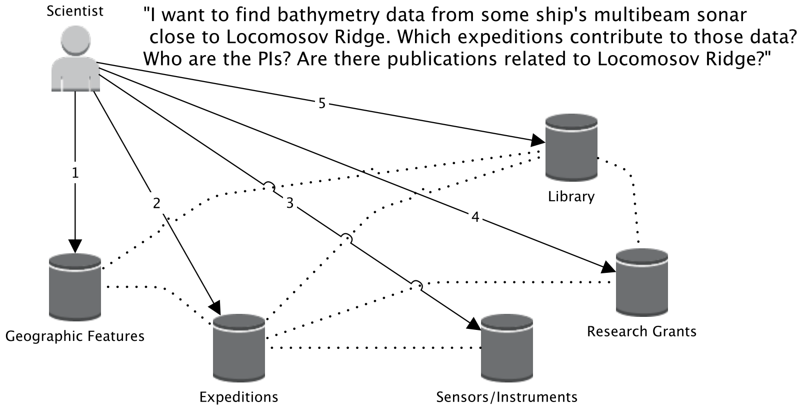
Modular Ontology Architecture for Data Integration in the GeoLink Project

Adila Krisnadhi

Wright State University

Ontology Summit 2016

Motivation



Needed!

Data integration: providing unified view over data at different sources.

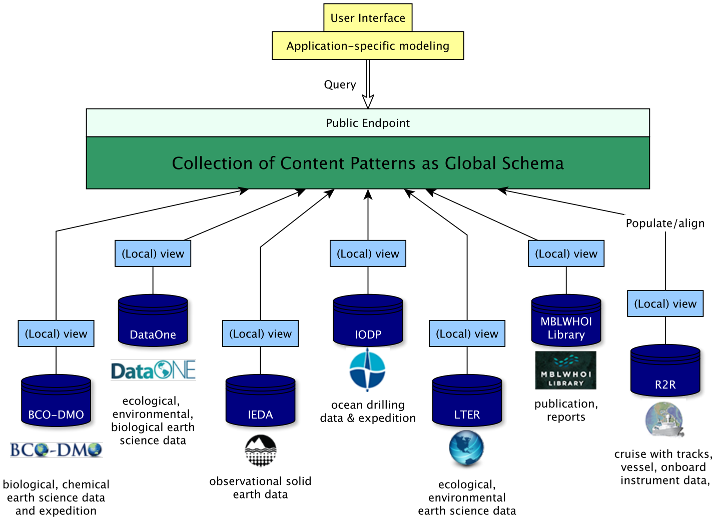
DI Challenges and GeoLink Project

- Challenges (regardless of architecture):
  - Syntactic heterogeneity: different data formats, serializations.
  - Semantic heterogeneity: different vocabulary, different level of granularity in data, different conceptualization.
  - Social/non-technical: inability/unwillingness to participate, fear of unanticipated cost, worry with major changes in their local system, skeptic with scalability
- GeoLink Project (`www.geolink.org`)
  - Part of NSF's EarthCube Program – one among dozens of building block projects.
  - Linked Data + Ontology design pattern-based integration.

Why Ontology Patterns?

- Upper-level and many domain ontologies are:
  - Hard to understand — too many terms, too abstract, too complicated axioms, too far from real data
  - Impose ontological commitments that may not be acceptable by all parties.
  - Brittle — costly/hard to extend, carelessly extending may cause the whole thing breaks.
- Ontology design pattern (ODP): a ("reusable") solution of a frequently occurring modeling problem in the domain and can act as a building block of a more complex ontology.
- Content pattern (CP): an ODP that models a particular generic notion in a particular domain.
- Community engagement via collaborative modeling

