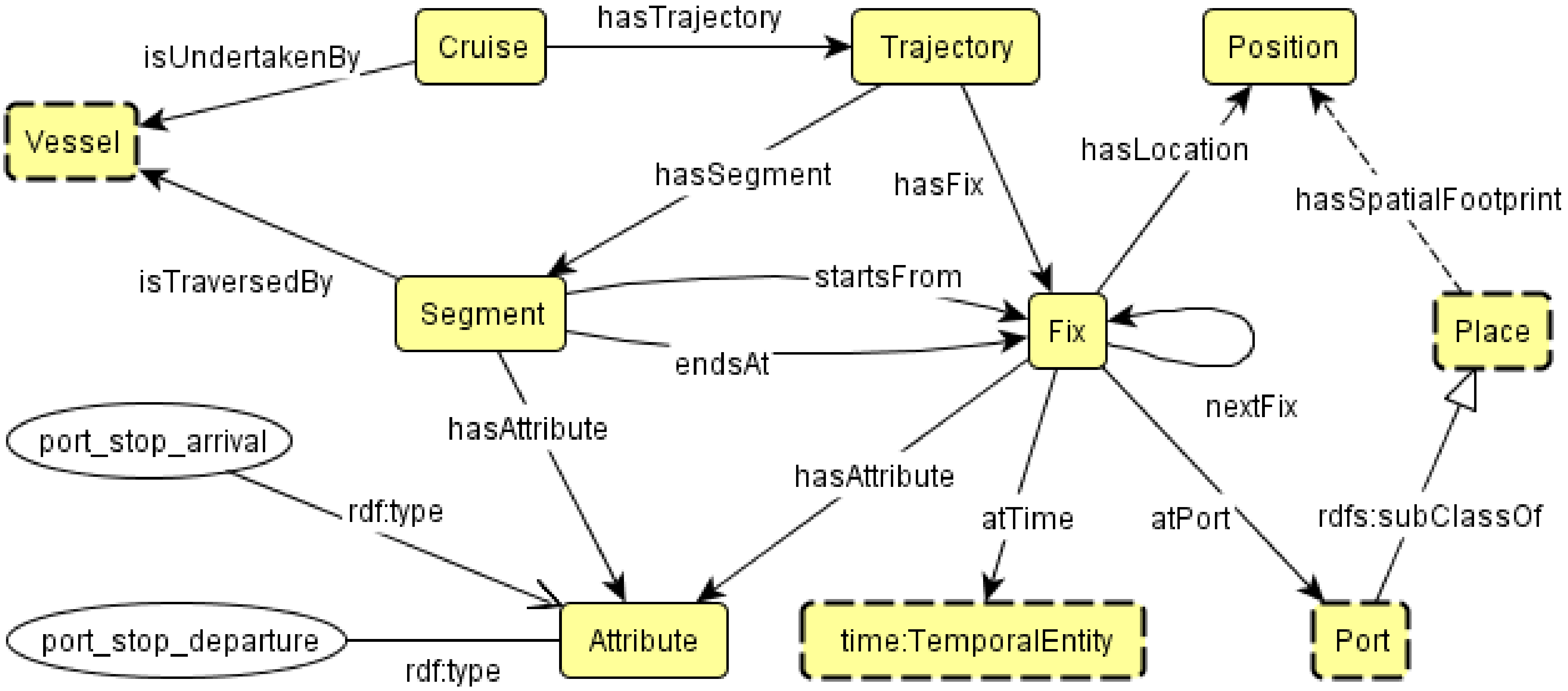
Cruise Pattern: Competency Questions

- Find all cruises passing through Gulf of Maine in August 2013.
- Show the trajectories of cruises in operation in September 2013.
- List all cruise vessels that departed from Woods Hole in 2012.
- Find the chief scientists of any cruise that collected samples of carbon-isotope data in Lake Superior.
- What datasets were produced by the cruise AE0901?
- Which cruises are funded by the NSF award DBI-0424599?
- List all cruises under the Ocean Flux Program.

