Knowledge Modeling for Data Sharing, Integration, and Reuse

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

Our Lab

Data Semantics (DaSe) Lab

Wright State University, Dayton, Ohio, USA

Directors: Michelle Cheatham & Pascal Hitzler

Post-Doc: Adila Krisnadhi

PhD students:
- Reihaneh Amini
- David Carral
- Amit Joshi
- Nazifa Karima
- Raghava Mutharaju
- Stella Sam
- Md. Kamruzzaman Sarker
- Cong Wang
- Lu Zhou

Master students:
- Pawel Grzebala
- Kylyn Magee
- Brooke McCurdy
- Jacob Miracle
- Chandan Patel
- Cogan Shimizu

Our Lab

Current focus topics:

Semantic Web & Ontologies:

- **ontology modeling**
- **ontology design patterns**
- **ontology and data alignment**
- **semantic web languages**

Knowledge Representation and Reasoning

- **use of formal semantics in applications**
- **logical foundations**
- **efficient reasoning algorithms**

data and information integration

data security and privacy

applications in the sciences and elsewhere

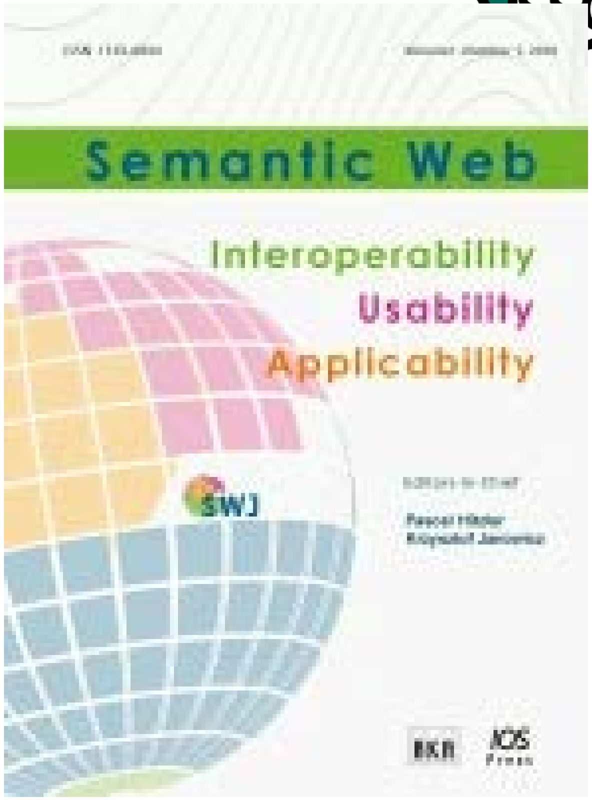
Data Science and Security Cluster at WSU

Membership includes 7 faculty and over 40 graduate and undergraduate students across 5 distinct research labs:

- Advanced Visual Data Analysis (AViDA), directed by Thomas Wischgoll.
- Bioinformatics Research Group (BiRG), directed by Travis Doom and Mike Raymer
- Cybersecurity Lab, directed by Junjie Zhang
- Data Semantics (DaSe) Lab, directed by Michelle Cheatham and Pascal Hitzler
- Web and Complex Systems (WaCS) Lab, directed by Derek Doran

Semantic Web journal

- EiCs: Pascal Hitzler
  Krzysztof Janowicz
- Funded 2010
- SCImago ranked us 18th worldwide in Computer Science in 2014



- 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/**

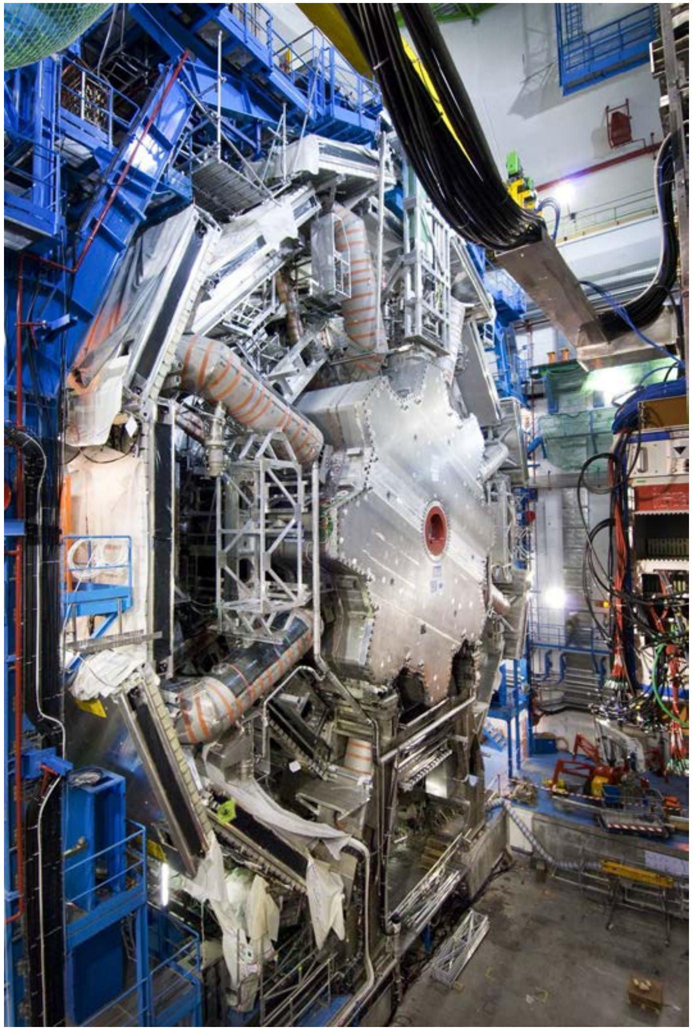
Some Primary Goals

- data/information *sharing* (should be simple!)
- data/information *discovery* (fine-grained, intelligent search)
- data/information *reuse* (preferably by software agents)
- data/information *integration* (on-the-fly, heterogeneous data)

