Research at the Data Semantics (DaSe) Laboratory: Data Management, Artificial Intelligence, and Applications

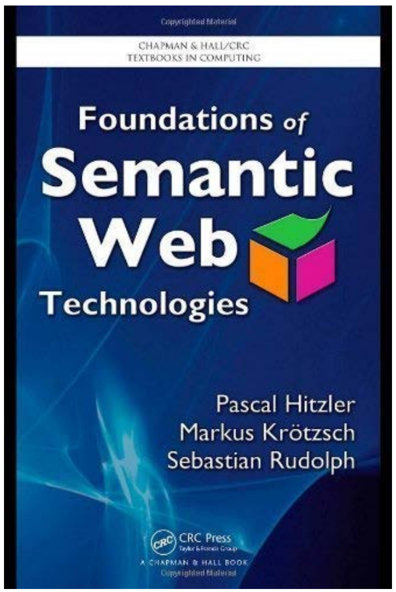
Pascal Hitzler
Data Semantics Laboratory (DaSe Lab)
Kansas State University

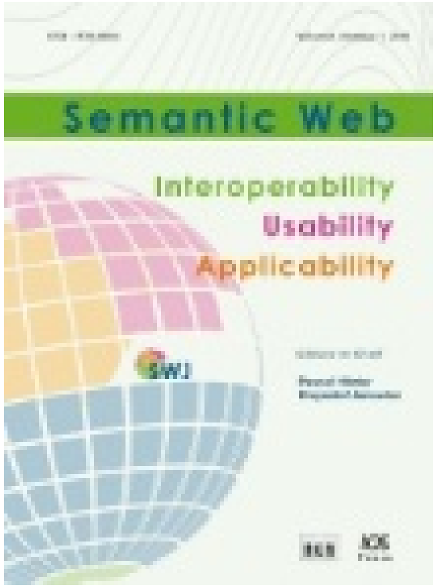
http://www.daselab.org

About me

DaSe Lab

- **I'm new here (joined 2019 as senior hire)**
- **I brought most of my lab (7 PhD students)**





Where (some) PhD students went

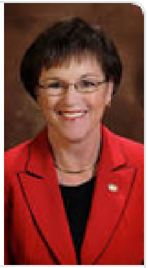
- Industry
  - Amazon
  - IBM
  - Apple
  - GE Global Research
- Academia
  - TU Dresden, Germany (several)
  - IIT Delhi, India
  - Universitas Indonesia, Jakarta
  - Wright State University, USA
- Elsewhere
  - UN Headquarters, New York

Knowledge Graphs

Google Knowledge Graph

Laura Kelly

Governor of Kansas



Laura Kelly is an American politician serving as the 48th governor of Kansas since 2019. A member of the Democratic Party, she represented the 18th district in the Kansas Senate from 2005 to 2019. Kelly ran for governor in the 2018 election and defeated the Republican nominee, Kansas Secretary of State Kris Kobach. Wikipedia

Born: January 24, 1950 (age 69 years), New York, NY

Spouse: Ted Daughety

Party: Democratic Party

Office: Governor of Kansas since 2019

Education: Indiana University, Bradley University, Indiana University Bloomington

Children: Kathleen Daughety, Molly Daughety

hasEducation

Indiana University

iu.edu

Indiana University is a multi-campus public university system in the state of Indiana, United States. Indiana University has a combined student body of more than 110,000 students, which includes approximately 46,000 students enrolled at the Indiana University Bloomington campus. Wikipedia

Mascot: Referred to as "The Hoosiers"

Endowment: 1.986 billion USD

Students: 110,436 university-wide

President: Michael McRobbie

Academic staff: 8,733 university-wide

Subsidiaries: Indiana University Bloomington, MORE

hasPresident

Michael McRobbie

President of Indiana University

president.iu.edu

Michael Alexander McRobbie AO is an Australian-American computer scientist, educator and academic administrator. He became the eighteenth president of Indiana University on July 1, 2007. Wikipedia

