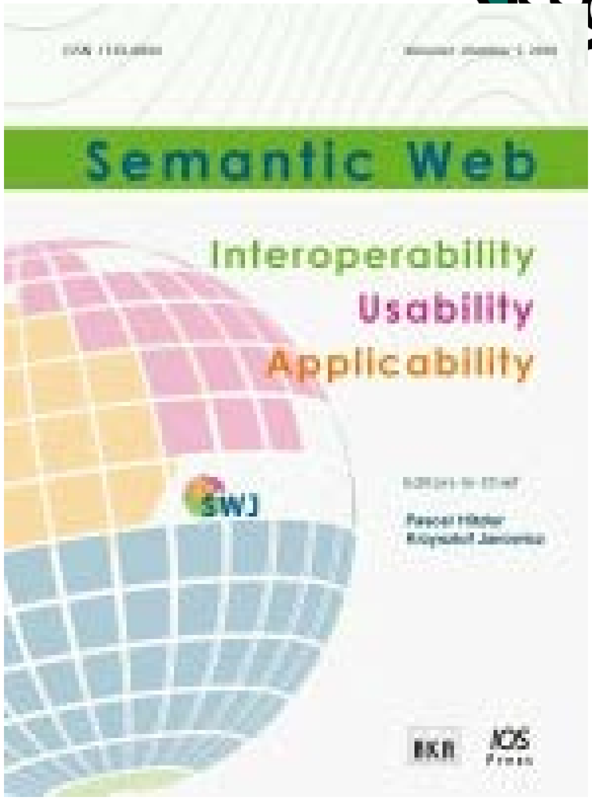
Semantic Technologies for Big Data Integration

Pascal Hitzler
Data Semantics Laboratory (DaSe Lab)
Data Science and Security Cluster (DSSC)
Wright State University
http://www.pascal-hitzler.de

Semantic Web journal

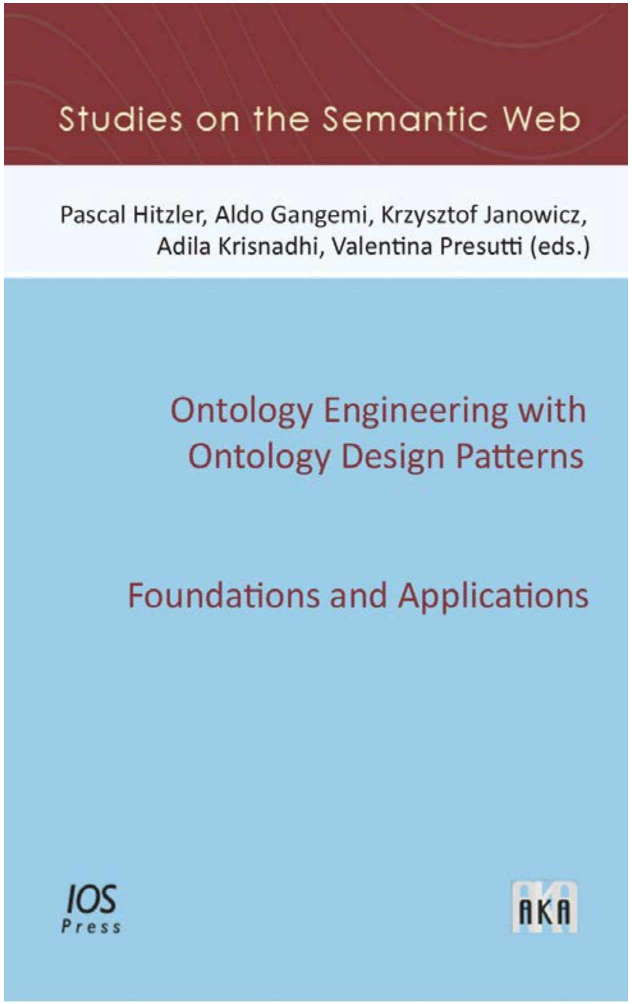
- **EiCs: Pascal Hitzler**
  Krzysztof Janowicz
- **Funded 2010**
- **2016 Impact factor of 1.786, top of all journals with "Web" in the title**

- **We very much welcome contributions at the "rim" of traditional Semantic Web research – e.g., work which is strongly inspired by a different field.**
- **Non-standard (open & transparent) review process.**



- **http://www.semantic-web-journal.net/**

Forthcoming book



Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnathi, Valentina Presutti (eds.), Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web. IOS Press/AKA Verlag, 2016/2017. To appear.

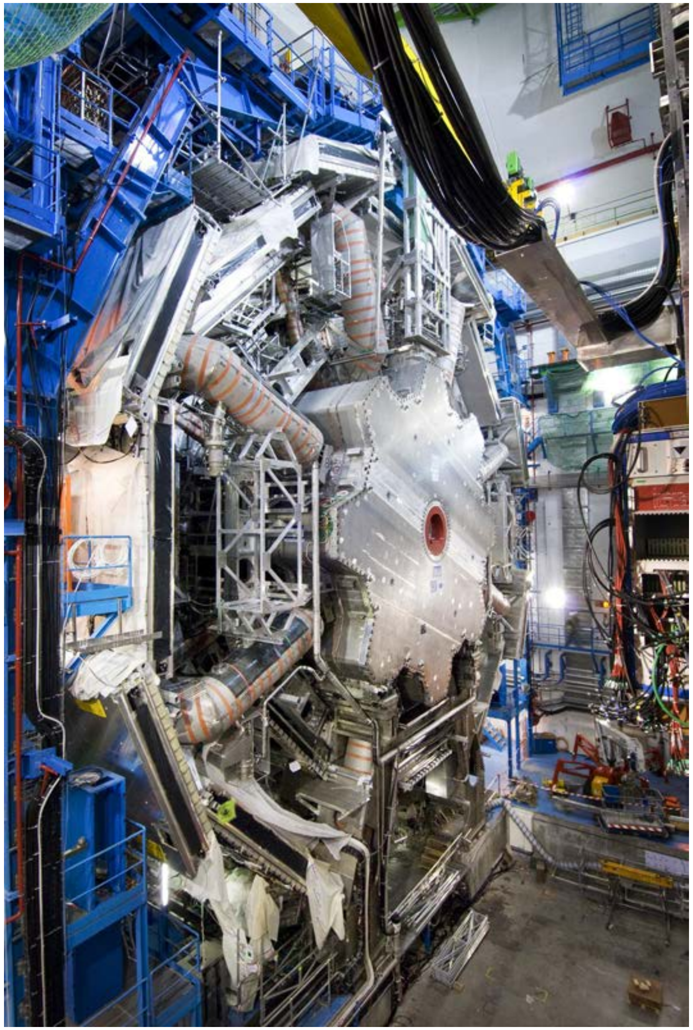
"Big" Data

What is "big"?

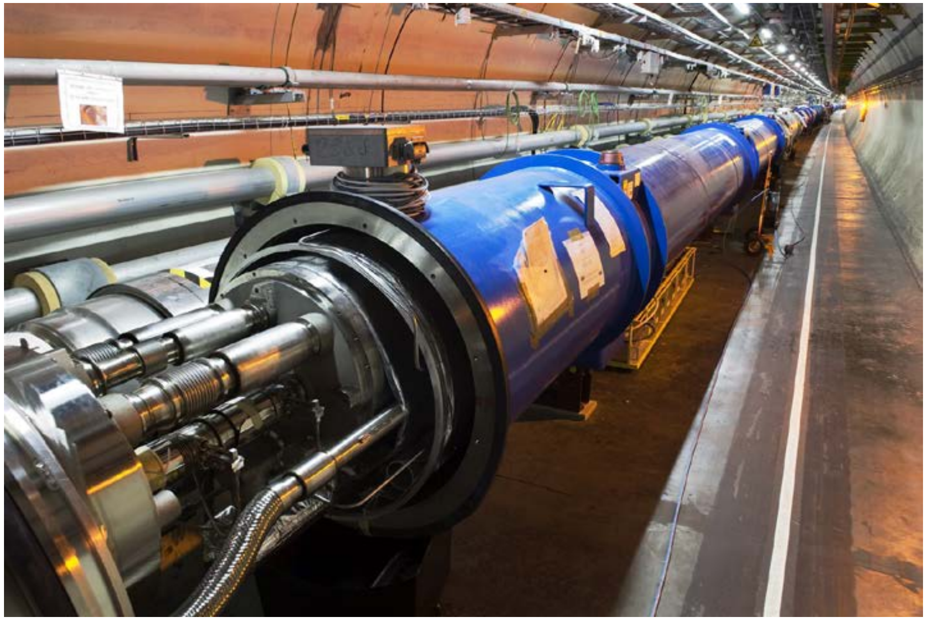
Some Primary Goals

- **data/information *sharing***
- **data/information *discovery***
- **data/information *integration***
- **data/information *reuse***

A Use Case Description



Large Hadron Collider (LHC) at CERN
experiments: ALICE
ATLAS
CMS
LHCb



Photos: ATLAS Experiment © 2014 CERN

A Use Case Description

At these experiments, billions or trillions of particle collisions are analyzed to determine probabilities or probability densities associated with a given physical process.