A Use Case Description

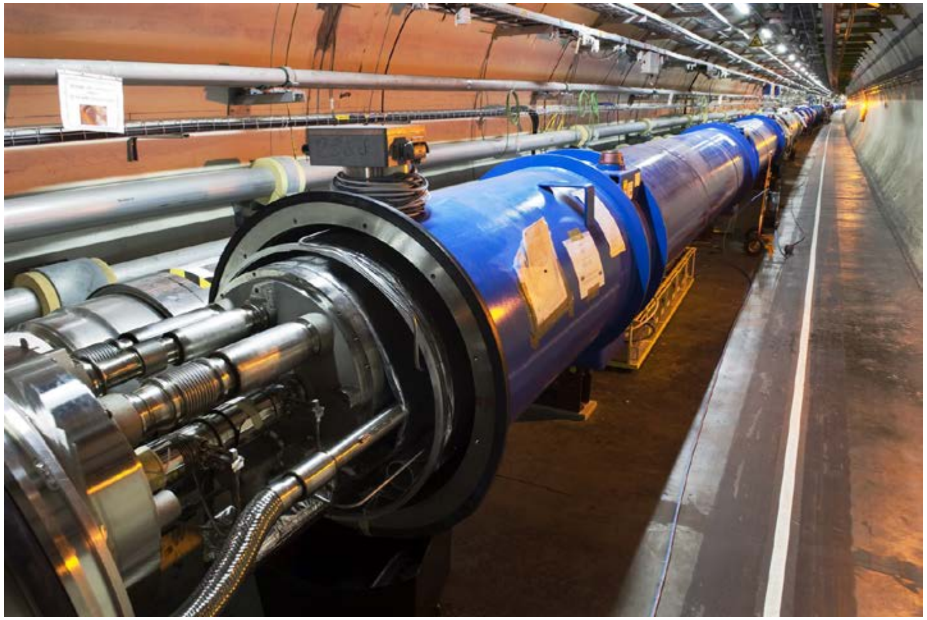
DaSe Lab

SSC



Large Hadron Collider (LHC) at CERN

experiments: ALICE
ATLAS
CMS
LHCb



Photos: ATLAS Experiment © 2014 CERN

WRIGHT STATE UNIVERSITY

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 $p_T > 40$ 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

EarthCube:

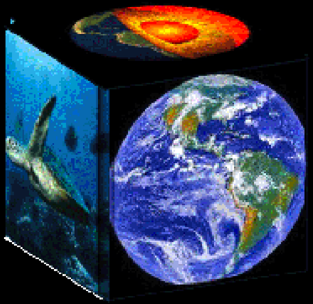
Developing a Community-Driven Data and Knowledge Environment for the Geosciences

“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 requires

- information integration
- interoperability
- conceptual modeling
- intelligent search
- data-model intercomparison
- data publishing support



Semantic Web studies

- information integration
- interoperability
- conceptual modeling
- intelligent search
- data-model intercomparison
- data publishing support

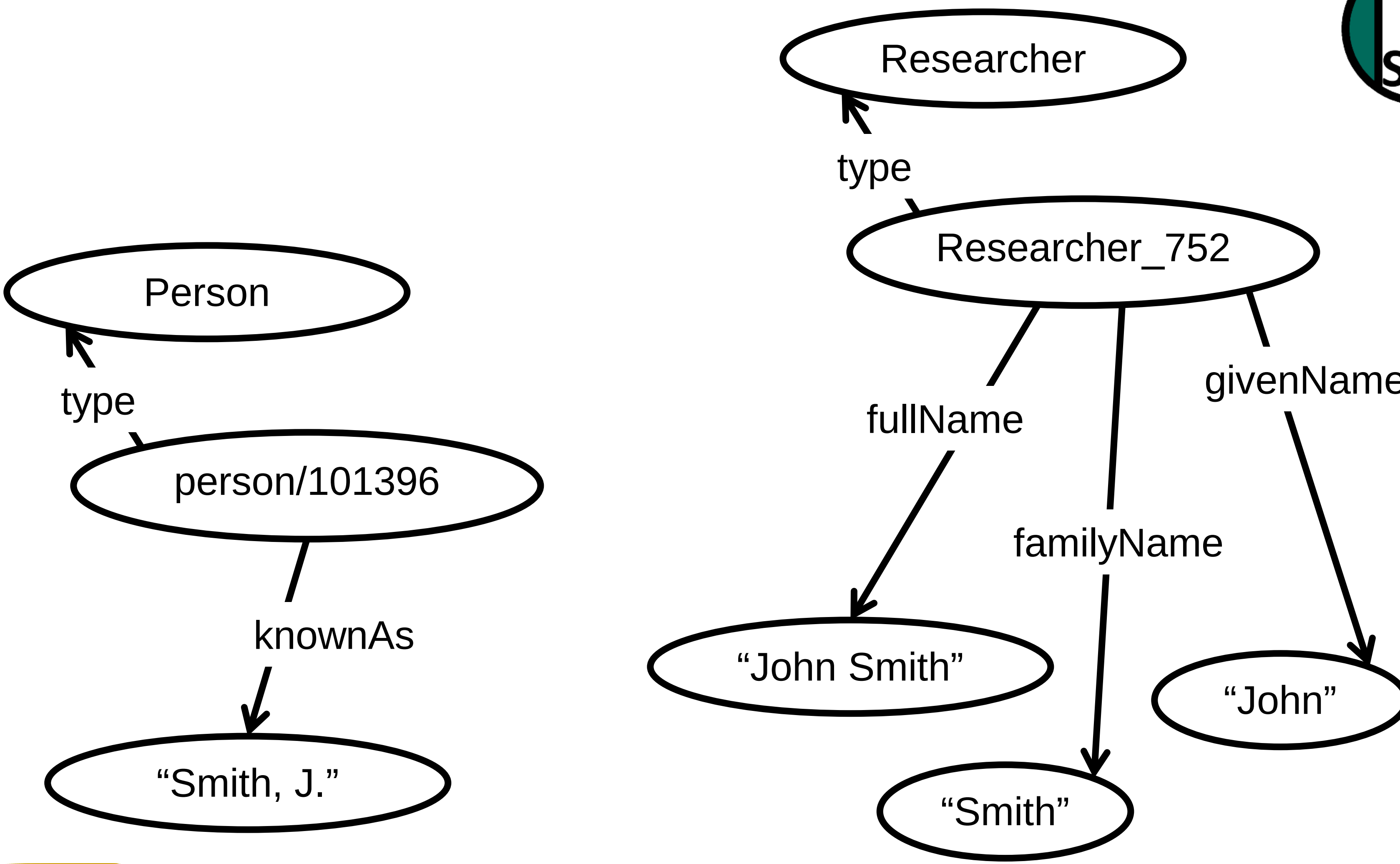
EarthCube Challenges

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**

More about this later.