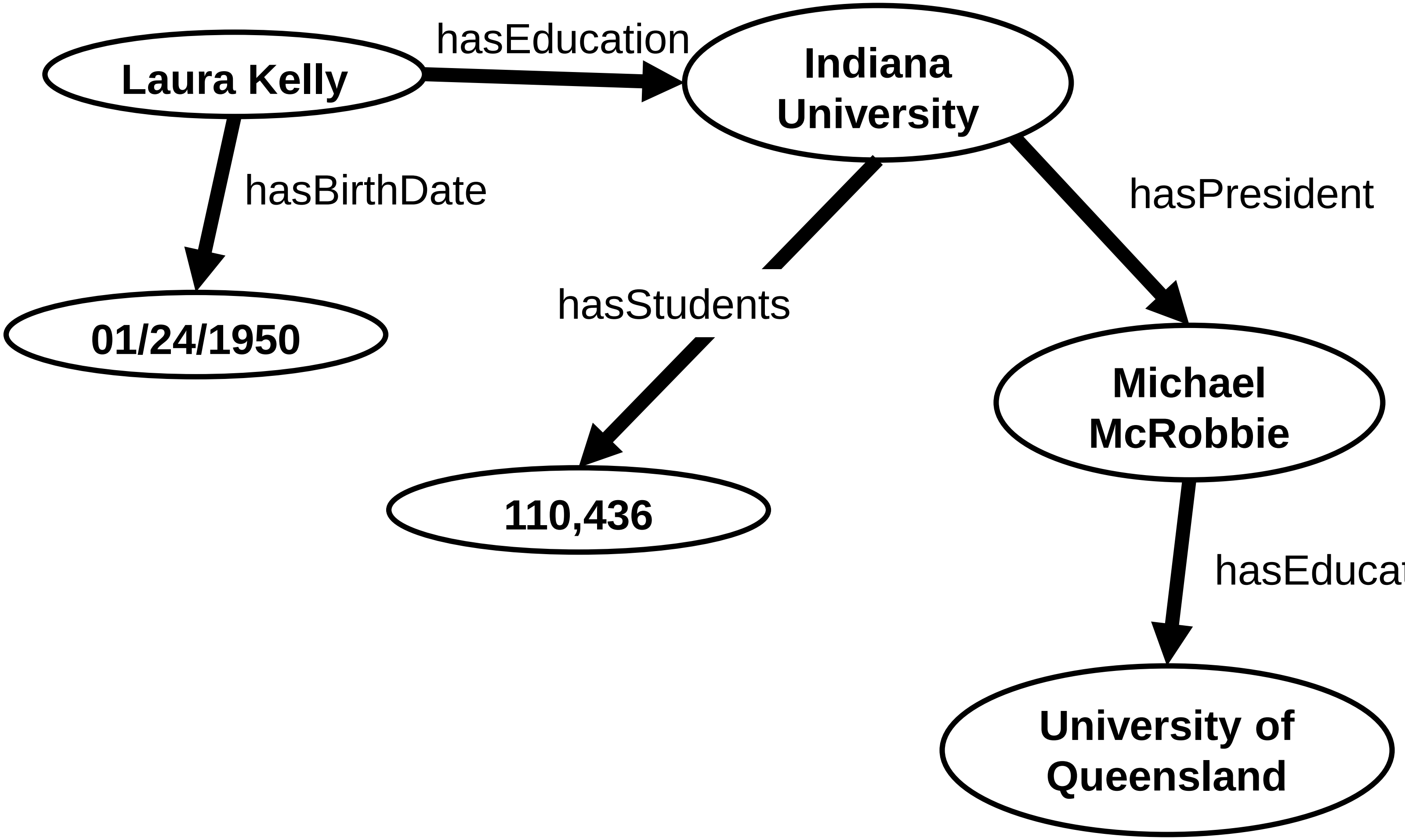
Born: October 11, 1950 (age 69 years), Melbourne, Australia

Spouse: Laurie Burns (m. 2005)

Education: The Australian National University, The University of Queensland

Books: Automated Theorem-proving in Non-classical Logics, Automated Deduction - Cade-13

Knowledge Graphs



Schema.org

- **Collaboratively launched in 2011 by Google, Microsoft, Yahoo, Yandex.**
  2011: 297 classes, 187 relations
  2015: 638 classes, 965 relations
- **Simple schema, request to web site providers to annotate their content with schema.org markup.**
  Promise: They will make better searches based on this.
- **2015: 31.3% of Web pages have schema.org markup, on average 26 assertions per page.**