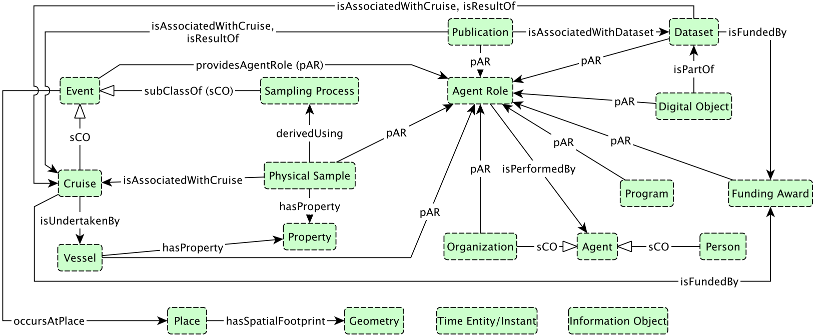
GeoLink Integration Architecture



Modeled Notions

- Content patterns corresponding to concrete domain notions:
  - Cruise, Vessel, Person, Organization, Funding Award, Program, Physical Sample, Dataset, Digital Object, Publication, Platform, Place, Time.
- Content patterns from abstraction in modeling:
  - Agent, Agent Role, Event, Information Object, Identifier, Personal Info Item, Person Name, Property Value.

Ontology Overview



Each node represents a content pattern.

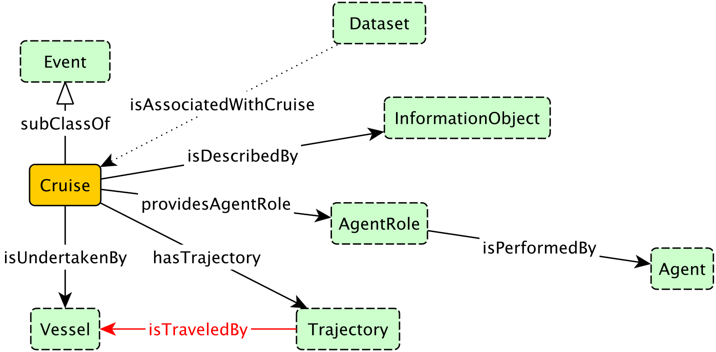
CP Example: Cruise

DaSe Lab

- Generate competency questions
  - "Find all cruises passing through Gulf of Maine in August 2013."
  - "Show the tracks 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?"
- Understand the nature of things we model.
  - Cruise .... is an Event
  - Track .... maybe complex, reuse Trajectory pattern?[1]
  - Vessel .... maybe complex
  - Chief scientist ..... a role of an agent
  - Dataset .... maybe complex
  - Funding award .... maybe complex

[1] Hu, et al. "A geo-ontology design pattern for semantic trajectories", COSIT 2013

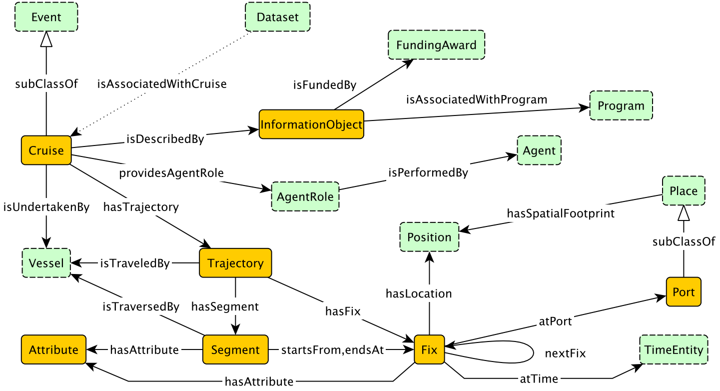
DaSe Lab


Dataset
Event
isAssociatedWithCruise
InformationObject
subClassOf
isDescribedBy
Cruise
providesAgentRole
AgentRole
isPerformedBy
Agent
isUndertakenBy
hasTrajectory
Vessel
isTraveledBy
Trajectory

Use informal natural language to model axioms together with domain experts/data providers.