Regarding Integration

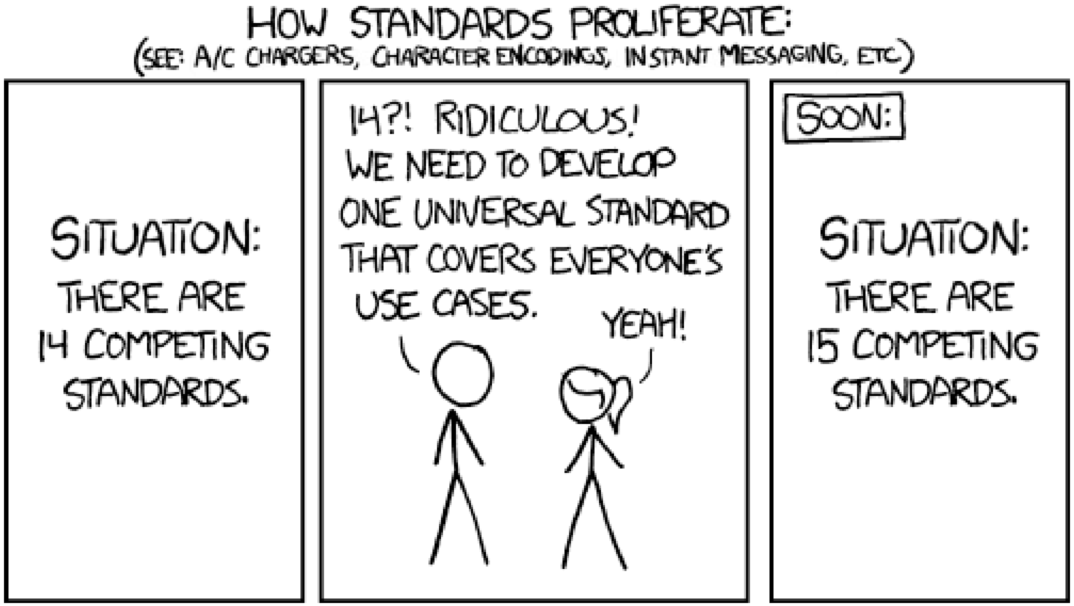


Use of standards

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.

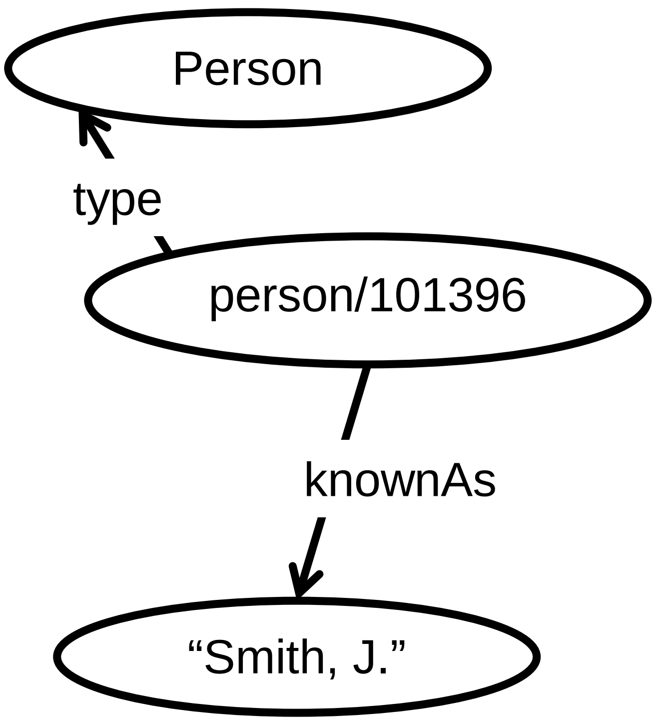
Serializations: several, some more human-readable, some more machine-readable. For the latter, mostly using RDF/XML.

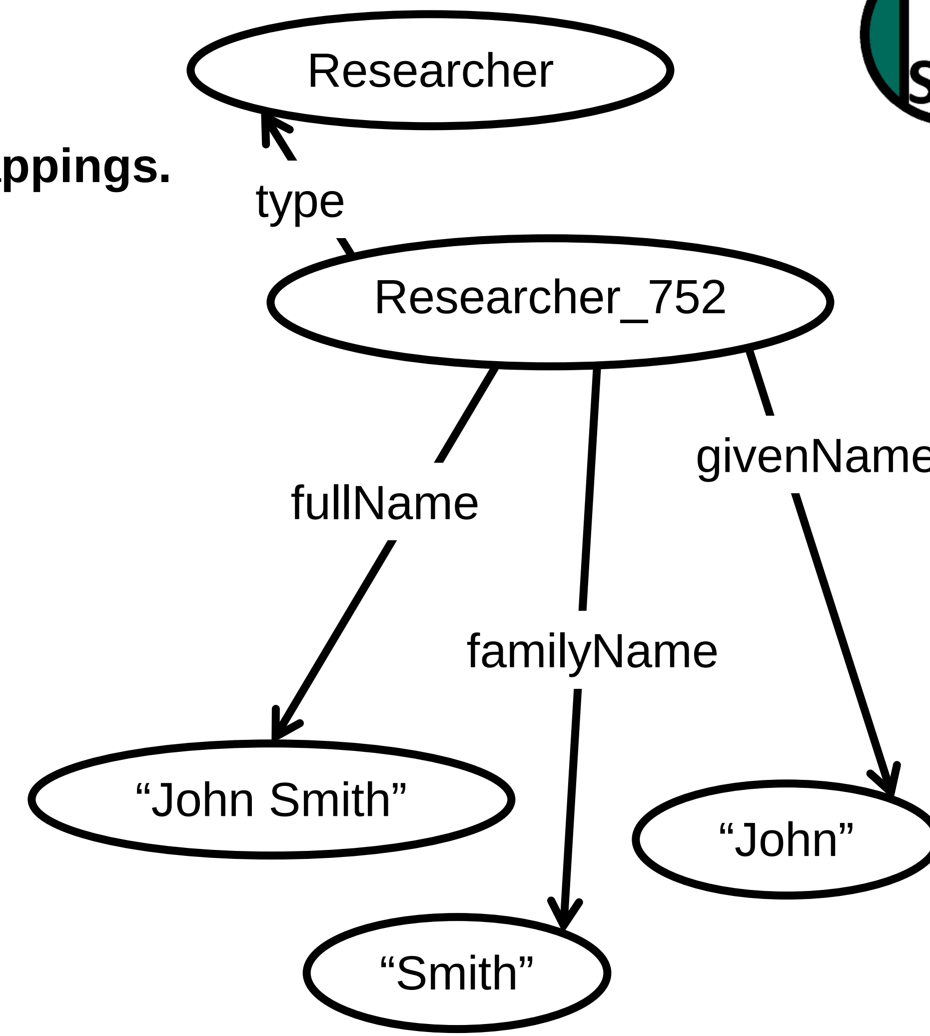
[W3C 2012]

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]