Very careful attention must be paid to defining the measurement that is to be made.

To date, there is no formal way of representing or classifying such experimental results, despite thousands of papers published since the 40s.

A Use Case Description

With a formal representation, e.g. an ATLAS physicist or a theorist could search an external database for previous work done by CMS in order to compare results.

Or even, say, an ATLAS researcher could search an internal database for previous examples similar to a planned analysis, saving substantial time and effort.

E.g.

- **Retrieve all analyses that used jets in the final state.**
- **Retrieve all analyses that veto extra leptons.**
- **Retrieve all analyses requiring large missing energy.**
- **Retrieve all analyses involving some electron with** $p_T > 40\ \mathrm{GeV}$**.**

Questions

- How do you set this up such that it does not only pertain to one particular CERN experiment, so that you can search across CERN experiments, across different accelerators, etc?
- How do you organize your data without knowing what types of questions will be asked in the future?
- How do you distinguish between base data and interpreted or computationally assessed data. What does this difference mean anyway in the context of HEP?

[Collaboration between DaSeLab and U. Notre Dame, CERN, U Washington, and others, in the context of the DASPOS NSF project]

[WOP 2015, ACAT 2016]

Another Scenario

The NSF EarthCube Program:

Developing a Community-Driven Data and Knowledge Environment for the Geosciences

"concepts and approaches to create integrated data management infrastructures across the Geosciences."

"EarthCube aims to create a well-connected and facile environment to share data and knowledge in an open, transparent, and inclusive manner, thus accelerating our ability to understand and predict the Earth system."

EarthCube GeoLink Scenario

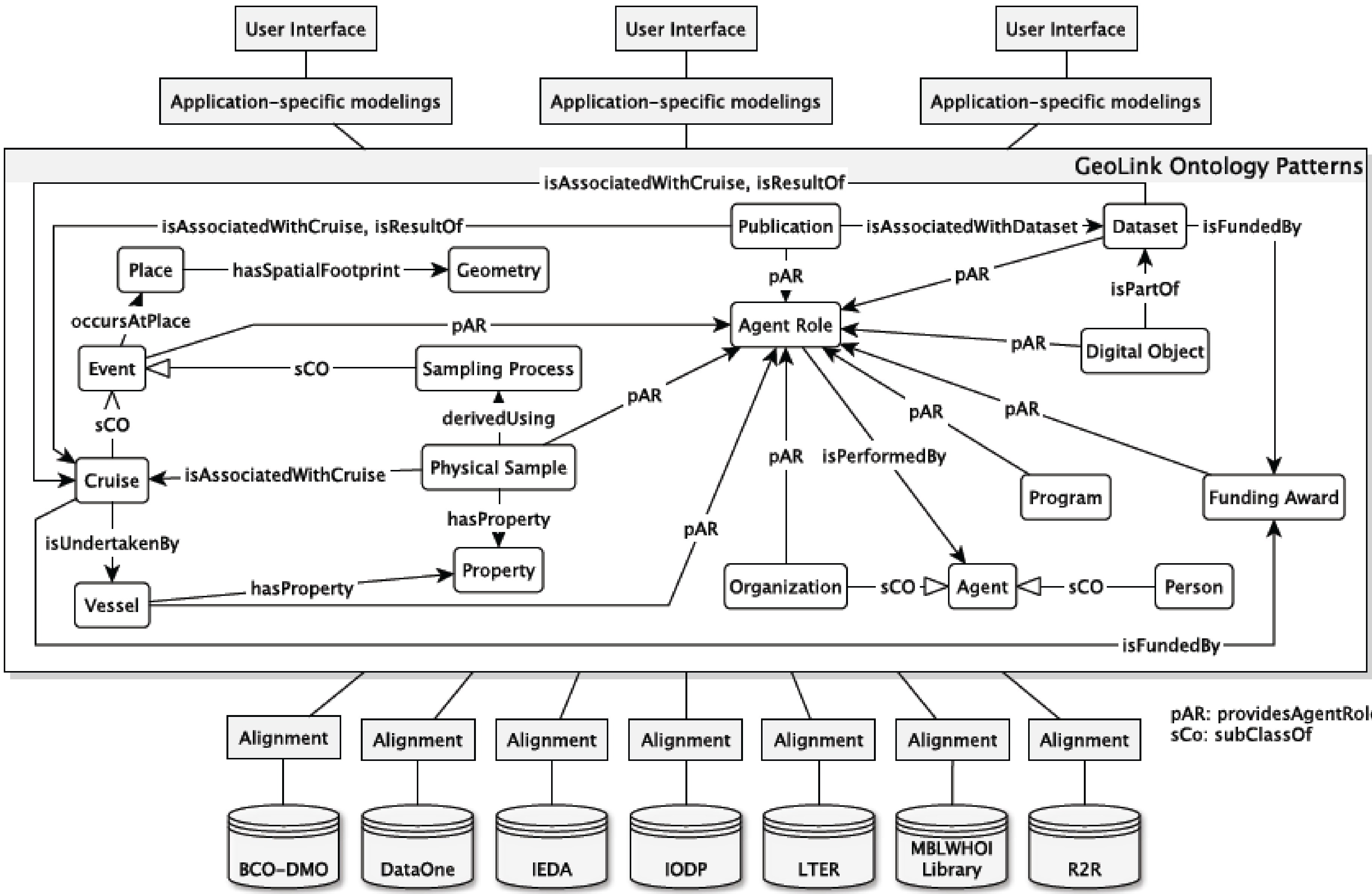
GeoLink: An EarthCube "Building Block" project (2014-2017)

How to realize data search across many large-scale geoscience data repositories, such that

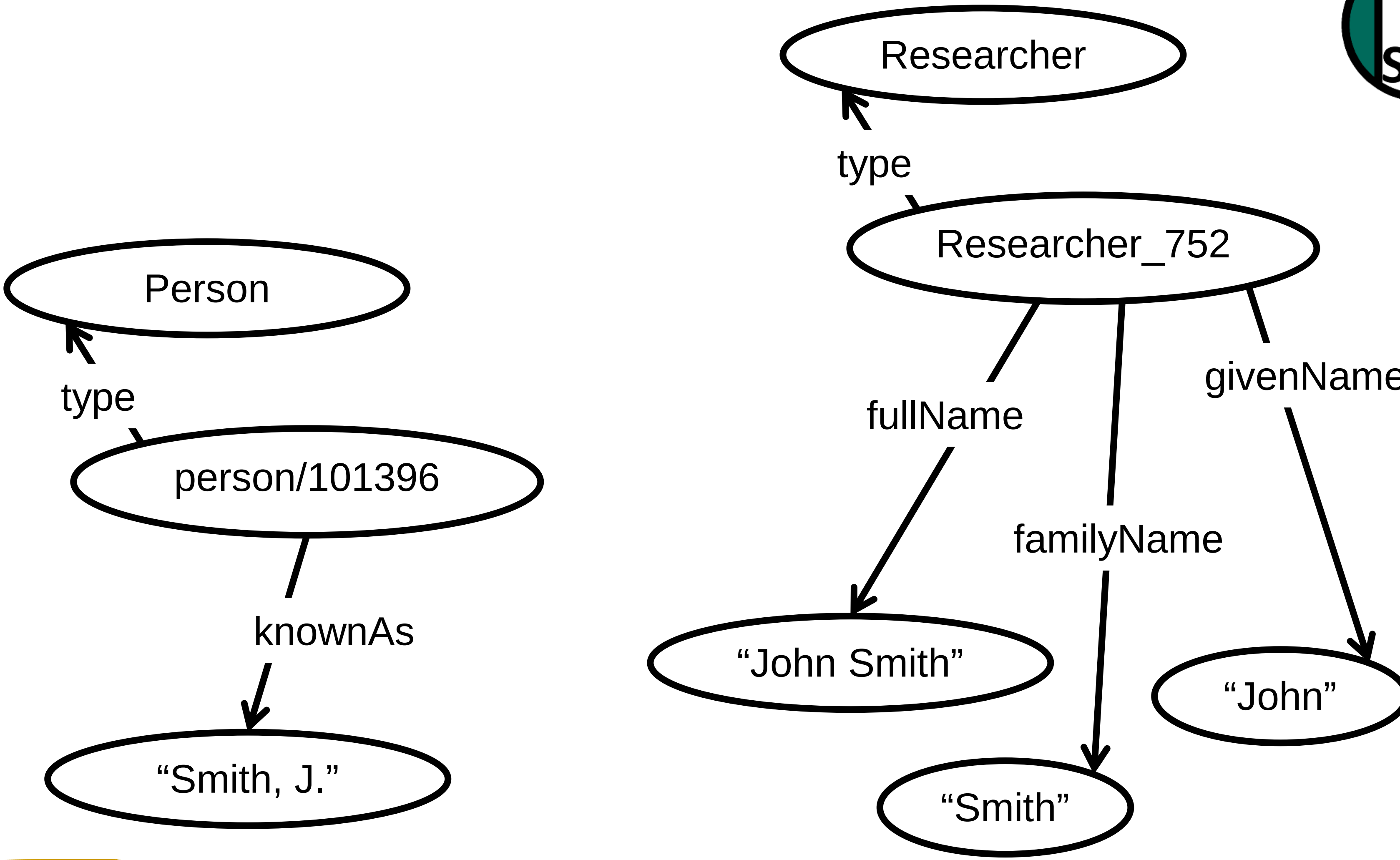
- **The approach is extendable to new repositories.**
- **The scope can extend across all of the Geosciences.**
- **The search capabilities can be made more fine-grained in the future if desired.**