Cruise Pattern

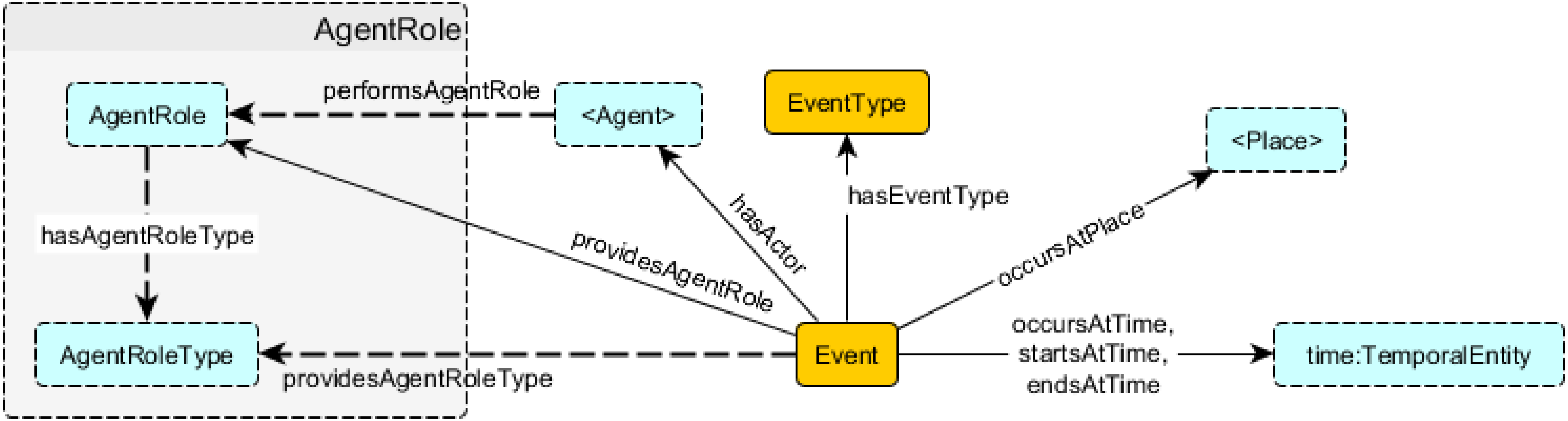


Cruise Trajectory

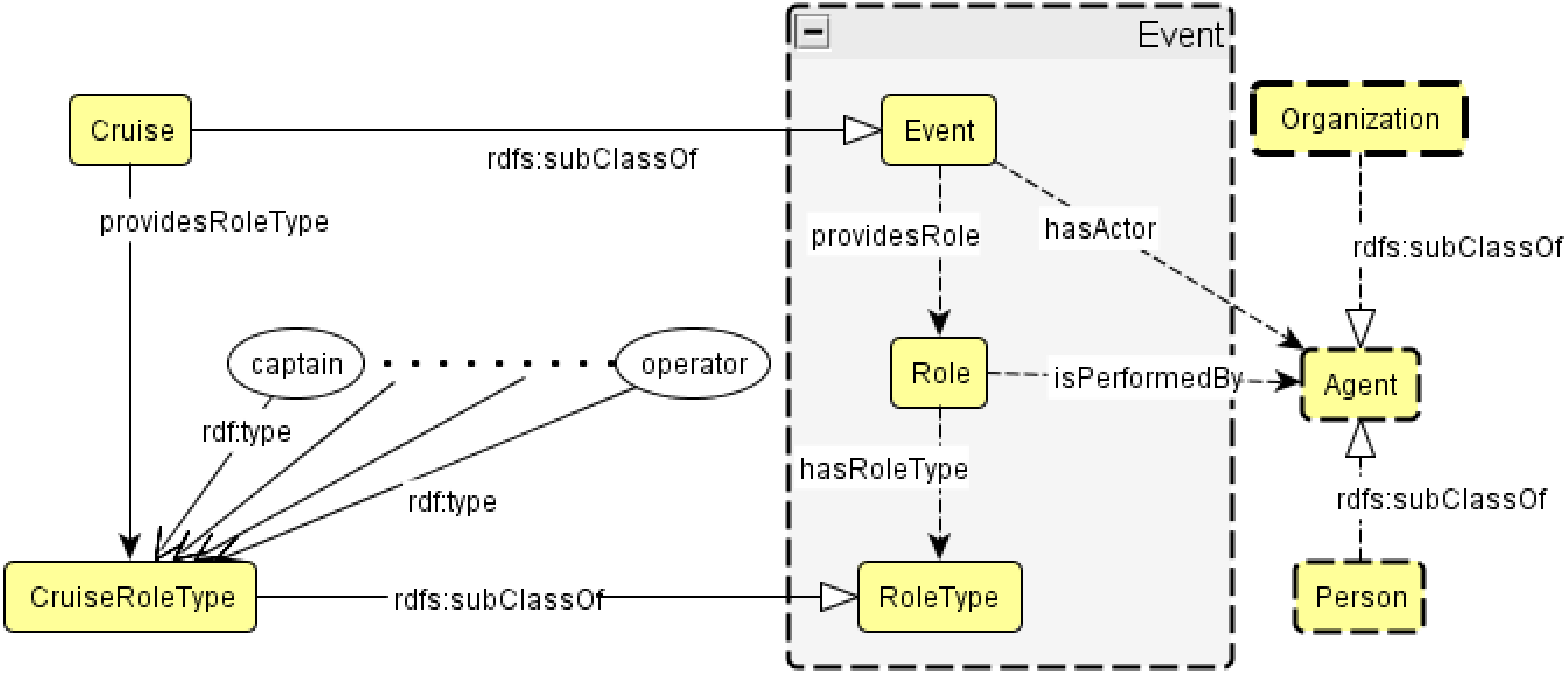


This reuses "Semantic Trajectory" pattern from Hu et al. COSIT 2013

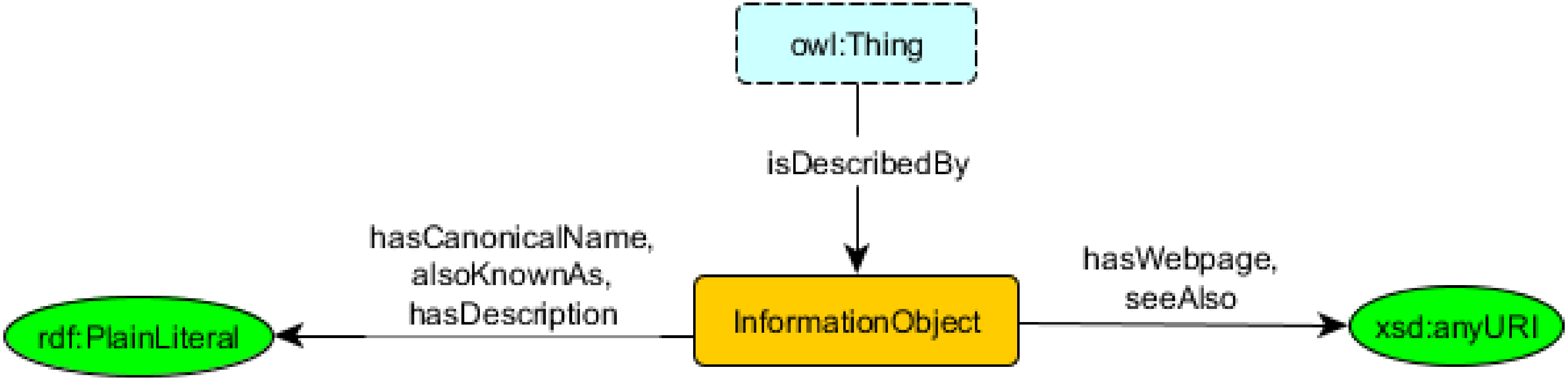
Event pattern



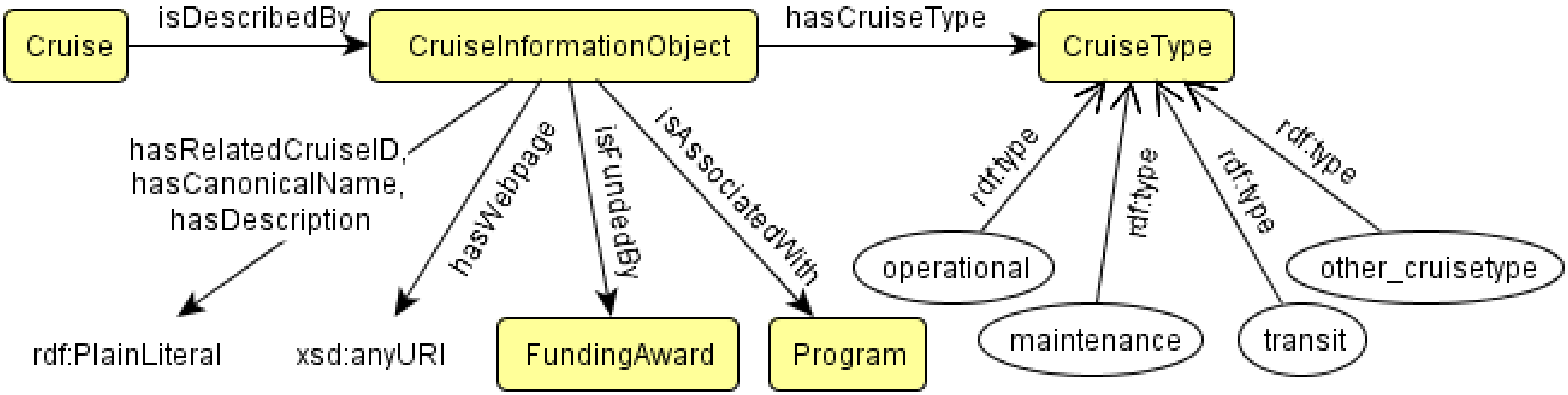
Roles (Cruise as Event)