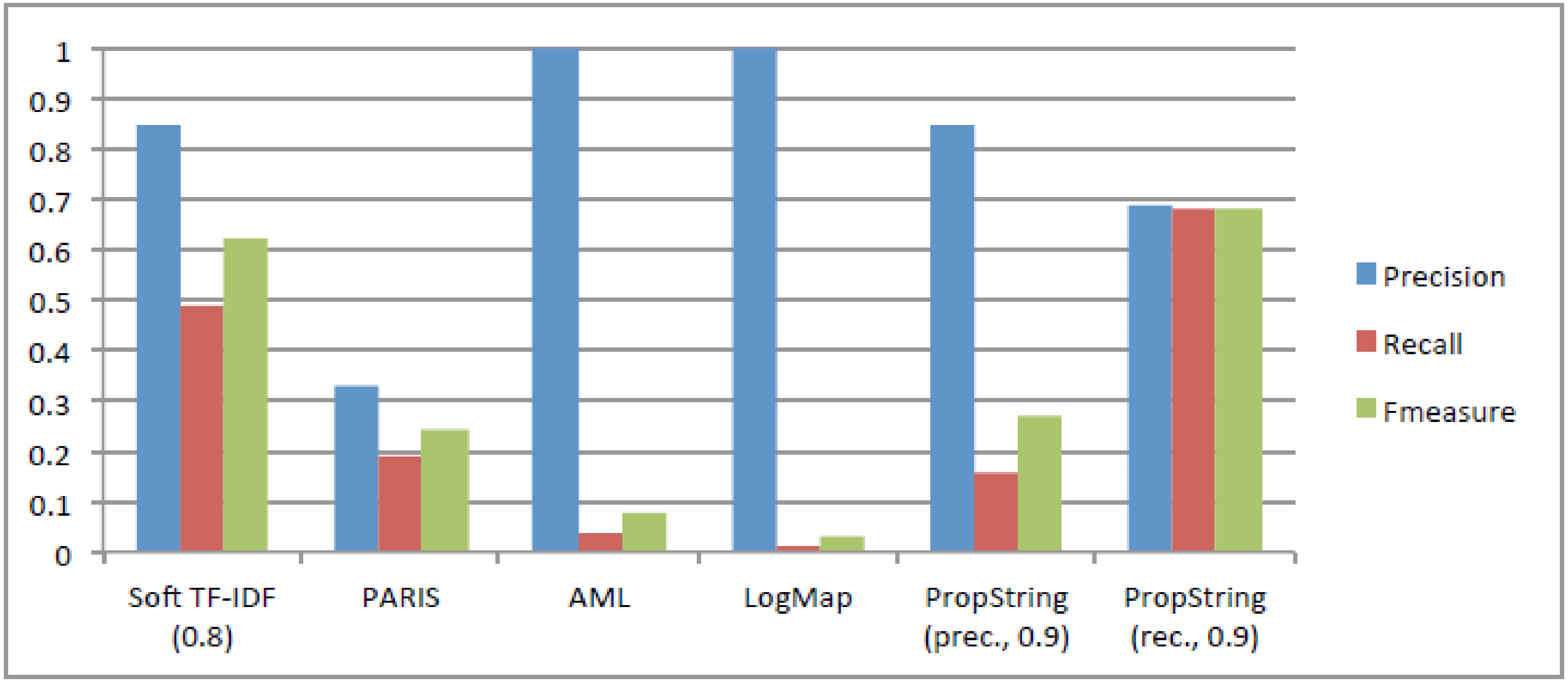


Fig. 1 Results of the YAGO-DBPedia alignment task

Dealing with bad mappings

[JIST 2014, ISWC 2015]

$$a{:}hasWife \sqsubseteq a{:}hasSpouse \quad (1)$$

$$symmetric(a{:}hasSpouse) \quad (2)$$

$$\exists a{:}hasSpouse.a{:}Female \sqsubseteq a{:}Male \quad (3)$$

$$\exists a{:}hasSpouse.a{:}Male \sqsubseteq a{:}Female \quad (4)$$

$$a{:}hasWife(\text{a:john}, a{:}mary) \quad (5)$$

$$a{:}Male(a{:}john) \quad (6)$$

$$a{:}Female(a{:}mary) \quad (7)$$

$$a{:}Male \sqcap a{:}Female \sqsubseteq \bot \quad (8)$$

$$symmetric(b{:}hasSpouse) \quad (9)$$

$$b{:}hasSpouse(b{:}mike, b{:}david) \quad (10)$$

$$b{:}Male(b{:}david) \quad (11)$$

$$b{:}Male(b{:}mike) \quad (12)$$

$$b{:}Female(b{:}anna) \quad (13)$$

$$a{:}hasSpouse \equiv b{:}hasSpouse \quad (14)$$

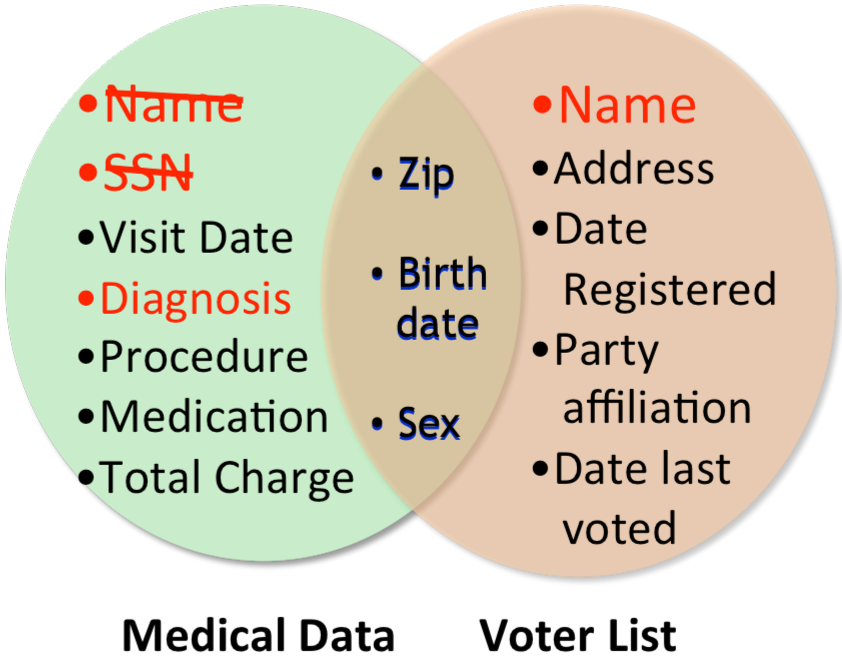
$$a{:}Male \equiv b{:}Male \quad (15)$$

$$a{:}Female \equiv b{:}Female \quad (16)$$

Fig. 1. Running example with selected axioms.