Central idea: Use a modular, extendable ontology for the integration of metadata.

The GeoLink Framework



Regarding Integration

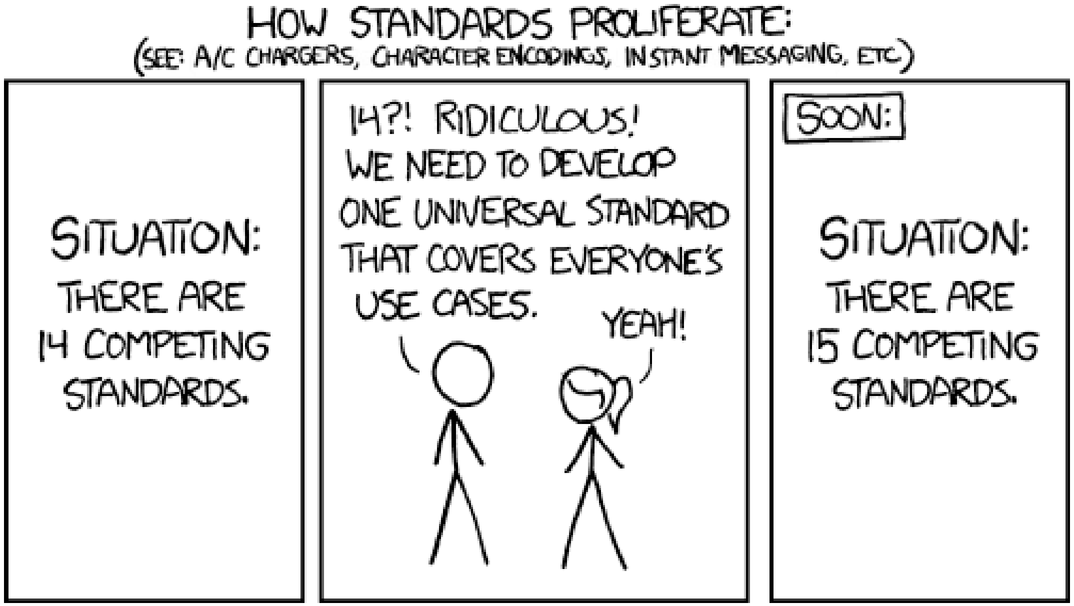


Use of standards

Standardization:

The traditional approach to data sharing, discovery, integration, reuse.

What are the limits of standardization?



Standards

- What is a road?
- What is a forest?
- What is marriage?
- What is a Higgs Boson?

We cannot standardize everything, it's too much.

We cannot standardize everything, because ambiguity is as much a feature as it is a bug.

Idea

- **Let's not establish a standard for everything.**
- **Instead, let's standardize a language *for making machine-readable definitions*.**

Definitions

Wikipedia:

A *forest* is a a large area of land covered with trees or other woody vegetation.

A *road* is a thoroughfare, route, or way on land between two places that has been paved or otherwise improved to allow travel by some conveyance, including a horse, cart, bicycle, or motor vehicle.

A *compactification* is the process or result of making a topological space into a compact space. A *compact space* is a topological space every open cover of which has a finite subcover.

We define terms by stating how they relate to other terms.

This is of course circular, but it's really the only way we can do it.

Web Ontology Language (OWL)

OWL is a (constrained, mathematically precise) language for stating definitions (i.e., relations between terms).

It is essentially a constrained version of first-order predicate logic.

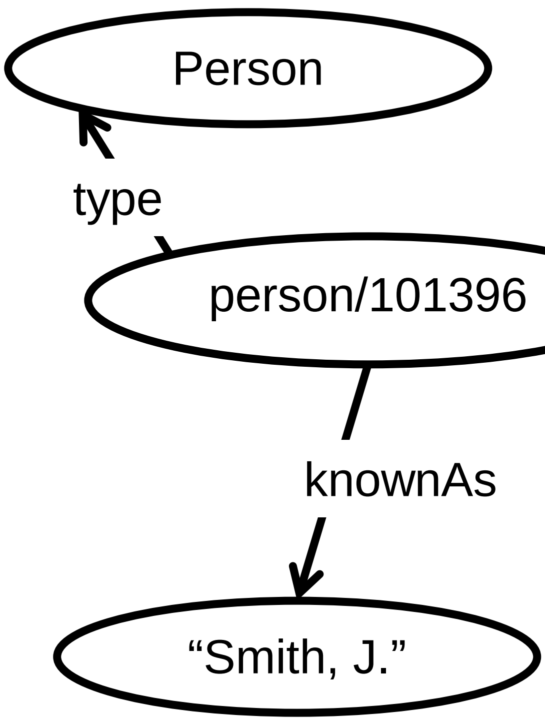
Serializations: several, some more human-readable, some more machine-readable. For the latter, mostly using RDF/XML.

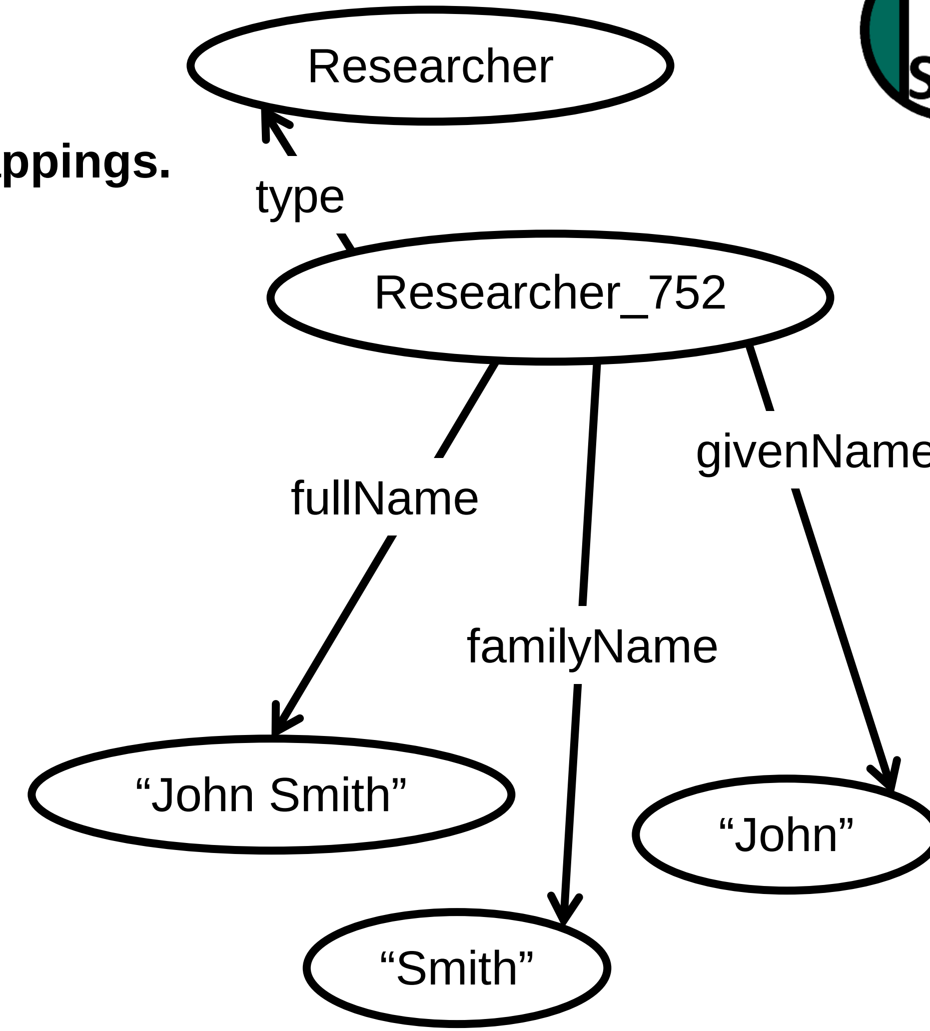
[W3C 2012]