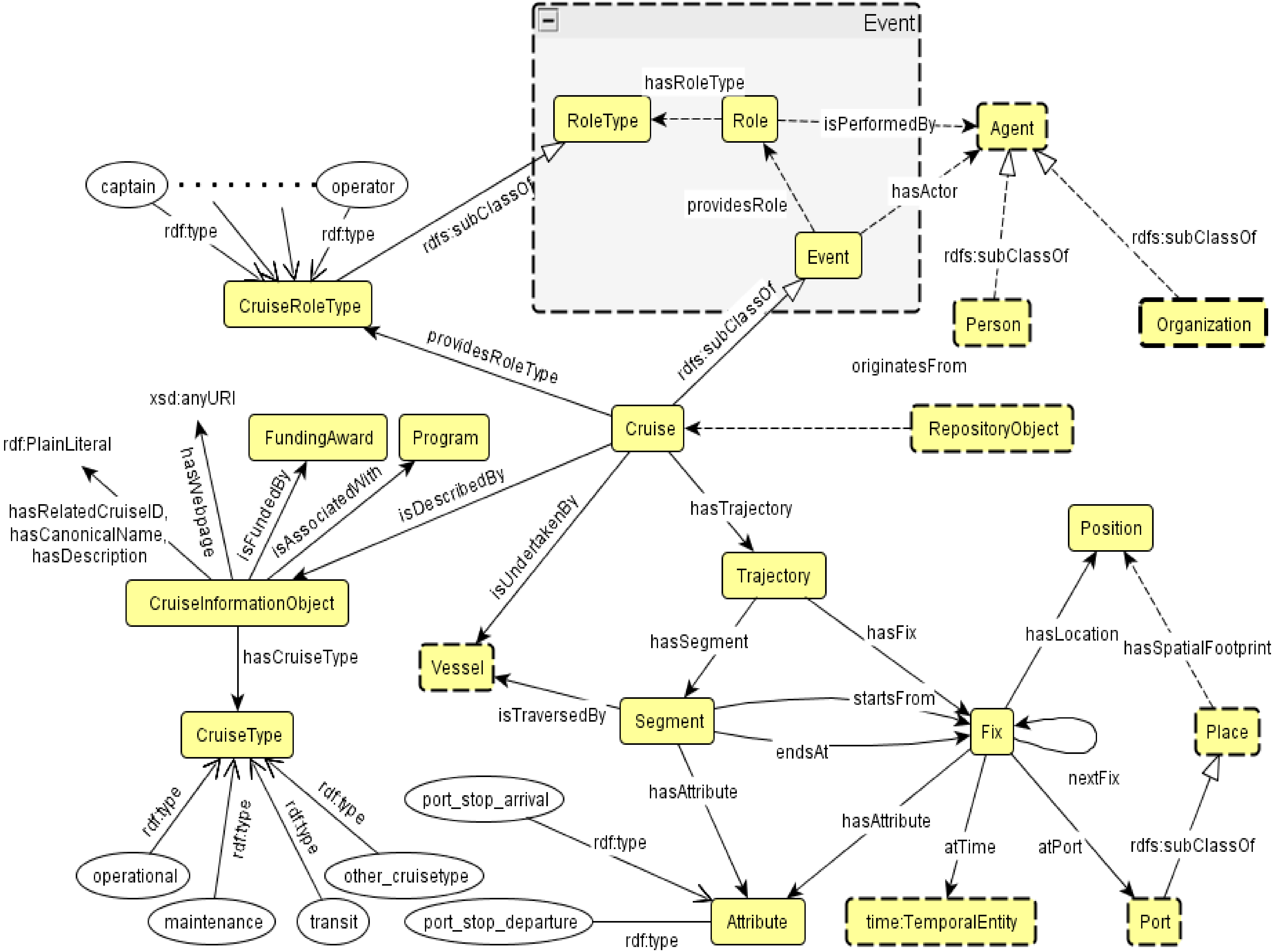


InformationObject pattern



Cruise Information Object




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

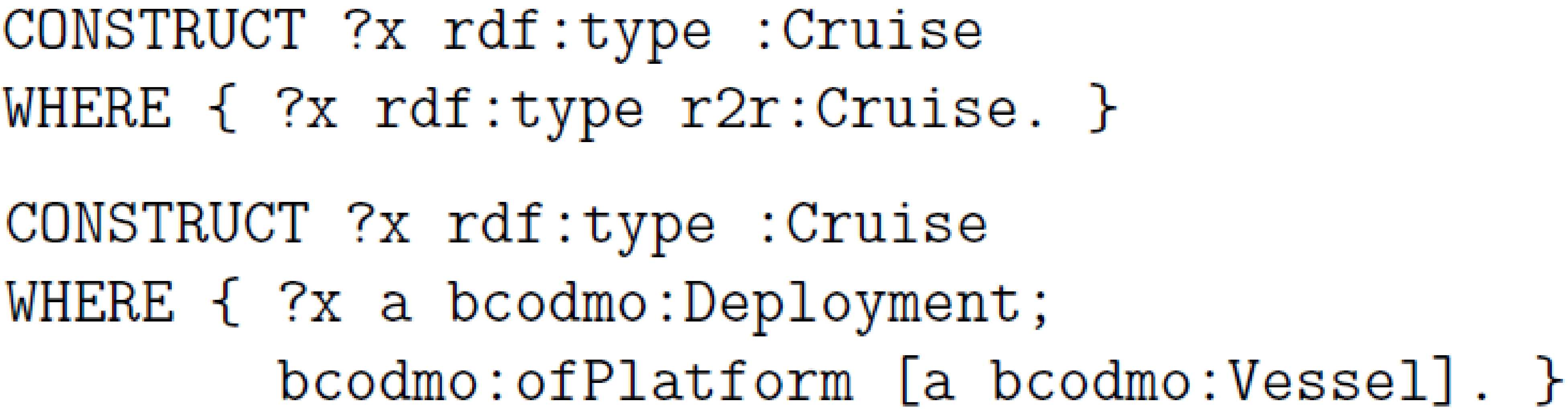
Views (shortcuts for common queries)

$$\begin{aligned}&\mathsf{Cruise}(x) \wedge \mathsf{providesRole}(x,y) \wedge \ \mathsf{isPerformedBy}(y,z)\\ &\wedge \mathsf{Person}(z) \wedge \mathsf{hasRoleType}(y, \mathtt{chief\_scientist})\\ &\qquad\qquad \rightarrow \mathsf{hasChiefScientist}(x,z) \qquad (30)\end{aligned}$$

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

$$\begin{aligned}&\mathsf{Cruise}(x) \wedge \mathsf{hasTrajectory}(x,y) \wedge \mathsf{hasFix}(y,z) \wedge \mathsf{StartingFix}(z)\\ &\qquad \wedge \mathsf{atPort}(z,p) \rightarrow \mathsf{hasStartingPort}(x,p) \qquad (32)\end{aligned}$$

Mappings (one way to do it)