Data Integration and Privacy

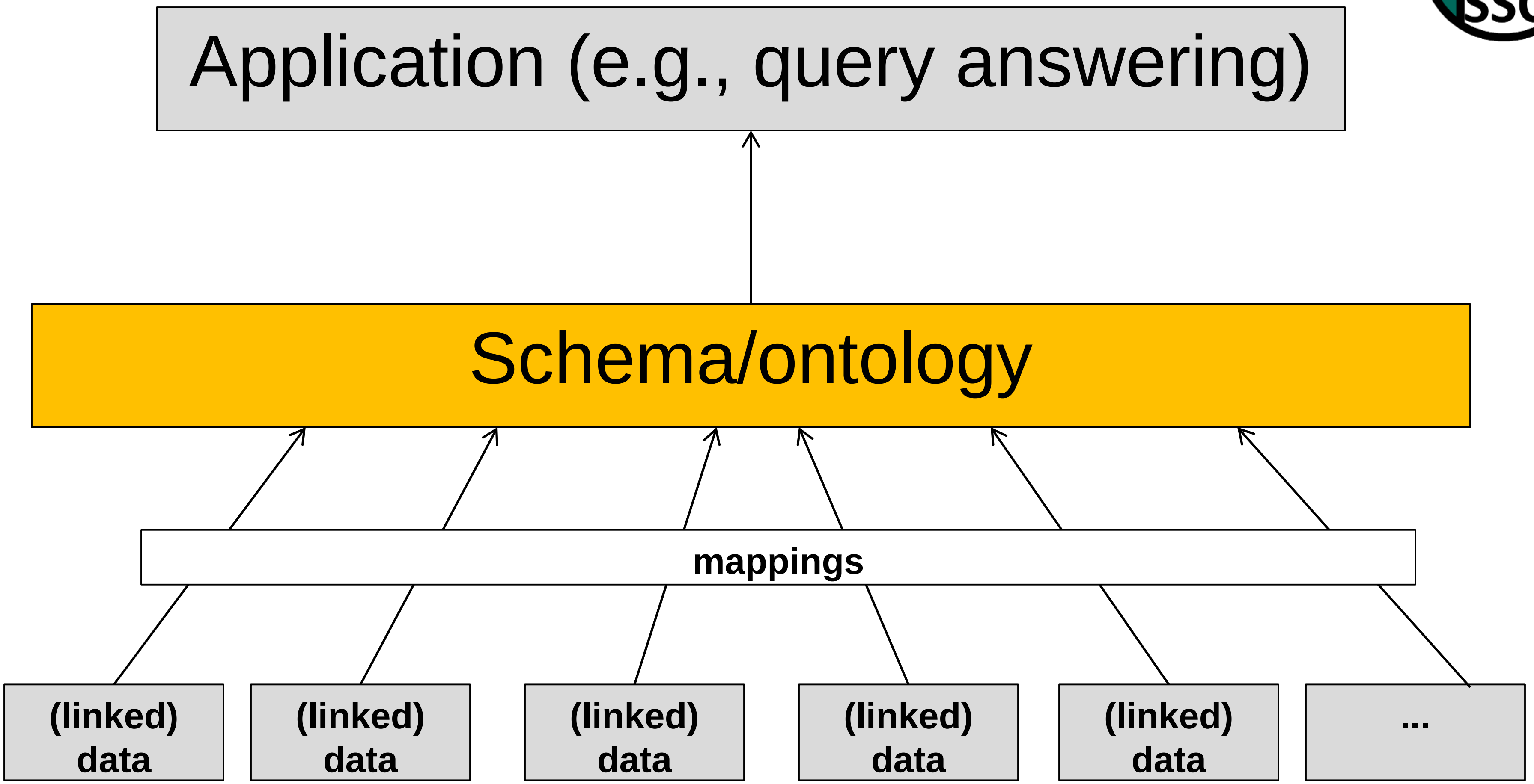
- **Research based on anonymized data can yield advances in medicine, economics, and more…**
- **But privacy must be respected**
- **Most privacy breaches of anonymized data occurs when two or more datasets are combined**
- **We are researching the potential for Semantic Web technologies to facilitate de-anonymization attacks**



[Cheatham 2016, in preparation]