So what about integration now?

Researcher(x) -> Person(x)

We may also want more complex mappings.





Ontology Alignment

Is about finding mappings between two different ontologies.

Let's look at the simplest case:

Class matching.

I.e. aligning classes (types) between the different ontologies, such as Person and Researcher in the previous example.

Some systems detect sub-class relationships.

Most systems detect same-class relationships.

Mostly string matching

[Cheatham, ISWC 2013]

Table 1. Results of strings only approaches and the competitors from the OAEI 2012 competition on the conference data set (left) and the anatomy data set (right)

Metric	Prec.	Recall	F-meas.
YAM++	0.81	0.69	0.75
LogMap	0.82	0.58	0.68
StringsOpt	**0.85**	**0.55**	**0.67**
StringsAuto	**0.79**	**0.57**	**0.66**
Optima	0.62	0.68	0.65
CODI	0.74	0.57	0.64
GOMMA	0.85	0.47	0.61
Wmatch	0.74	0.50	0.60
WeSeE	0.76	0.49	0.60
Hertuda	0.74	0.50	0.60
MaasMatch	0.63	0.57	0.60
LogMapLt	0.73	0.50	0.59
HotMatch	0.71	0.51	0.59
Baseline 2	0.79	0.47	0.59
ServOMap	0.73	0.46	0.56
Baseline 1	0.80	0.43	0.56
ServOMapLt	0.88	0.40	0.55

Metric	Prec.	Recall	F-meas.
GOMMA-bk	0.92	0.93	0.92
YAM++	0.94	0.86	0.90
CODI	0.97	0.83	0.89
StringsOpt	**0.88**	**0.87**	**0.88**
LogMap	0.92	0.85	0.88
GOMMA	0.96	0.80	0.87
StringsAuto	**0.86**	**0.84**	**0.85**
MapSSS	0.94	0.75	0.83
WeSeE	0.91	0.76	0.83
LogMapLt	0.96	0.73	0.83
TOAST*	0.85	0.76	0.80
ServOMap	1.00	0.64	0.78
ServOMapLt	0.99	0.64	0.78
HotMatch	0.98	0.64	0.77
AROMA	0.87	0.69	0.77
StringEquiv	1.00	0.62	0.77
Wmatch	0.86	0.68	0.76

Mostly string matching

[Cheatham, under review]

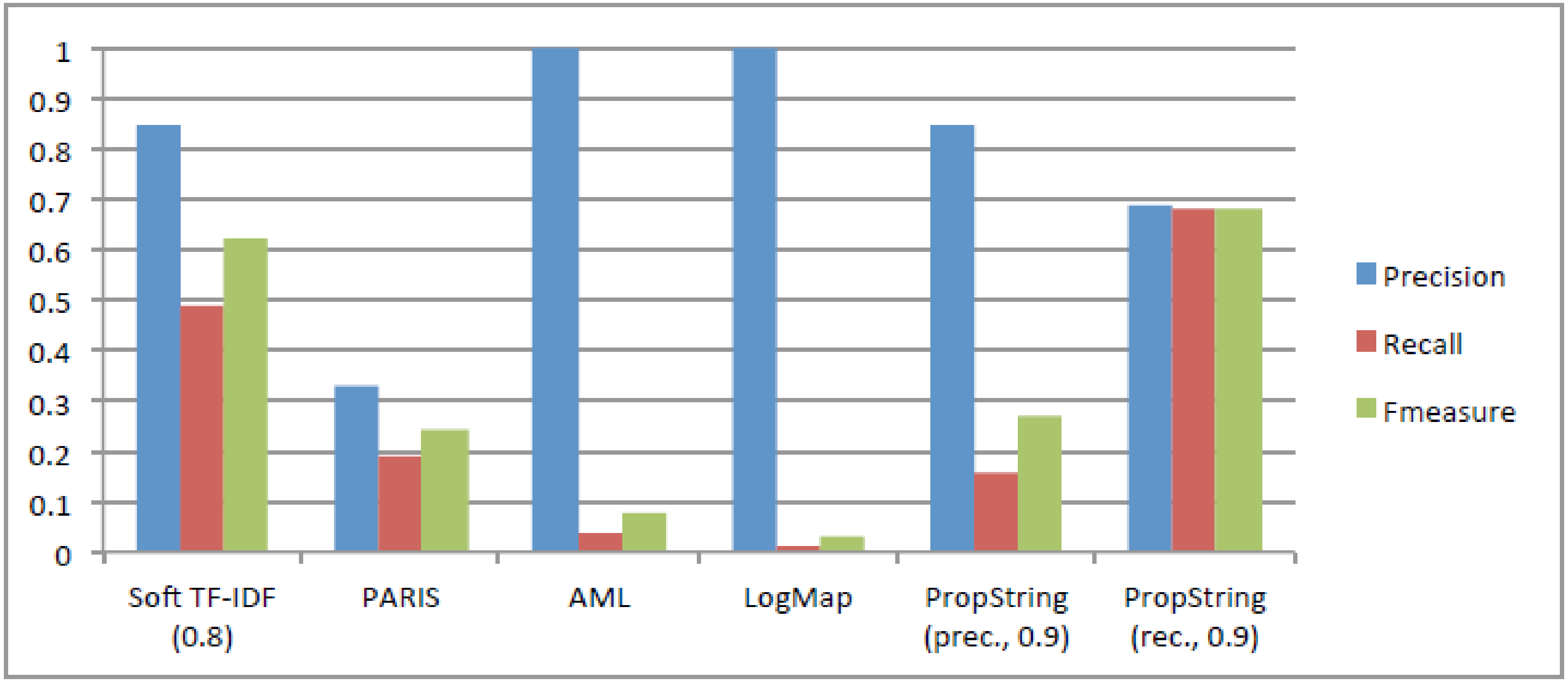
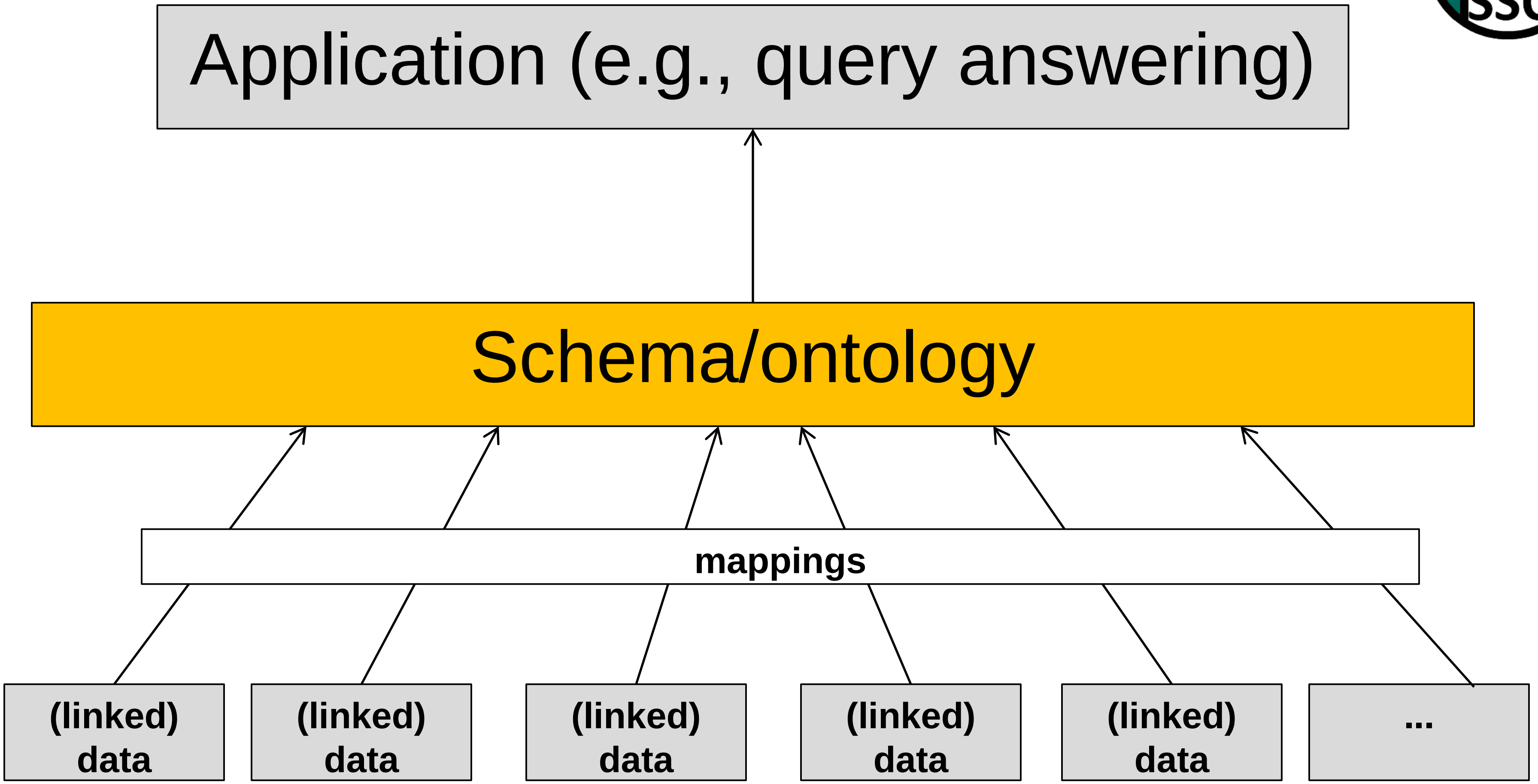