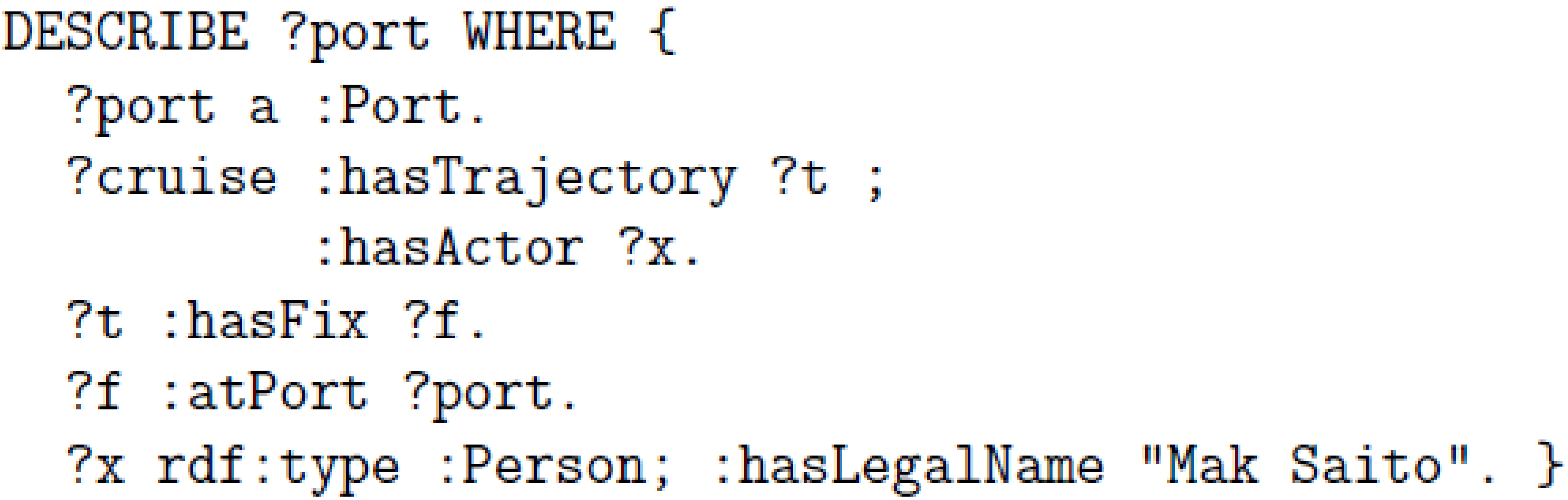


```
CONSTRUCT ?x rdf:type :Cruise
WHERE { ?x rdf:type r2r:Cruise. }

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

Queries

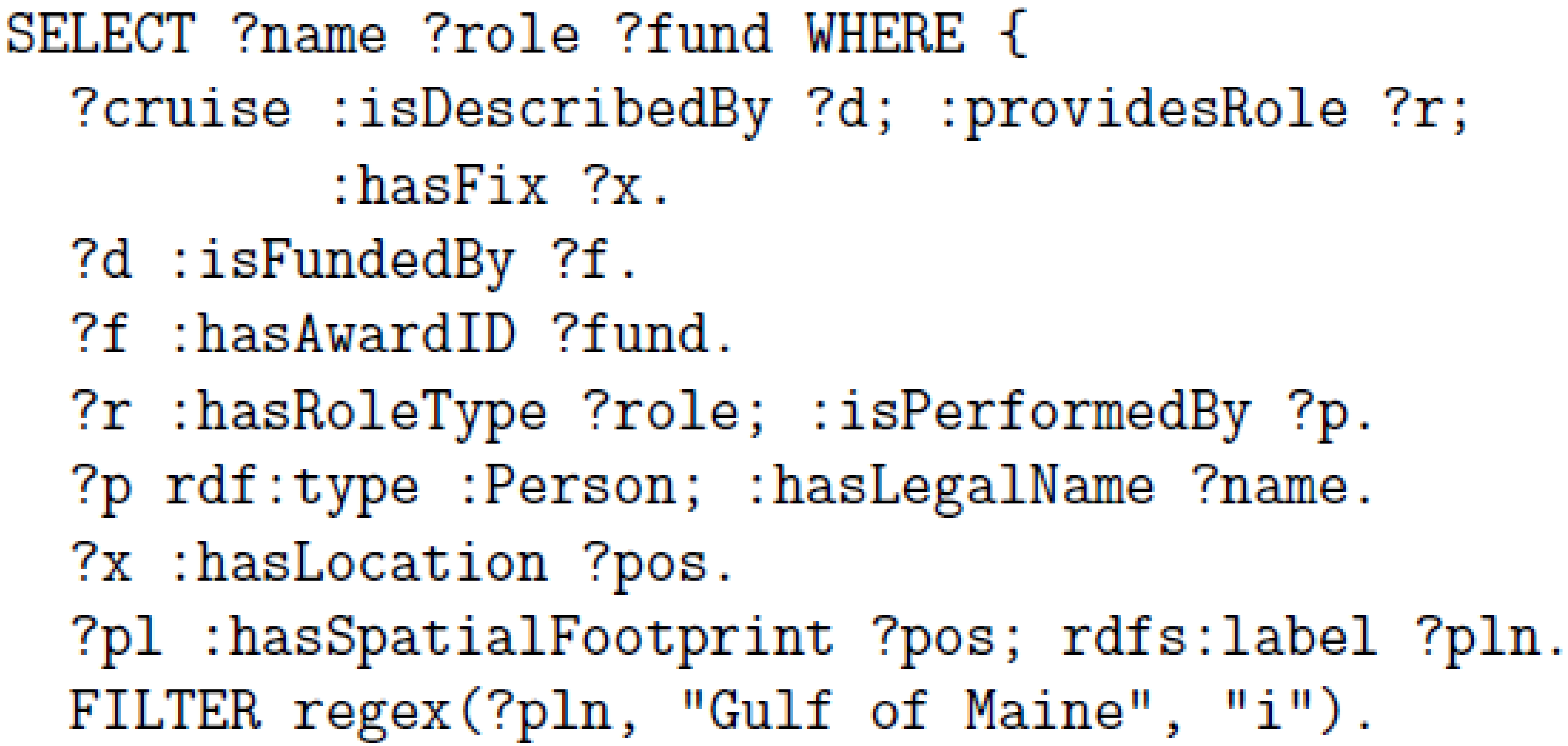
- Find all ports at which the researcher “Mak Saito” stopped by in any of his expeditions.



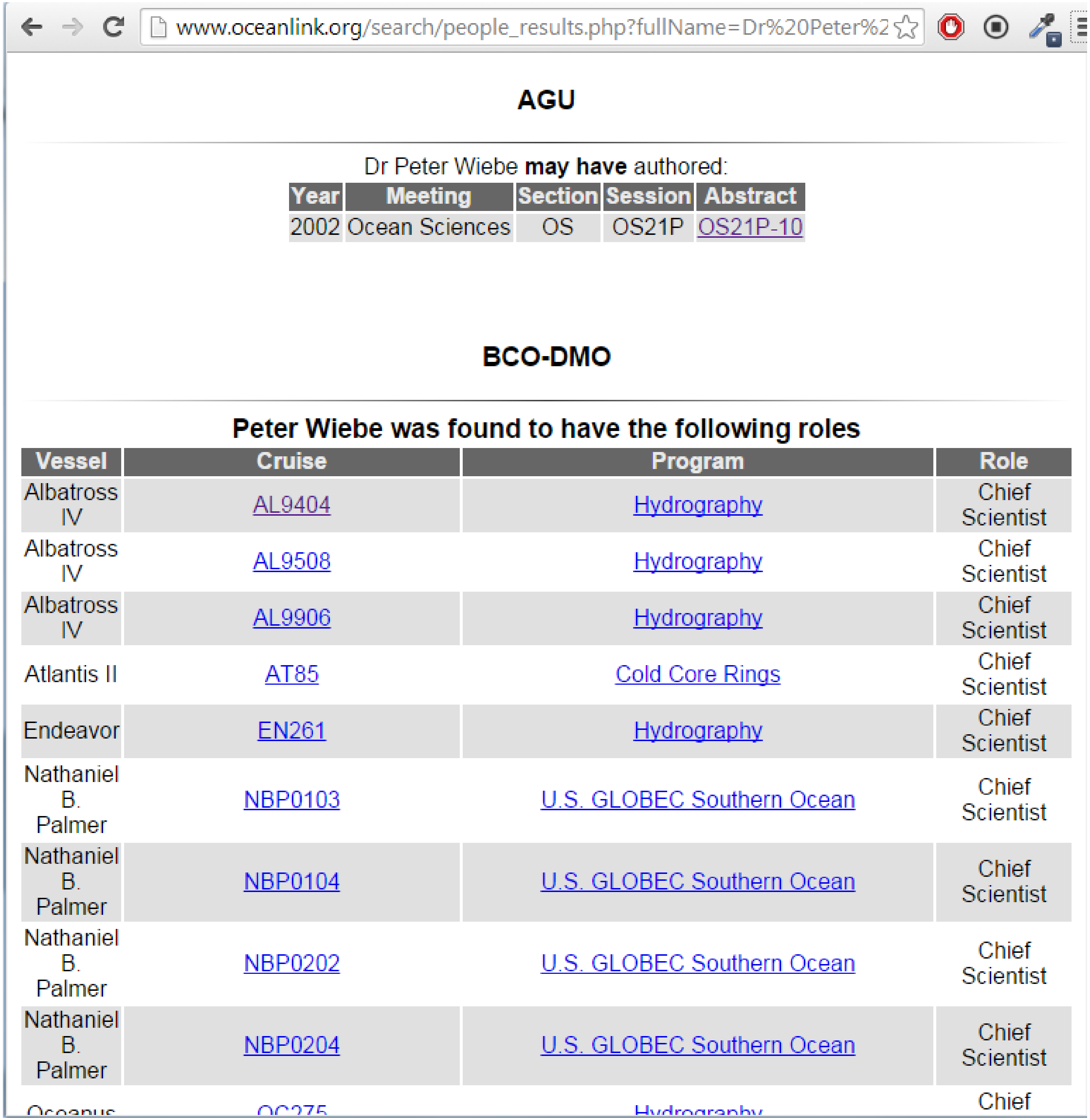
```
DESCRIBE ?port WHERE {
  ?port a :Port.
  ?cruise :hasTrajectory ?t ;
          :hasActor ?x.
  ?t :hasFix ?f.
  ?f :atPort ?port.
  ?x rdf:type :Person; :hasLegalName "Mak Saito". }
```

Queries

- 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").
```



www.oceanlink.org/search/people_results.php?fullName=Dr%20Peter%2

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

- Evaluation
  - Does the pattern approach succeed in meeting both the 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.
- Data-to-Pattern Mappings
  - Abstraction may sometimes be more complex than the modeling on the data level, so simple query unfolding may not work.
- Reasoning
  - Entailment in queries
  - Integrity checking on data (missing or errorneous data)

OceanLink Collaborators

- 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

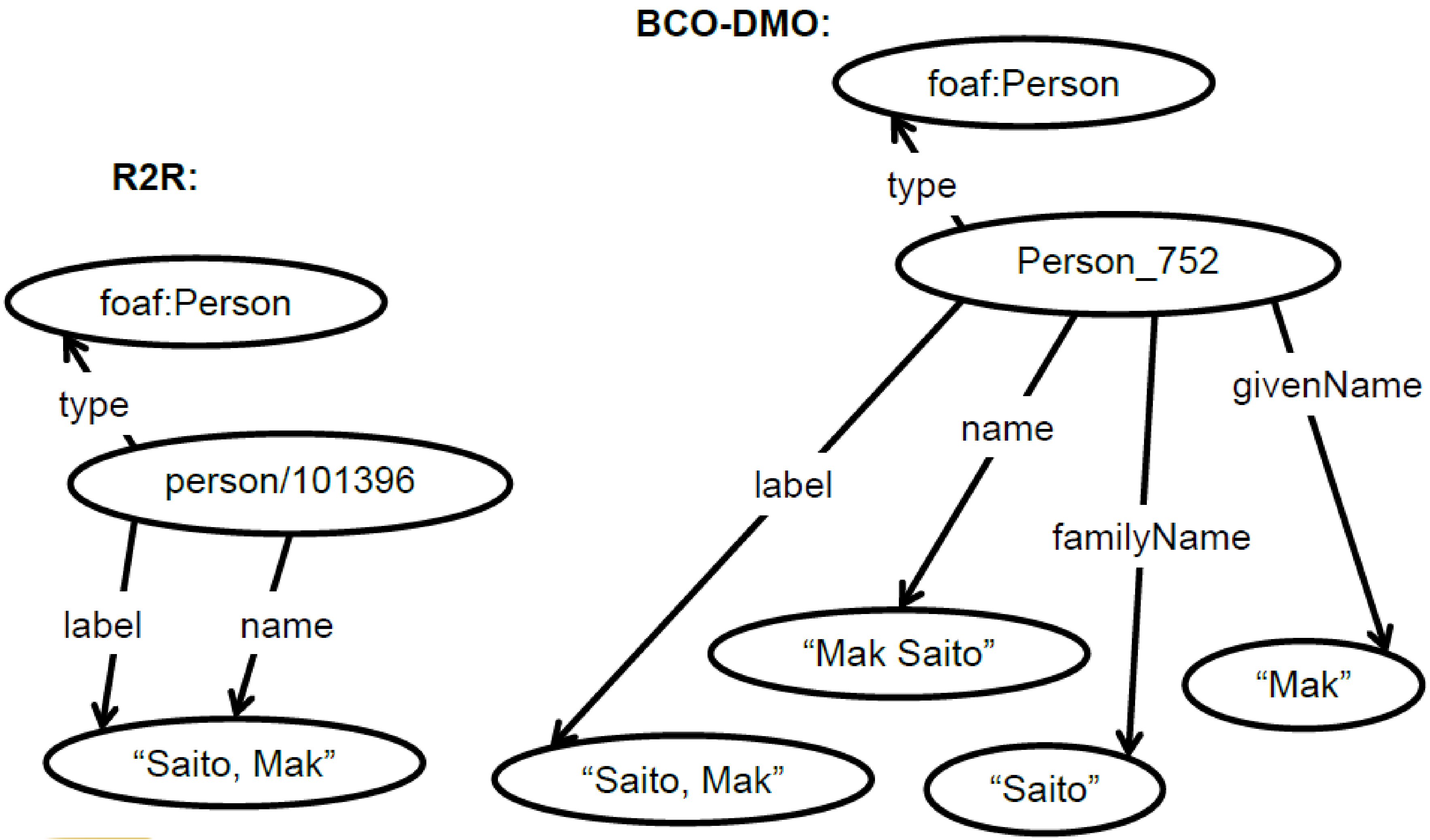
- The presented work is part of the NSF OceanLink project: “EAGER: Collaborative Research: EarthCube Building Blocks, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery.”