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**

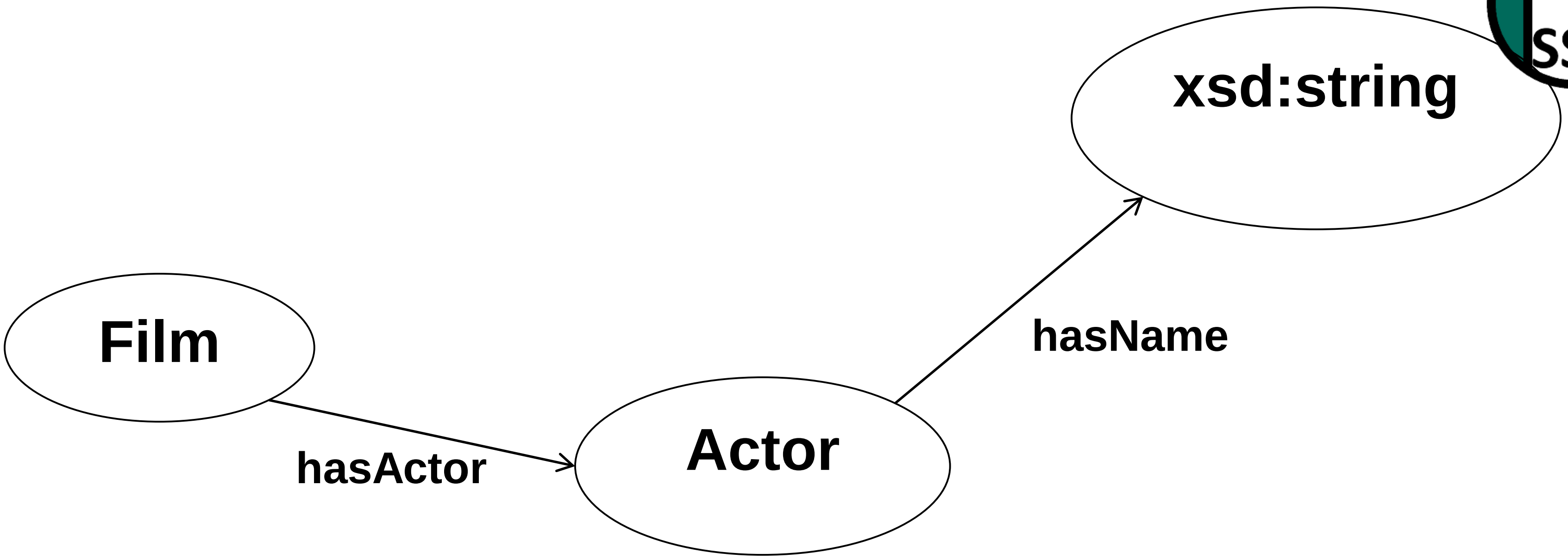
Three modeling principles

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

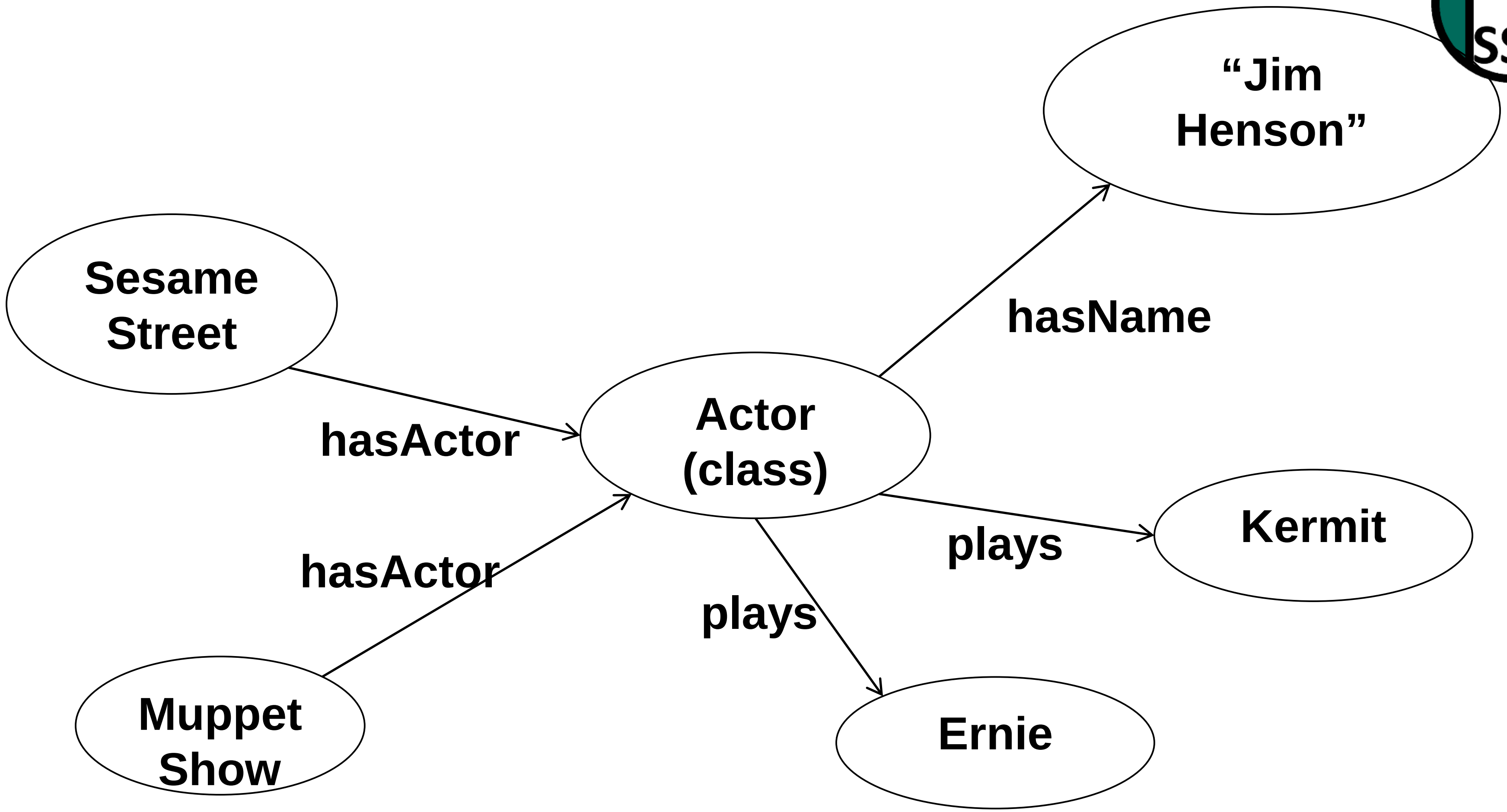
2. **Modularize your ontology to make it manageable and flexible (e.g. by modifying/replacing independent modules, extending with new modules, etc.).**

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

Example from Linked MDB

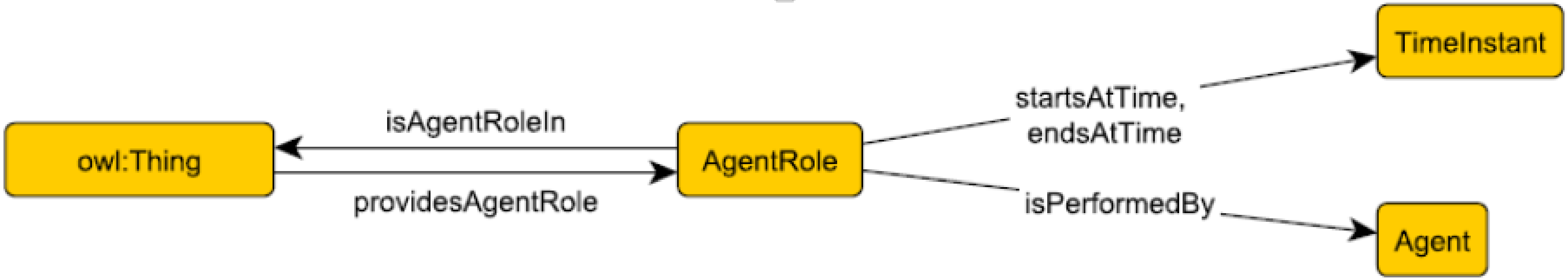


Problem!