Fig. 1 Results of the YAGO-DBPedia alignment task

Schema integration

Goal: making manual integration easier.



Definitions

- **What is a road?**
- **What is a forest?**
- **What is marriage?**
- **What is a Higgs Boson?**

They may mean (slightly, or very) different things for different data sources.

How do we integrate that?

EarthCube Challenges

The EarthCube "Architecture" must be

- **modular**
- **extensible**
- **sustainable**
- **sliceable (i.e. you can adopt part of it without adopting all)**
- **simple enough for easy adoption**
- **complex enough to solve real problems**
- **scalable in terms of breadth of topic coverage**
- **elastic, in that it allows partners to decide how much they want to share**
- **respectful of individual modeling choices**

Three modeling principles

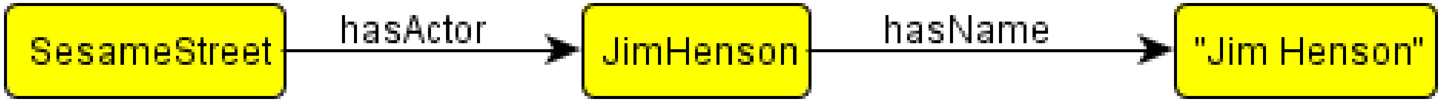
1. Borrow from best practices to make generic schema which fits (relatively) many purposes.
   I.e. which respects heterogeneity.

2. Modularize your ontology to make it manageable and flexible (e.g. by modifying/replacing independent modules, extending with new modules, etc.).

3. Provide simplified views on your ontology for different users if needed.

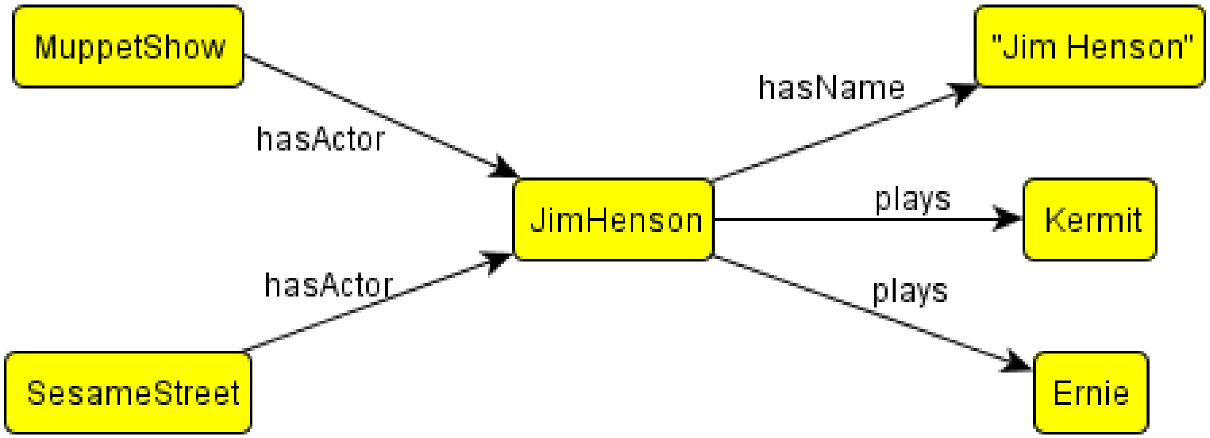
ABoxes as graphs

SesameStreet	**has Actor**	**JimHenson .**
JimHenson	**hasName**	**“Jim Henson” .**

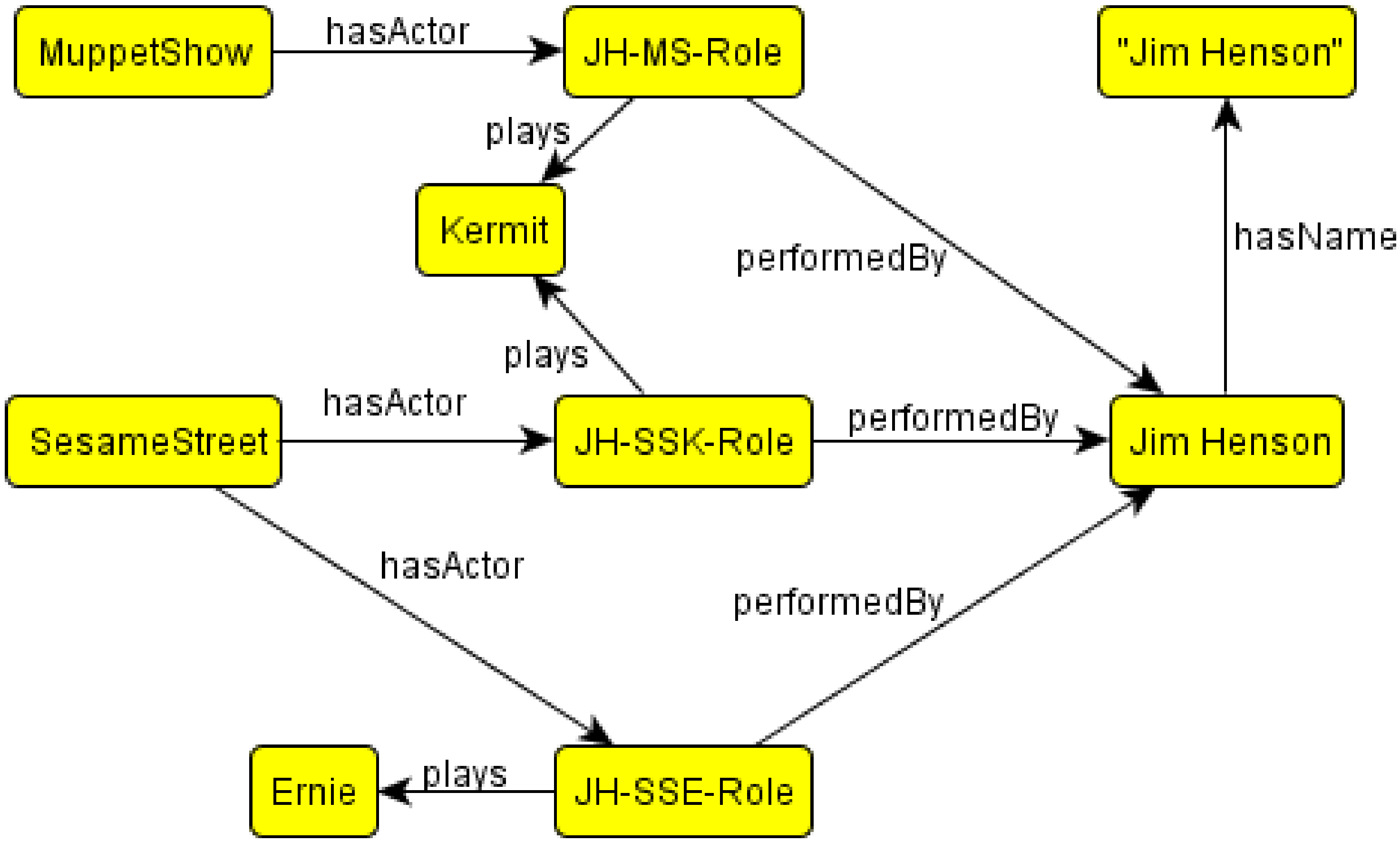


Problem!