Thanks!

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

Information representation choices



BCO-DMO:

foaf:Person

type

Person_752

label

name

familyName

givenName

“Mak Saito”

“Mak”

“Saito, Mak”

“Saito”

R2R:

foaf:Person

type

person/101396

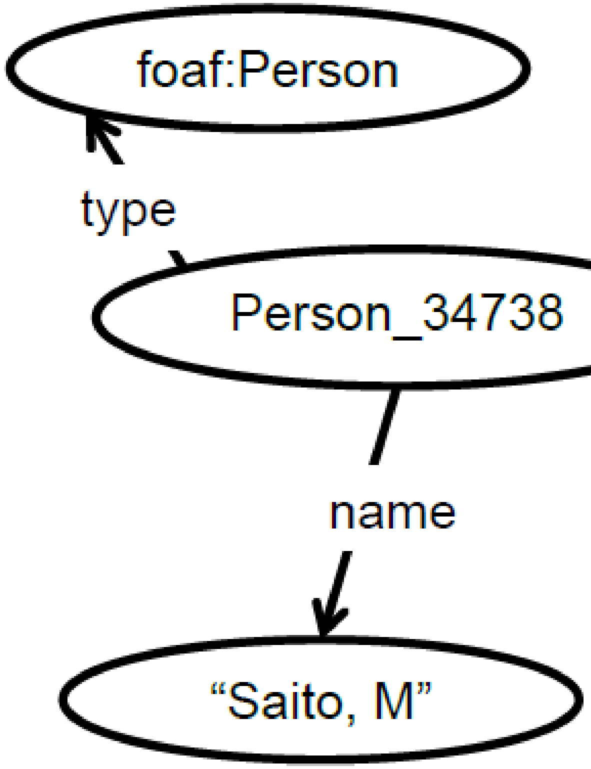
label

name

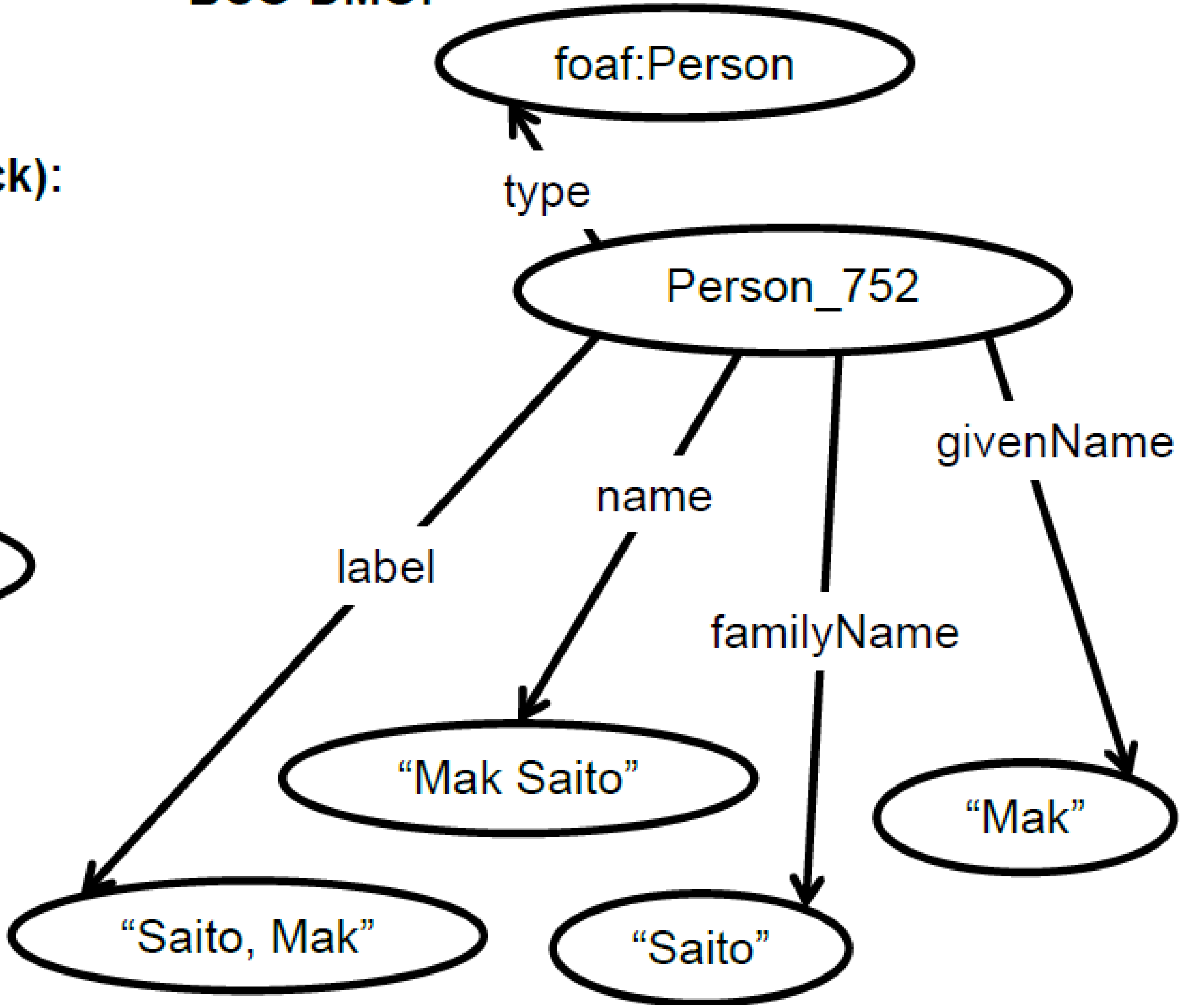
“Saito, Mak”

Information representation choices

AGU Abstracts (Tom Narock):