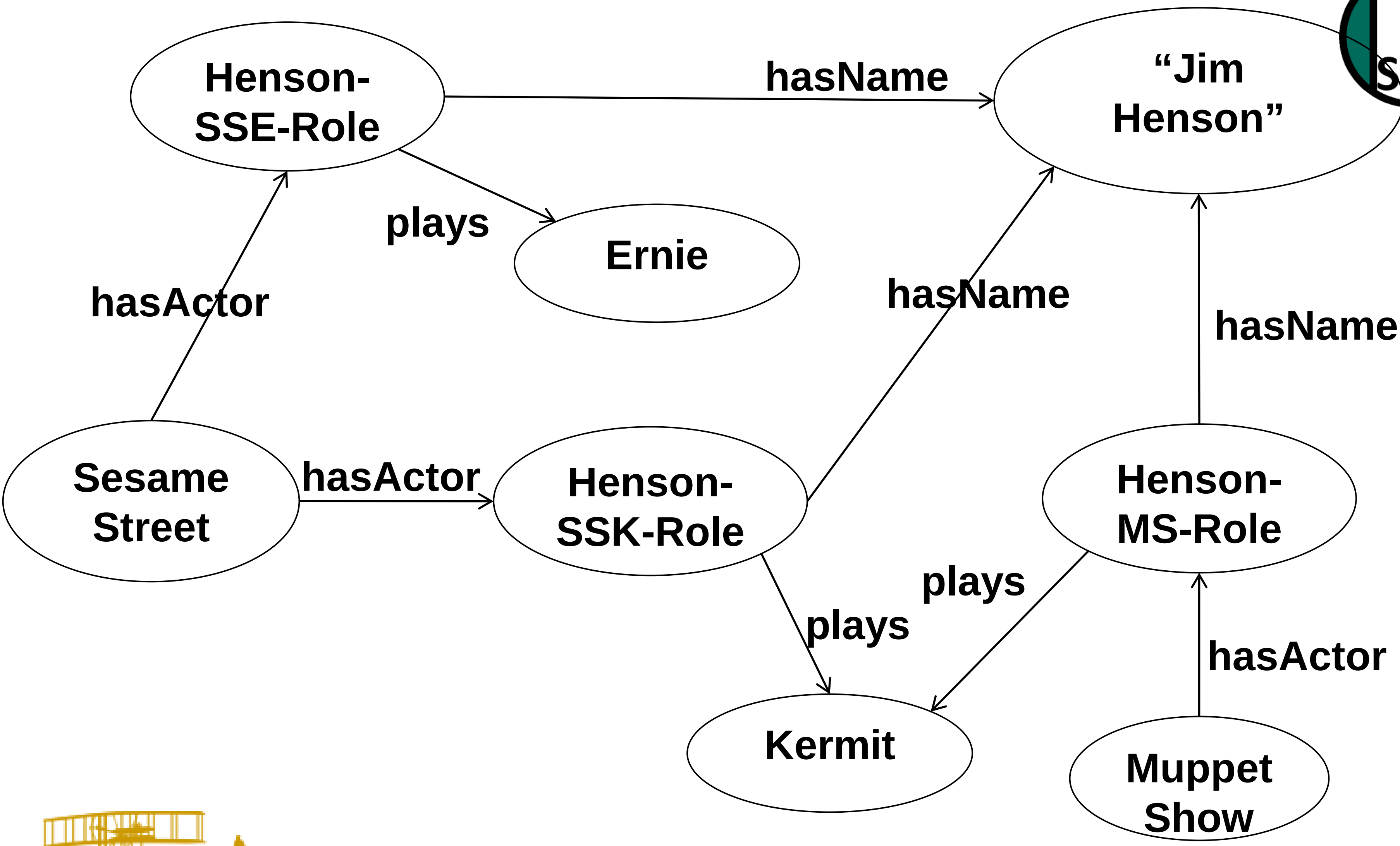


“Jim Henson”
Sesame Street
hasName
Actor (class)
hasActor
Kermit
plays
hasActor
plays
Muppet Show
Ernie

Best Practices: Agent Roles



Solution!



Henson-SSE-Role
hasName
"Jim Henson"
plays
Ernie
hasActor
hasName
hasName
Sesame Street
hasActor
Henson-SSK-Role
Henson-MS-Role
plays
plays
hasActor
Kermit
Muppet Show

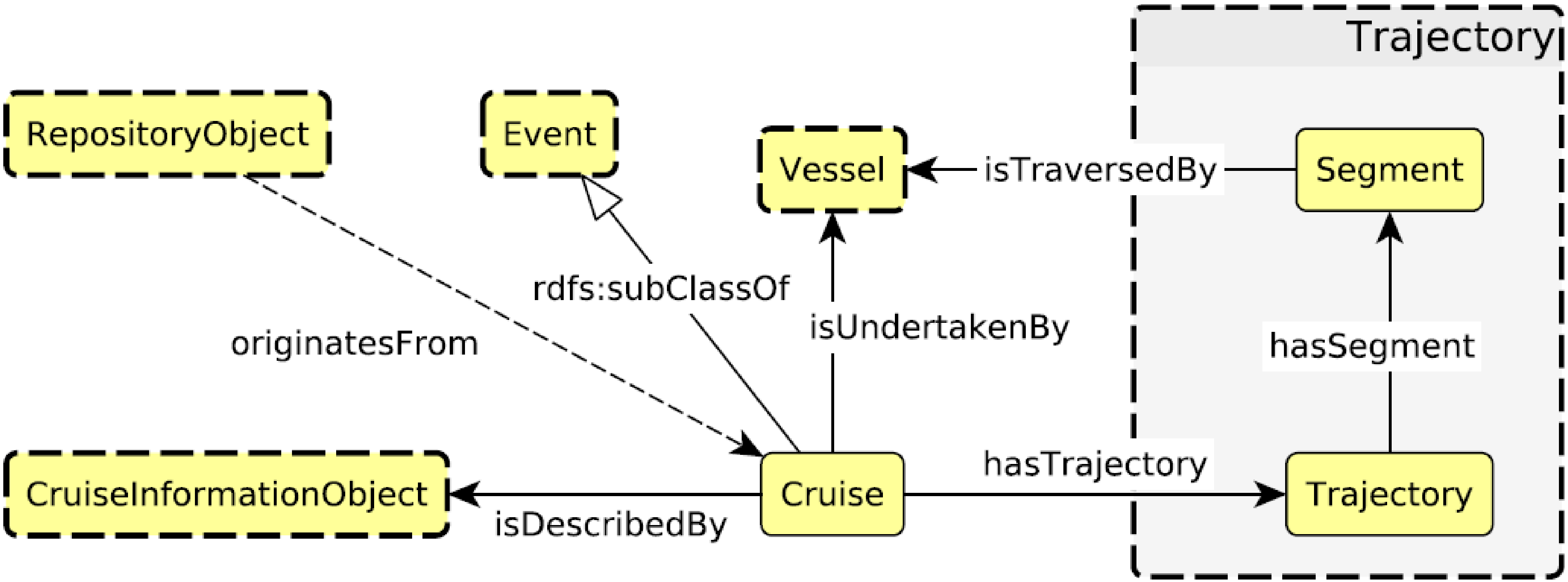
Ontology Design Patterns

"An ontology design pattern 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.