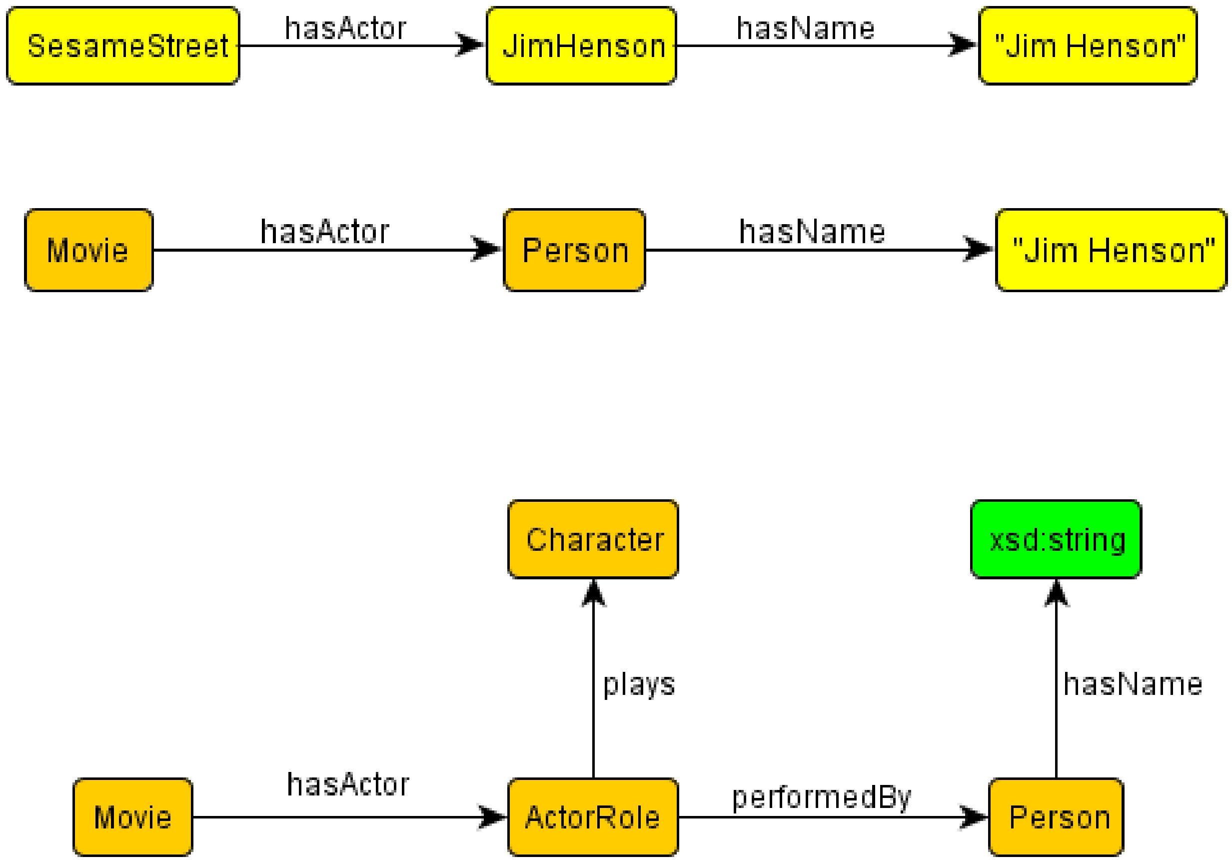
SesameStreet	**has Actor**	**JimHenson .**
MuppetShow	**has Actor**	**JimHenson .**
JimHenson	**plays**	**Kermit .**
JimHenson	**plays**	**Ernie .**
JimHenson	**hasName**	**“Jim Henson” .**



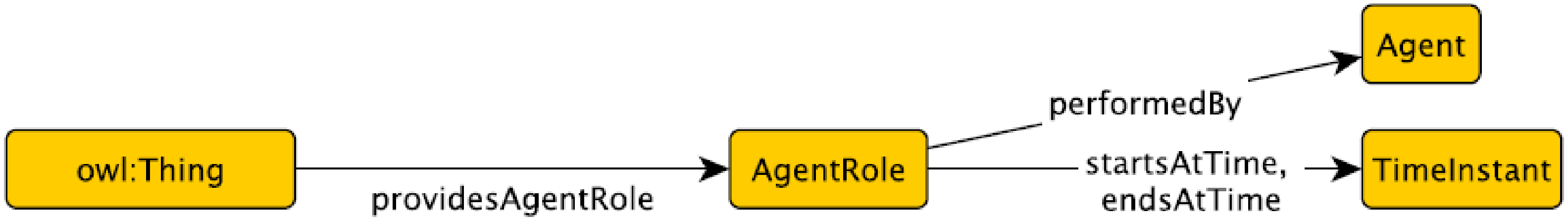
Solution!


MuppetShow
hasActor
JH-MS-Role
"Jim Henson"
plays
Kermit
performedBy
hasName
plays
hasActor
SesameStreet
JH-SSK-Role
performedBy
Jim Henson
hasActor
performedBy
Ernie
plays
JH-SSE-Role

Schematically



Ontology Design Patterns



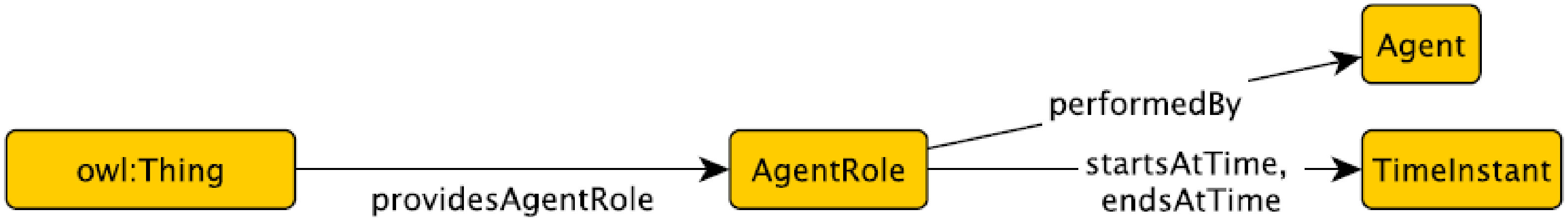
An *Ontology Design Pattern* (ODP) is a reusable successful solution to a recurrent ontology modeling problem.

[Gangemi 2005]

So-called *content patterns* usually encode specific abstract notions, such as process, event, agent, etc.