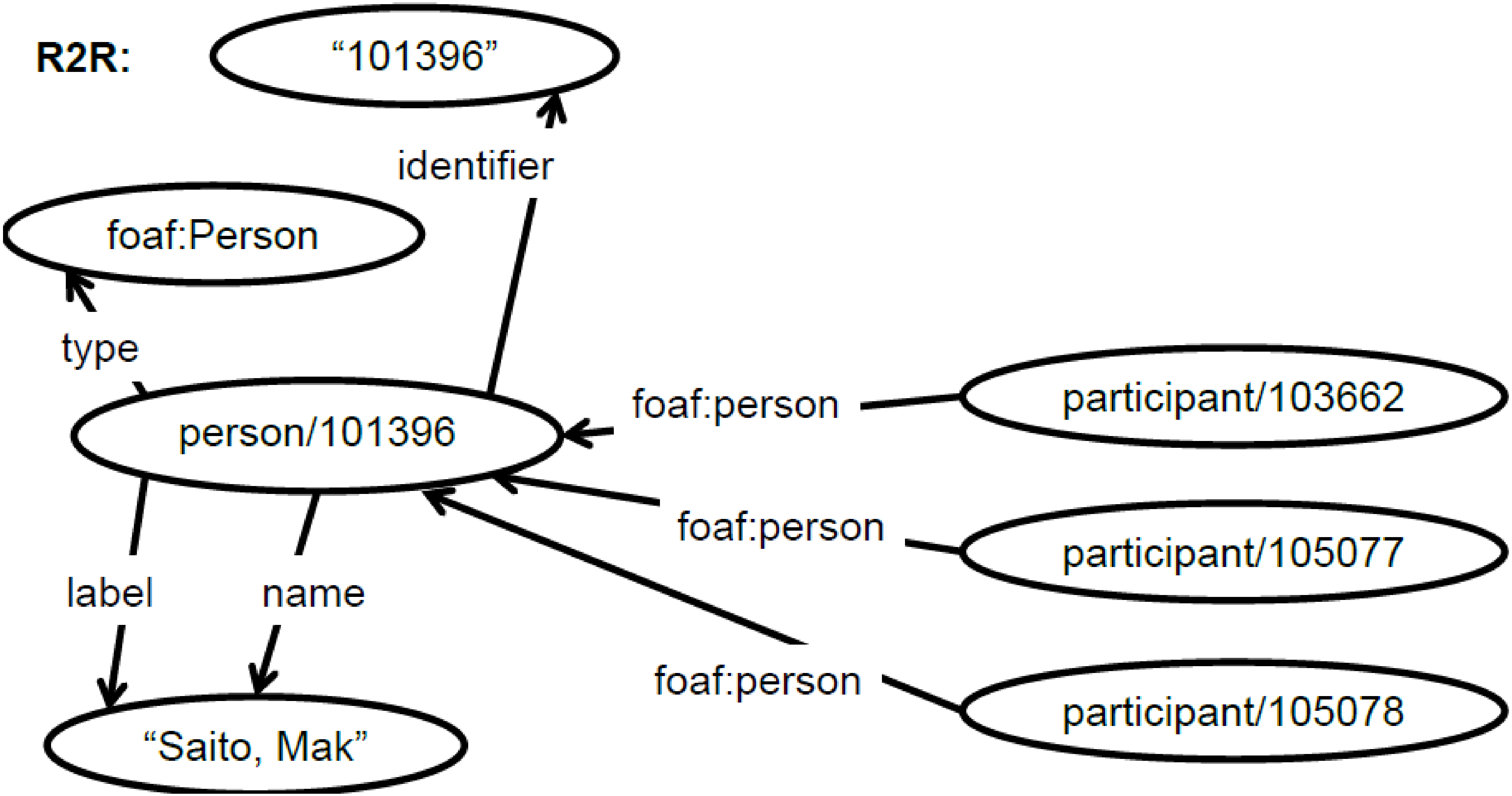


BCO-DMO:

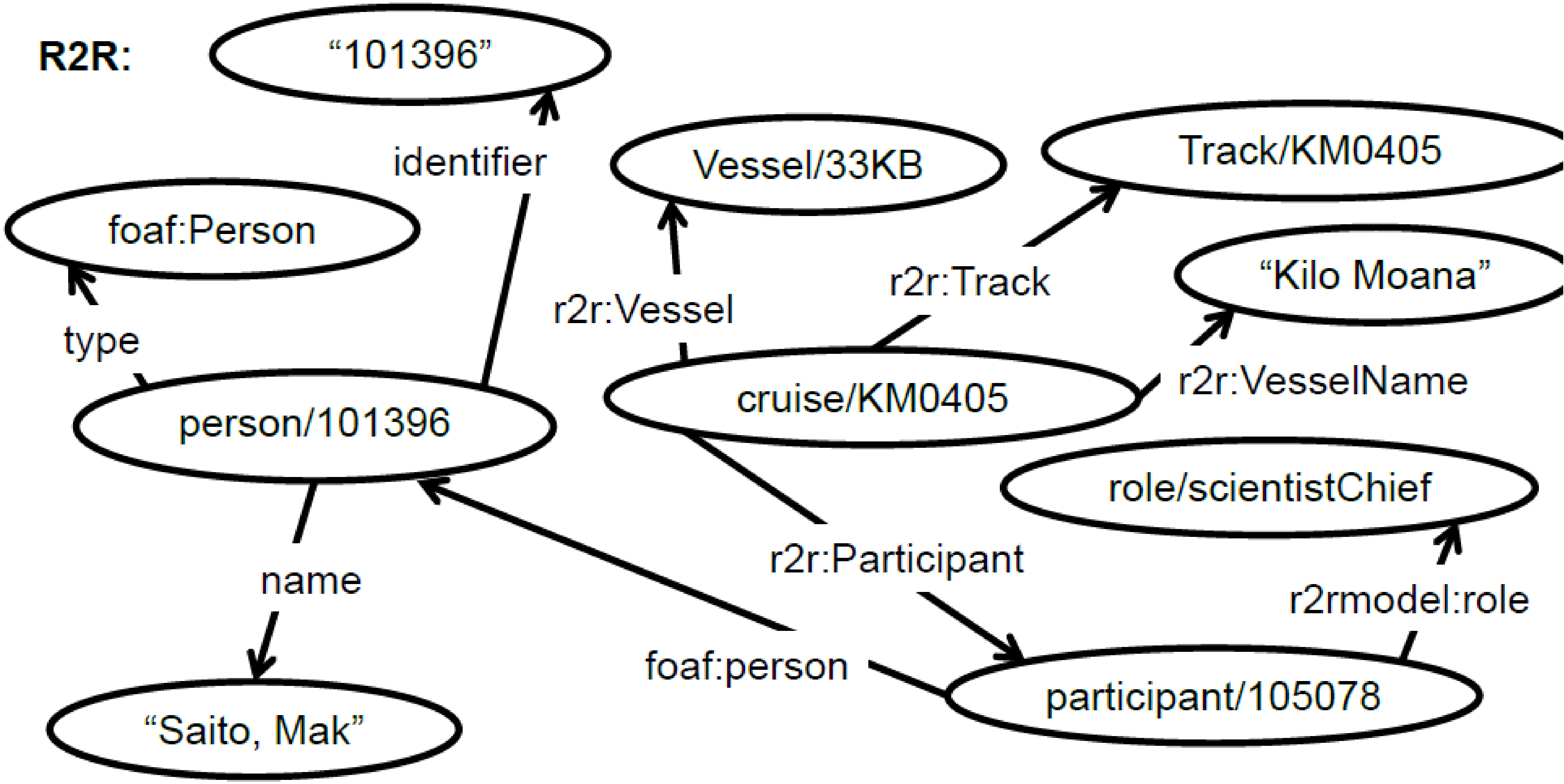


Information Representation Choices

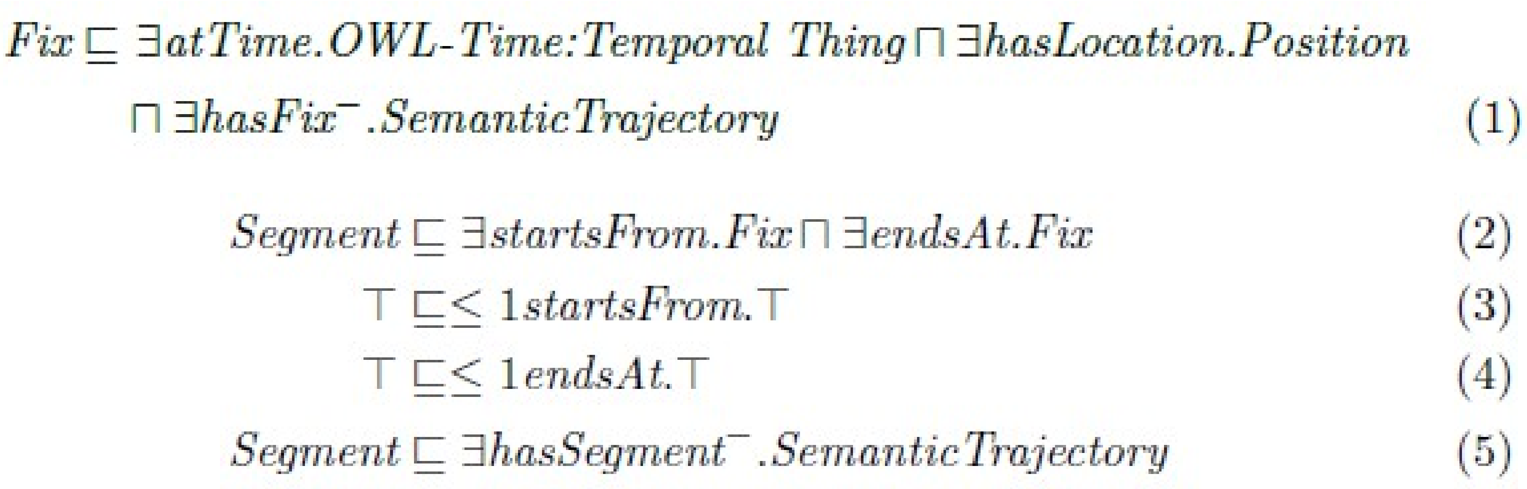
R2R:



Information Representation Choices

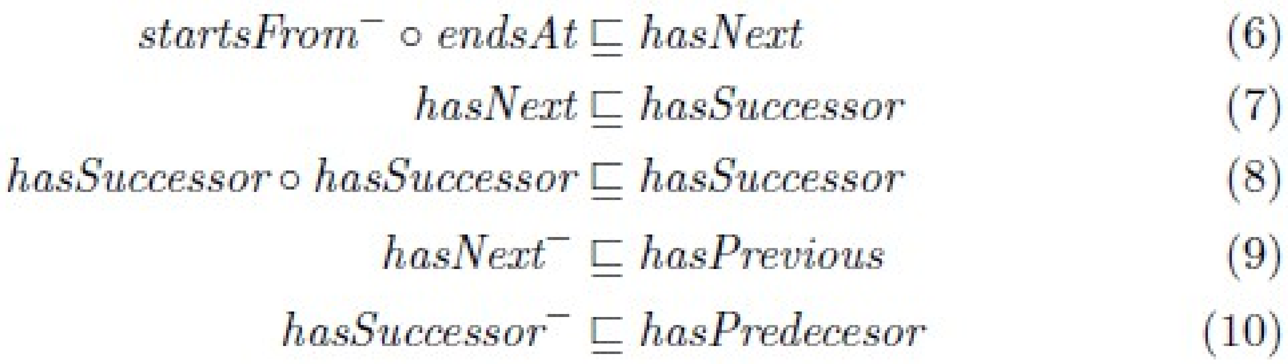


Semantic Trajectory (OWL)



$$
\begin{aligned}
\mathit{Fix} \sqsubseteq\ & \exists \mathit{atTime}.\mathit{OWL\text{-}Time{:}Temporal\ Thing} \sqcap \exists \mathit{hasLocation}.\mathit{Position} \\
& \sqcap \exists \mathit{hasFix}^{-}.\mathit{SemanticTrajectory} && (1)
\end{aligned}
$$

$$
\begin{aligned}
\mathit{Segment} &\sqsubseteq \exists \mathit{startsFrom}.\mathit{Fix} \sqcap \exists \mathit{endsAt}.\mathit{Fix} && (2) \\
\top &\sqsubseteq \leq 1 \mathit{startsFrom}.\top && (3) \\
\top &\sqsubseteq \leq 1 \mathit{endsAt}.\top && (4) \\
\mathit{Segment} &\sqsubseteq \exists \mathit{hasSegment}^{-}.\mathit{SemanticTrajectory} && (5)
\end{aligned}
$$



$$
\begin{aligned}
\mathit{startsFrom}^{-} \circ \mathit{endsAt} &\sqsubseteq \mathit{hasNext} && (6) \\
\mathit{hasNext} &\sqsubseteq \mathit{hasSuccessor} && (7) \\
\mathit{hasSuccessor} \circ \mathit{hasSuccessor} &\sqsubseteq \mathit{hasSuccessor} && (8) \\
\mathit{hasNext}^{-} &\sqsubseteq \mathit{hasPrevious} && (9) \\
\mathit{hasSuccessor}^{-} &\sqsubseteq \mathit{hasPredecesor} && (10)
\end{aligned}
$$

Semantic Trajectory (OWL)

$$
\begin{align}
\mathit{Fix} \sqcap \neg\exists \mathit{endsAt}.\mathit{Segment} &\sqsubseteq \mathit{StartingFix} \tag{11}\\
\mathit{Fix} \sqcap \neg\exists \mathit{startsFrom}.\mathit{Segment} &\sqsubseteq \mathit{EndingFix} \tag{12}\\
\mathit{Segment} \sqcap \exists \mathit{startsFrom}.\mathit{StartingFix} &\sqsubseteq \mathit{StartingSegment} \tag{13}\\
\mathit{Segment} \sqcap \exists \mathit{endsAt}.\mathit{EndingFix} &\sqsubseteq \mathit{EndingSegment} \tag{14}
\end{align}
$$

$$
\begin{align}
\mathit{SemanticTrajectory} &\sqsubseteq \exists \mathit{hasSegment}.\mathit{Segment} \tag{15}\\
\mathit{hasSegment} \circ \mathit{startsFrom} &\sqsubseteq \mathit{hasFix} \tag{16}\\
\mathit{hasSegment} \circ \mathit{endsAt} &\sqsubseteq \mathit{hasFix} \tag{17}
\end{align}
$$

$$
\begin{align}
\exists \mathit{hasSegment}.\mathit{Segment} &\sqsubseteq \mathit{SemanticTrajectory} \tag{18}\\
\exists \mathit{hasSegment}^{-}.\mathit{SemanticTrajectory} &\sqsubseteq \mathit{Segment} \tag{19}\\
\exists \mathit{hasFix}.\mathit{Segment} &\sqsubseteq \mathit{SemanticTrajectory} \tag{20}\\
\exists \mathit{hasFix}^{-}.\mathit{SemanticTrajectory} &\sqsubseteq \mathit{Fix} \tag{21}
\end{align}
$$

Cruise Trajectory (OWL)

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

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

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

$$\begin{aligned}\mathsf{Segment} \sqsubseteq\ & (=1\ \mathsf{startsFrom}.\mathsf{Fix}) \sqcap (=1\ \mathsf{endsAt}.\mathsf{Fix}) \\ & \sqcap \exists \mathsf{hasSegment}^{-}.\mathsf{Trajectory}\end{aligned} \tag{4}$$

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

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

$$\mathsf{startsFrom} \circ \mathsf{nextFix} \sqsubseteq \mathsf{endsAt} \tag{7}$$

Cruise Trajectory (OWL)

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

$$\begin{aligned}&\mathsf{Attribute}(\texttt{port\_stop\_arrival})\\ &\mathsf{Attribute}(\texttt{port\_stop\_departure})\end{aligned} \tag{9a,b}$$

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

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

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

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

$$\begin{aligned}&\mathsf{hasTrajectory} \circ \mathsf{hasSegment} \circ \mathsf{isTraversedBy}\\ &\qquad\qquad\qquad \sqsubseteq \mathsf{isUndertakenBy}\end{aligned} \tag{13}$$

Cruise as Event (OWL)

$$\begin{aligned}
\mathsf{Role} \sqcap \exists \mathsf{providesRole}^{-}.\mathsf{Event} \sqsubseteq (=1\ \mathsf{hasRoleType}.\mathsf{RoleType}) & \\
\sqcap \exists \mathsf{isPerformedBy}.\mathsf{Agent} & \quad (14) \\
\mathsf{providesRole} \circ \mathsf{isPerformedBy} \sqsubseteq \mathsf{hasActor} & \quad (15) \\
\mathsf{Cruise} \sqsubseteq \mathsf{Event} & \quad (16) \\
\mathsf{CruiseRoleType} \sqsubseteq \mathsf{RoleType} & \quad (17) \\
\mathsf{CruiseRoleType}(x) \text{ for every role type } x \text{ in } (*) & \quad (18\text{a-t}) \\
R_{\mathsf{Cruise}} \circ \mathsf{owl{:}topObjectProperty} \circ R_{\mathsf{CruiseRoleType}} & \\
\sqsubseteq \mathsf{providesRoleType} & \quad (19) \\
\mathsf{Cruise} \equiv \exists R_{\mathsf{Cruise}}.\mathsf{Self}, & \quad (20) \\
\mathsf{CruiseRoleType} \equiv \exists R_{\mathsf{CruiseRoleType}}.\mathsf{Self} & \quad (21)
\end{aligned}$$

Cruise as Event (OWL)

- 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

Cruise Information Object (OWL)

DaSe Lab

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

$$\begin{aligned}&\mathsf{CruiseInformationObject} \sqsubseteq \\ &\qquad ({=}1\ \mathsf{hasCruiseType}.\mathsf{CruiseType})\end{aligned} \tag{23}$$

$$\begin{aligned}&\mathsf{CruiseType} \equiv \{\texttt{operational}, \texttt{transit}, \\ &\qquad \texttt{maintenance}, \texttt{other\_cruisetype}\}\end{aligned} \tag{24}$$

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

WRIGHT STATE
UNIVERSITY

Disjointness, Domain & Range

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

$$\exists \mathsf{hasFix}.\mathsf{Fix} \sqsubseteq \mathsf{Trajectory},\ \mathsf{Trajectory} \sqsubseteq \forall \mathsf{hasFix}.\mathsf{Fix} \tag{27}$$