- Cruise $\sqsubseteq$ Event
- Cruise has exactly 1 trajectory and is undertaken by exactly 1 vessel.
  $\text{Cruise} \sqsubseteq (=1\ \text{hasTrajectory.Trajectory}) \sqcap (=1\ \text{isUndertakenBy.Vessel})$
- Cruise is described by exactly 1 information object.
  $\text{Cruise} \sqsubseteq (=1\ \text{isDescribedBy.InformationObject})$
- Trajectory of a cruise must be traveled by the vessel by which the cruise is undertaken.
  $\text{hasTrajectory}^{-} \circ \text{isUndertakenBy} \sqsubseteq \text{isTraveledBy}$

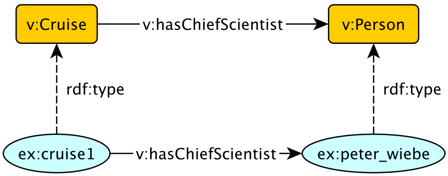
Cruise Trajectory



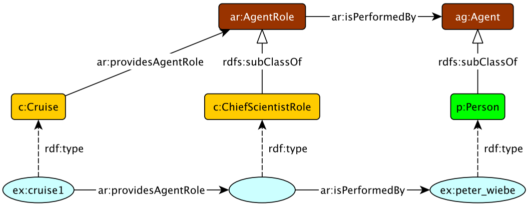
- Since patterns represent key notions as understood by domain experts and data providers, intuitively an appropriate mapping/alignment exists between "local" vocabulary and the patterns.
- A (local) pattern view between a data source and the patterns makes such a mapping explicit.
  - View is a very minimalistic schema (class names, property names, simple domain and range axioms)
  - Separating "core conceptualization" and "nomenclature" issues: vocabulary terms in a local view may be repository-specific and need not be the same as the patterns.
  - Mapping can be expressed in rules that help populating the patterns.
  - Data providers can populate the global schema (pattern collection) by simply populating a local view.
  - Existing controlled vocabulary can also be accommodated as a pattern view.

Pattern View Example

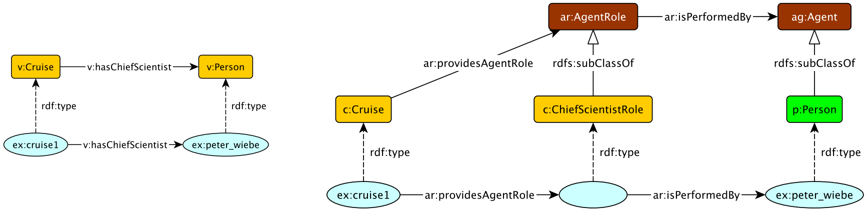
Producer populates view:



to populate Cruise, Agent Role, and Person patterns:

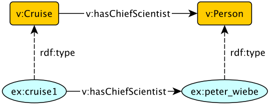


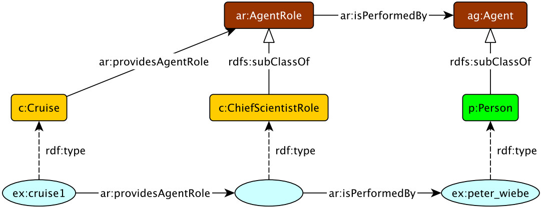
Mapping Rule



$$
\begin{aligned}
&\text{v:Cruise}(X) \wedge \text{v:hasChiefScientist}(X, Y) \wedge \text{v:Person}(Y) \\
&\quad \longrightarrow \exists Z.\big(\text{c:Cruise}(X) \wedge \text{ar:providesAgentRole}(X, Z) \\
&\qquad \wedge \text{c:ChiefScientistRole}(Z) \\
&\qquad \wedge \text{ar:isPerformedBy}(Z, Y) \wedge \text{p:Person}(Y)\big)
\end{aligned}
$$

DaSe Lab






```
CONSTRUCT {
  ?X a c:Cruise ;
     ar:providesAgentRole [ a c:ChiefScientistRole ;
                            ar:isPerformedBy ?Y ] .
  ?Y a p:Person .
}
WHERE {
  ?X a v:Cruise ; v:hasChiefScientist ?Y .
  ?Y a v:Person .
}
```

Conclusion and Outlook

DaSe Lab

- The GeoLink modular oceanography ontology = collection of content patterns in oceanography.
- Collaborative modeling approach.
- Two-layered ontology architecture with patterns and local views helps semantic interoperability across different data sources, while allowing data providers to retain their own local vocabulary and schema.
- See also: `http://www.geolink.org`, `http:/schema.geolink.org`

Questions?

Acknowledgement:

- NSF for GeoLink funding