[SWJ 2016]

Axiomatization



$$\top \sqsubseteq \forall \mathsf{providesAgentRole}.\mathsf{AgentRole}$$

$$\mathsf{AgentRole} \sqsubseteq \forall \mathsf{performedBy}.\mathsf{Agent}$$

$$\exists \mathsf{performedBy}.\mathsf{Agent} \sqsubseteq \mathsf{AgentRole}$$

$$\mathsf{AgentRole} \sqsubseteq \forall \mathsf{startsAtTime}.\mathsf{TimeInstant}$$

$$\mathsf{AgentRole} \sqsubseteq \forall \mathsf{endsAtTime}.\mathsf{TimeInstant}$$

$$\mathsf{AgentRole} \sqsubseteq \exists \mathsf{providesAgentRole}^{-}.\top$$

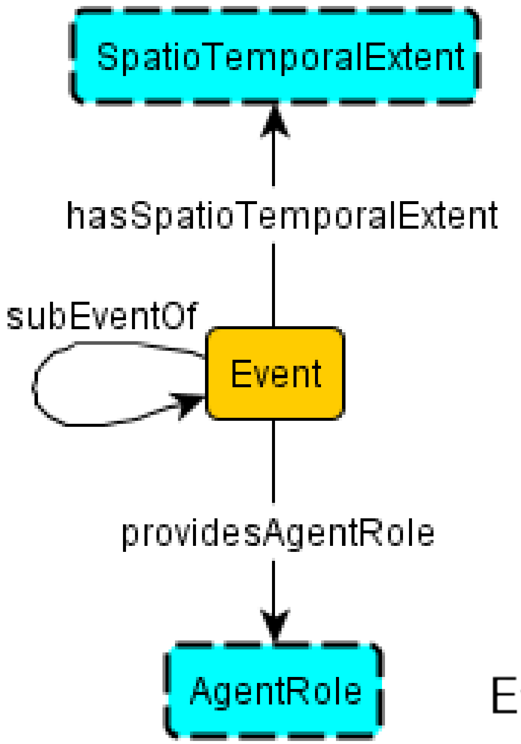
$$\mathsf{AgentRole} \sqsubseteq {=}1 \mathsf{performedBy}.\mathsf{Agent}$$

$$\mathsf{AgentRole} \sqsubseteq {=}1 \mathsf{startsAtTime}.\mathsf{TimeInstant}$$

$$\mathsf{AgentRole} \sqsubseteq {=}1 \mathsf{endsAtTime}.\mathsf{TimeInstant}$$

$$\mathsf{DisjointClasses}(\mathsf{AgentRole}, \mathsf{Agent}, \mathsf{TimeInstant})$$

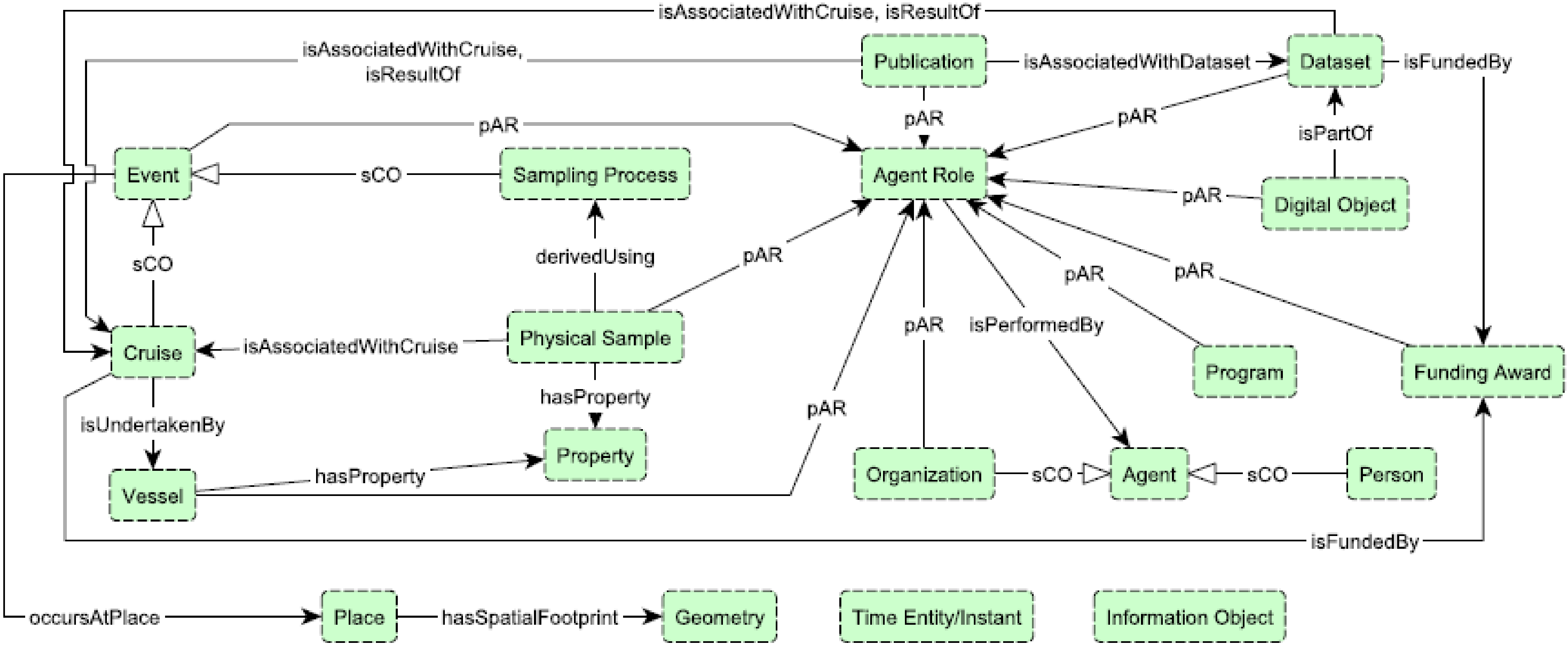
A Minimal Event Pattern



$$\top \sqsubseteq \forall \mathsf{hasSpatioTemporalExtent.SpatioTemporalExtent}$$
$$\top \sqsubseteq \forall \mathsf{providesAgentRole.AgentRole}$$
$$\mathsf{Event} \sqsubseteq \exists \mathsf{hasSpatioTemporalExtent.SpatioTemporalExtent}$$
$$\mathsf{Event} \sqsubseteq \forall \mathsf{subEventOf.Event}$$
$$\exists \mathsf{subEventOf.Event} \sqsubseteq \mathsf{Event}$$
$$\mathsf{subEventOf} \circ \mathsf{subEventOf} \sqsubseteq \mathsf{subEventOf}$$
$$\mathsf{DisjointClasses(Event, AgentRole, SpatioTemporalExtent)}$$

Modularization

DaSe Lab



High-level overview of the GeoLink Modular Ontology (GMO).

Each box stands for a module, which has been modeled in its own right.

[ISWC 2015]

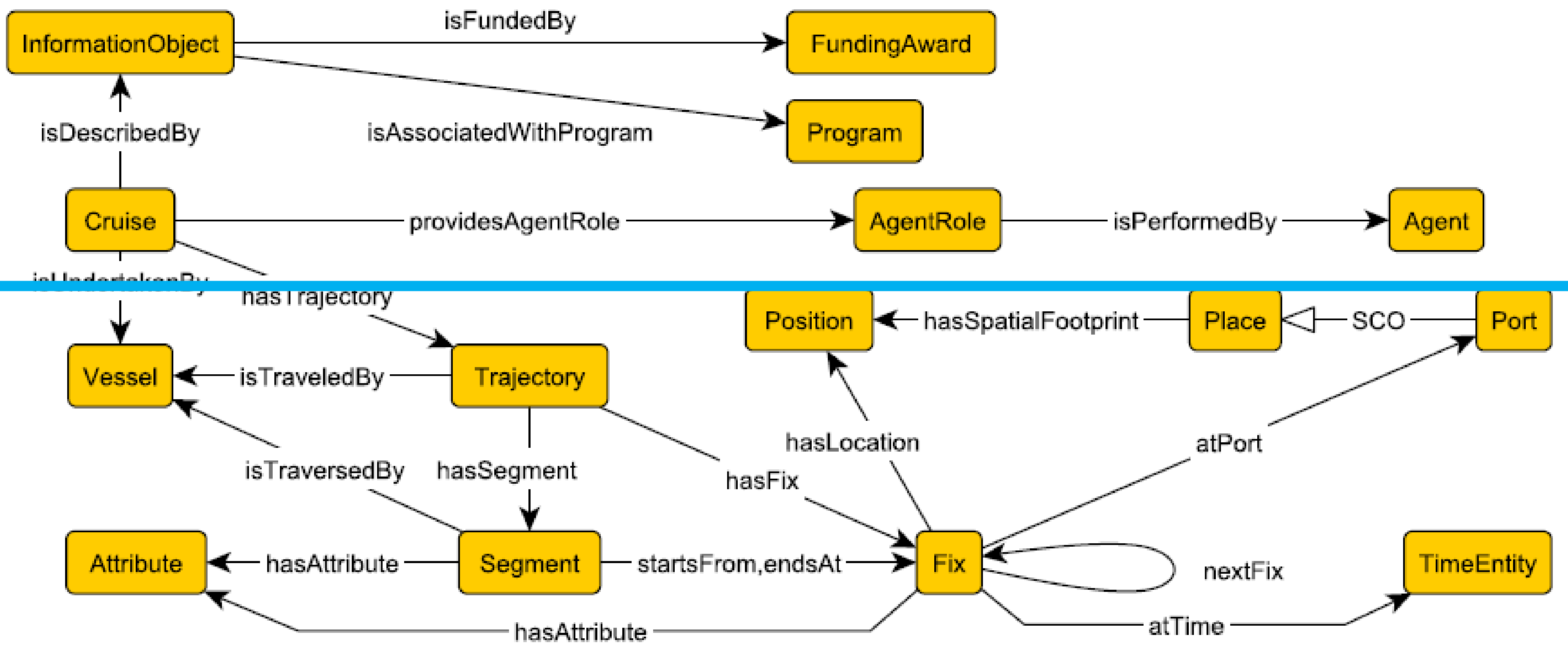
Example Module: Cruise

Cruise reused e.g. the generic patterns

AgentRole

Trajectory

and conceptually cruises are understood to be events.