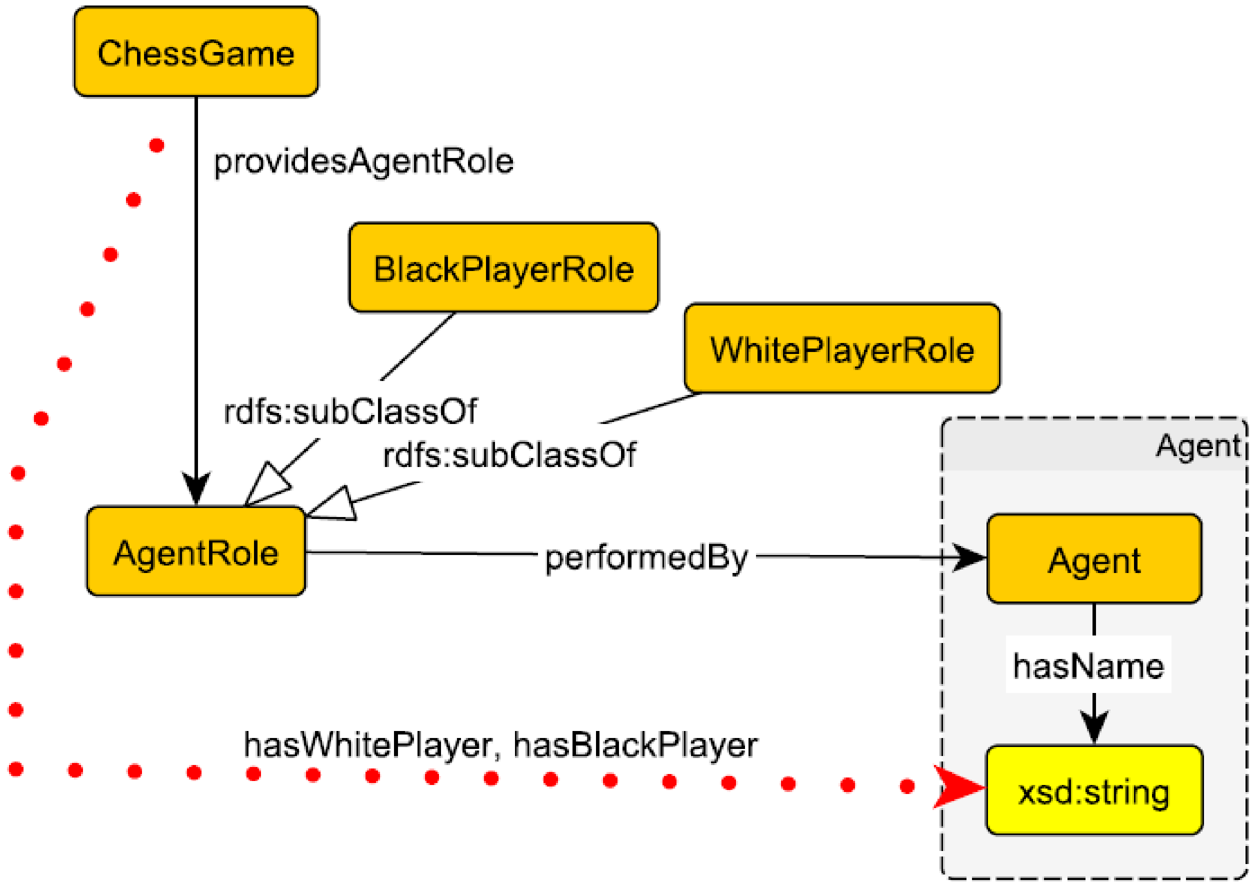
[SWJ 2016]

Modularization



[ISWC 2015]

Simplified views



[COLD 2015; Krisnadhi Dissertation 2015]

Representing Ontologies

The Web Ontology Language (OWL)

- **W3C Standard 2004/2009.**
- **Based on a subset of first-order predicate logic.**
- **Logic is used because its formal semantics is mathematically defined, thus unambiguous.**
- **Logic furthermore enables reasoning, the derivation of deductive inferences based on given data, which can be used to enhance search, for debugging data, etc.**
- **However the complexity of deductive reasoning is high.**

Distributed memory reasoning

- **Controlling communication overhead is central.**
  System for the logic EL: [ESWC2015]

	GO	SNOMED	SNOMEDx2	SNOMEDx3	SNOMEDx5	Traffic
Before	87,137	1,038,481	2,076,962	3,115,443	5,192,405	7,151,328
After	868,996	14,796,555	29,593,106	44,389,657	73,982,759	21,840,440

Table 2: Number of axioms, before and after classification, in ontologies.

Ontology	ELK	jCEL	Snorocket	Pellet	HermiT	FaCT++
GO	23.5	57.4	40.3	231.4	91.7	367.89
SNOMED	31.8	126.6	52.34	620.46	1273.7	1350.5
SNOMEDx2	77.3	OOM[a]	OOM[a]	OOM[a]	OOM[a]	OOM[a]
SNOMEDx3	OOM[a]	OOM[a]	OOM[a]	OOM[a]	OOM[a]	OOM[a]
SNOMEDx5	OOM[a]	OOM[a]	OOM[a]	OOM[a]	OOM[a]	OOM[a]
Traffic	OOM[b]	OOM[c]	OOM[c]	OOM[b]	OOM[b]	OOM[c]

Table 3: Classification times in seconds. OOM[a]: reasoner runs out of memory. OOM[b]: reasoner runs out of memory during incremental classification. OOM[c]: ontology too big for OWL API to load in memory.

Distributed memory reasoning

Ontology	8 nodes	16 nodes	24 nodes	32 nodes	64 nodes
GO	134.49	114.66	109.46	156.04	137.31
SNOMED	544.38	435.79	407.38	386.00	444.19
SNOMEDx2	954.17	750.81	717.41	673.08	799.07
SNOMEDx3	1362.88	1007.16	960.46	928.41	1051.80
SNOMEDx5	2182.16	1537.63	1489.34	1445.30	1799.13
Traffic	60004.54	41729.54	39719.84	38696.48	34200.17

Table 4: Classification time (in seconds) of `DistEL`

[ESWC 2015]

Horn-SRIQ reasoning

- Horn-SRIQ: a significant subset of OWL.
- We carried over and refined methods from reasoning with existential rules.
- As a consequence, we improved on state of the art reasoners by an order of magnitude.

[Carral et al, 2016, under review]

Take-homes

- Data integration and reuse – and data management – is a still growing in importance.
- Progress in this area requires advances and applications from many computer science and related disciplines, including logic-based knowledge representation and automated reasoning, machine learning and data mining, natural language processing and linguistics, cognitive science, etc.
- At DaSeLab vertical research, from fundamentals to applications, using multiple methods, is pursued, with some emphasis on modeling pragmatics, logical foundations, and ontology alignment.

DaSe Lab

SSC

Thanks!

References

Hitzler, Krötzsch, 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.

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.

References

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.

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

References

Kunal Sengupta, Pascal Hitzler, Krzysztof Janowicz, Revisiting default description logics – and their role in aligning ontologies. In: T. Supnithi et al. (eds.), Semantic Technology, 4th Joint International Conference, JIST 2014, Chiang Mai, Thailand, November 9-11, 2014. Revised Selected Papers. Lecture Notes in Computer Science, Vol. 8943, Springer, Heidelberg, 2015, pp. 3-18.

Kunal Sengupta, Pascal Hitzler, Towards Defeasible Mappings for Tractable Description Logics. In: Marcelo Arenas et al. (eds.), The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I. Lecture Notes in Computer Science 9366, Springer, Heidelberg, 2015, pp. 237-252.

Cheatham 2016, on data integration and privacy, in preparation.

References

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.

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.

References

Raghava Mutharaju, Pascal Hitzler, Prabhaker Mateti, Freddy Lecue, Distributed and Scalable OWL EL Reasoning. In: Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, Antoine Zimmermann, The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings. Lecture Notes in Computer Science Vol. 9088, Springer, Heidelberg, 2015, pp. 88-103.

David Carral et al., Novel Acyclicity Notions for Horn Ontologies Improving Efficiency of Conjunctive Query Answering by an Order of Magnitude. Under review.