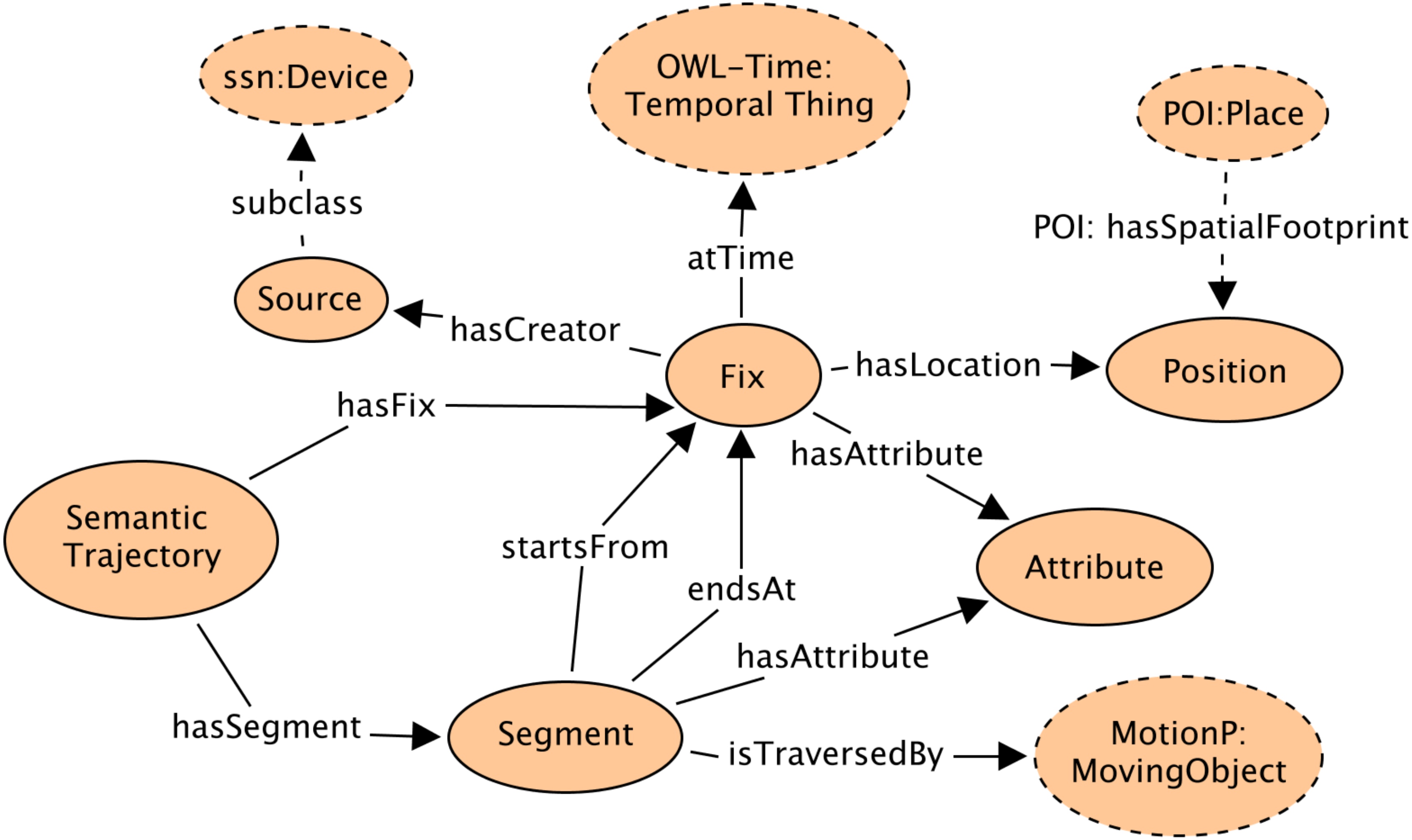
$$\begin{aligned} &\exists \mathsf{hasRelatedCruiseID}.\mathsf{rdf{:}PlainLiteral} \\ &\qquad \sqsubseteq \mathsf{CruiseInformationObject} \end{aligned} \tag{28}$$

$$\begin{aligned} &\mathsf{CruiseInformationObject} \\ &\qquad \sqsubseteq \forall \mathsf{hasRelatedCruiseID}.\mathsf{rdf{:}PlainLiteral} \end{aligned} \tag{29}$$

The EarthCube wish list

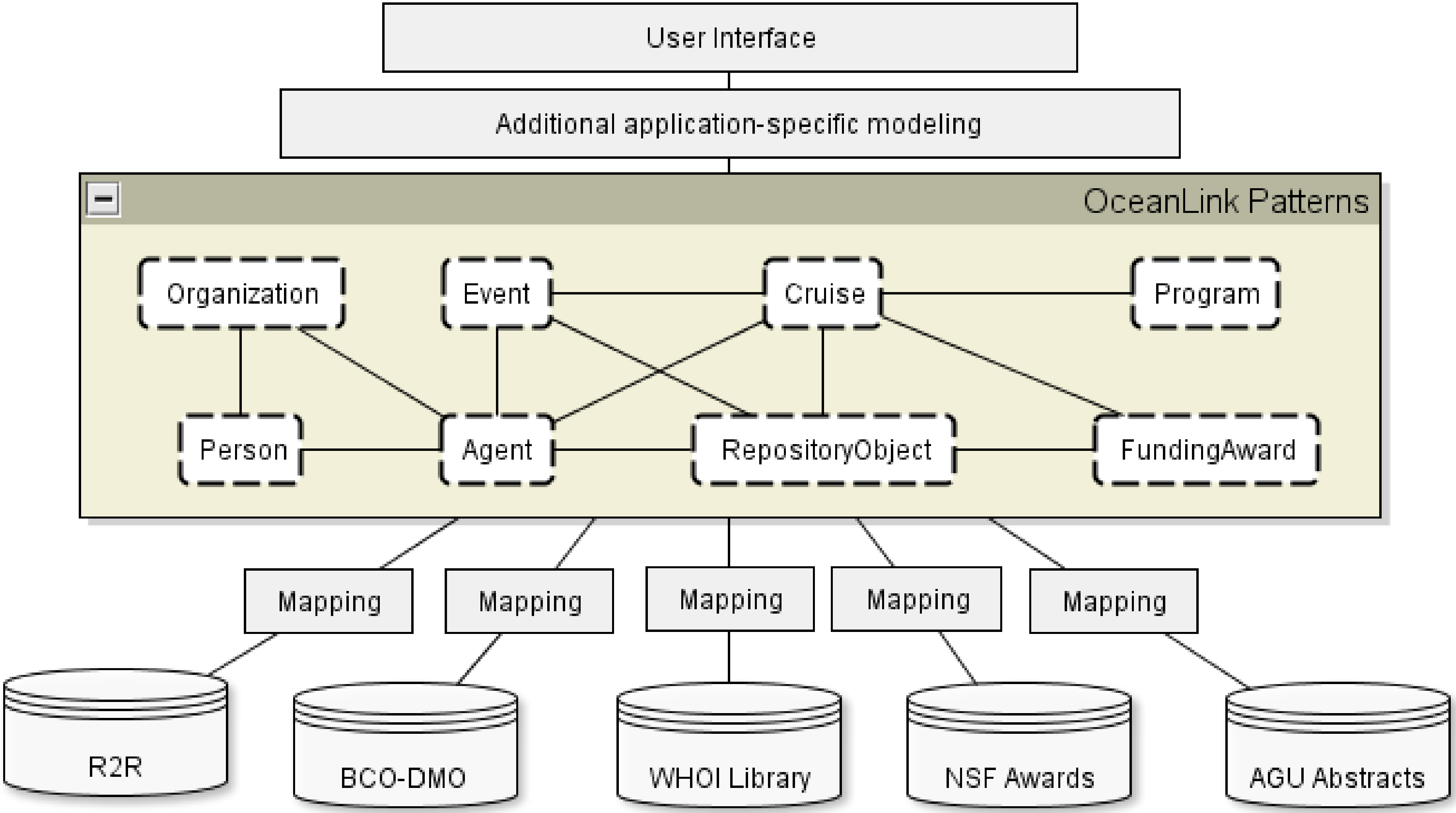
- Modular
- Extendible
- Sustainable
- Sliceable (you can adopt part of it without adopting all)
- Simple enough for easy adoption
- Complex enough to solve real problems
- Scalable and broad enough to cover multiple topics/domains
- Elastic/flexible enough to allow partners to decide how much they want to share
- Respectful of individual modeling choices

Example: Semantic Trajectory



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

OceanLink Architecture


User Interface
Additional application-specific modeling
OceanLink Patterns
Organization
Event
Cruise
Program
Person
Agent
RepositoryObject
FundingAward
Mapping
Mapping
Mapping
Mapping
Mapping
R2R
BCO-DMO
WHOI Library
NSF Awards
AGU Abstracts