GeoLink

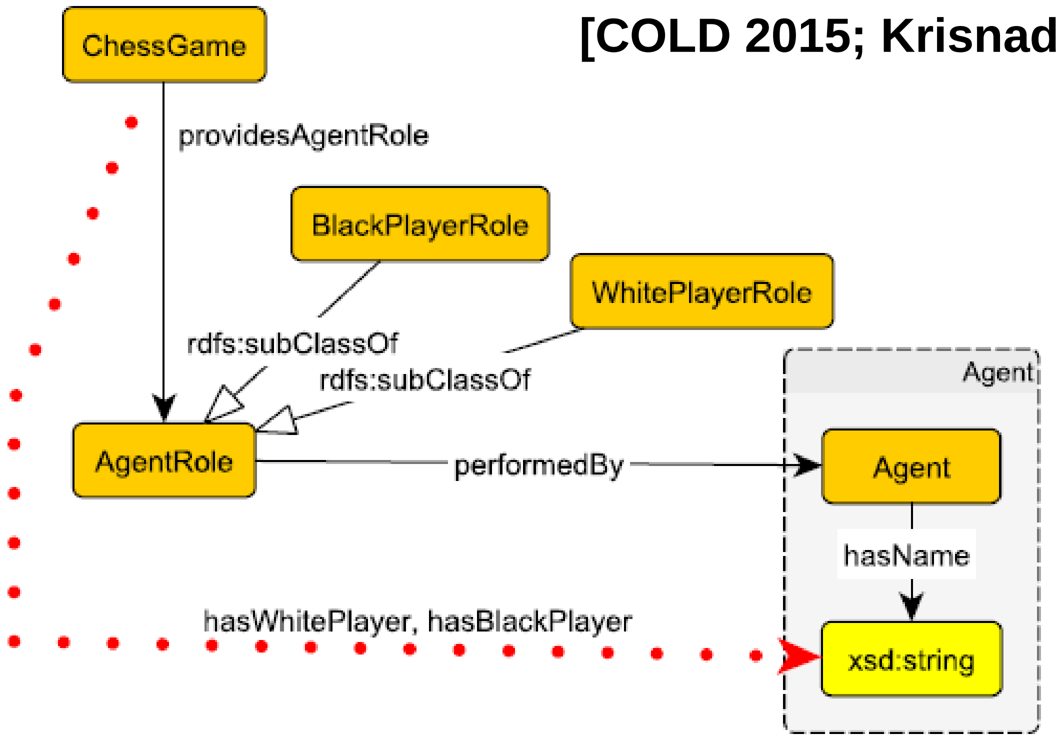
An (preliminary) interactive demonstration of the integrated GeoLink data is available at

http://demo.geolink.org

At http://www.geolink.org/ there are links to the complete schema, a SPARQL Endpoint, publications, etc.

Simplified Views

[COLD 2015; Krisnadhi Dissertation 2015]



$$\mathsf{ChessGame}(x) \wedge \mathsf{pAR}(x,y) \wedge \mathsf{WhitePlayerRole}(y) \wedge \mathsf{performedBy}(y,z) \wedge \mathsf{Agent}(z) \wedge \mathsf{hasName}(z,s) \rightarrow \mathsf{hasWhitePlayer}(x,s)$$

$$\mathsf{ChessGame}(x) \wedge \mathsf{pAR}(x,y) \wedge \mathsf{BlackPlayerRole}(y) \wedge \mathsf{performedBy}(y,z) \wedge \mathsf{Agent}(z) \wedge \mathsf{hasName}(z,s) \rightarrow \mathsf{hasBlackPlayer}(x,s)$$

Take-homes

- Data integration and reuse – and data management – is a still growing in importance.
- Reuse of data is much easier if data is published according to well-designed schemas, in the form of modular ontologies.
- Best practices for modular ontology

Thanks!

References

Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnathi, Valentina Presutti (eds.), Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web. IOS Press/AKA Verlag, 2016/2017. To appear.

Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, Foundations of Semantic Web Technologies, CRC/Chapman & Hall, 2010

David Carral, Michelle Cheatham, Sunje Dallmeier-Tiessen, Patricia Herterich, Michael D. Hildreth, Pascal Hitzler, Adila Krisnadhi, Kati Lassila-Perini, Elizabeth Sexton-Kennedy, Gordon Watts, Charles Vardeman, An Ontology Design Pattern for Particle Physics Analysis. In: Eva Blomqvist et al. (eds.), Proceedings of the 6th Workshop on Ontology and Semantic Web Patterns (WOP 2015) co-located with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, Pensylvania, USA, October 11, 2015. CEUR Workshop Proceedings 1461, CEUR-WS.org, 2015.

References

Gordon Watts, Pascal Hitzler, Charles Vardeman, David Carral, The Detector Final State pattern: Using the Web Ontology Language to describe a Physics Analysis. ACAT 2016, 18-22 January 2016, Valpariso, Chile.

Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, Sebastian Rudolph, OWL 2 Web Ontology Language: Primer (Second Edition). W3C Recommendation, 11 December 2012.

Michelle Cheatham, Pascal Hitzler, String Similarity Metrics for Ontology Alignment. In: H. Alani et al. (eds.), The Semantic Web - ISWC 2013. 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II. Lecture Notes in Computer Science Vol. 8219, Springer, Heidelberg, 2013, pp. 294-309.

References

Cheatham, Oliveira, Pesquita, McCurdy, The Properties of Property Alignment on the Semantic Web, under review

A. Gangemi. Ontology design patterns for semantic web content. In Y. Gil et al. (eds), The Semantic Web - ISWC 2005 – 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings, volume 3729 of Lecture Notes in Computer Science, pages 262-276. Springer, 2005

Eva Blomqvist, Pascal Hitzler, Krzysztof Janowicz, Adila Krisnadhi, Thomas Narock, Monika Solanki, Considerations regarding Ontology Design Patterns. Semantic Web 7 (1) 1-7.

References

Adila A. Krisnadhi, Yingjie Hu, Krzysztof Janowicz, Pascal Hitzler, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Tim Finin, Peng Ji, Matthew Jones, Nazifa Karima, Audrey Mickle, Tom Narock, Margaret O'Brien, Lisa Raymond, Adam Shepherd, Mark Schildhauer, Peter Wiebe, The GeoLink Modular Oceanography Ontology. In: Marcelo Arenas, Óscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul T. Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, Steffen Staab (eds.), The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II. Lecture Notes in Computer Science 9367, Springer, Heidelberg, 2015, 301-309.

References

Víctor Rodríguez-Doncel, Adila A. Krisnadhi, Pascal Hitzler, Michelle Cheatham, Nazifa Karima, Reihaneh Amini, Pattern-Based Linked Data Publication: The Linked Chess Dataset Case. In: Olaf Hartig, Juan Sequeda, Aidan Hogan (eds.), Proceedings of the 6th International Workshop on Consuming Linked Data co-located with 14th International Semantic Web Conference (ISWC 2105), Bethlehem, Pennsylvania, US, October 12th, 2015. CEUR Workshop Proceedings 1426, CEUR-WS.org, 2015.

Adila Krisnadhi, Ontology Pattern-Based Data Integration. Dissertation, Department of Computer Science and Engineering, Wright State University, 2015.