Ramanathan V. Guha, Dan Brickley, Steve Macbeth: Schema.org: Evolution of Structured Data on the Web. ACM Queue 13(9): 10 (2015)

- TrainTrip
- Organization
    - Airline
    - Corporation
    - EducationalOrganization
        - CollegeOrUniversity
        - ElementarySchool
        - HighSchool
        - MiddleSchool
        - Preschool
        - School
    - GovernmentOrganization
    - LocalBusiness
        - AnimalShelter
        - AutomotiveBusiness
            - AutoBodyShop
            - AutoDealer
            - AutoPartsStore
            - AutoRental
            - AutoRepair
            - AutoWash
            - GasStation
            - MotorcycleDealer
            - MotorcycleRepair
        - ChildCare
        - Dentist
        - DryCleaningOrLaundry
        - EmergencyService
            - FireStation
            - Hospital
            - PoliceStation
        - EmploymentAgency
        - EntertainmentBusiness
            - AdultEntertainment
            - AmusementPark
            - ArtGallery
            - Casino
            - ComedyClub
            - MovieTheater
            - NightClub
        - FinancialService
            - AccountingService
            - AutomatedTeller
            - BankOrCreditUnion
            - InsuranceAgency
        - FoodEstablishment
            - Bakery
            - BarOrPub
            - Brewery
            - CafeOrCoffeeShop
            - FastFoodRestaurant

WIKIDATA

Main Page | Discussion | Read | View source | View history | Search Wikidata

- Main page
- Community portal
- Project chat
- Create a new item
- Recent changes
- Random item
- Query Service
- Nearby
- Help
- Donate

Print/export

- Create a book
- Download as PDF
- Printable version

In other projects

- Wikimedia Commons
- MediaWiki
- Meta-Wiki
- Wikispecies
- Wikibooks
- Wikinews
- Wikipedia
- Wikiquote
- Wikisource
- Wikiversity
- Wikivoyage
- Wiktionary

Tools

- What links here

open

collaborative

multilingual

free

Welcome to Wikidata

the free knowledge base with 29,012,184 data items that anyone can edit

Introduction • Project Chat • Community Portal • Help

Welcome!

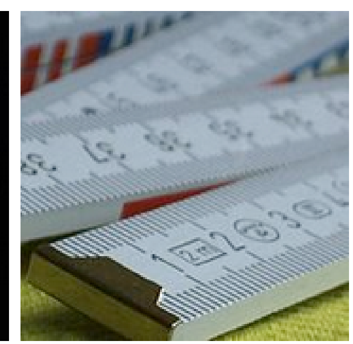
Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

Learn about data

New to the wonderful world of data? Develop and improve your data literacy through content designed to get you up to speed and feeling comfortable with the fundamentals in no time.



Past and current external sponsors

- Federal and State
  - NSF (main source of funding to date) – CISE, GEO and OIA directorates
  - NIST / Department of Commerce
  - USGS
  - Ohio Board of Regents
- Defense
  - DARPA
  - DoD / Air Force
  - AFRL/RY
  - AFOSR
  - Defense Associated Graduate Student Innovation program
- Foundations
  - The Andrew W. Mellon Foundation
  - Henry M. Jackson Foundation
  - Sloan Foundation
- Industry
  - IOS Press (Publisher, several)
  - Lockheed-Martin
- International
  - DFG (Germany)
  - DAAD (Germany)

Plenty of open questions

- What makes good knowledge graphs?
- What are good processes and tools for making them?
- What are strong intelligent algorithms for managing them, including
  - Automatic construction
  - Integration
  - Querying
- How do I make them self-explanatory?
- How do I use them in or with intelligent systems?
- What is the underlying theory/mathematics of the representation languages and (complex) algorithms?

Knowledge Graph Standards

W3C Recommendation

RDF 1.1 Concepts and Abstract Syntax

W3C Recommendation 25 February 2014

This version:
http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/
Latest published version:
http://www.w3.org/TR/rdf11-concepts/
Previous version:
http://www.w3.org/TR/2014/PR-rdf11-concepts-20140109/
Previous Recommendation:
http://www.w3.org/TR/rdf-concepts
Editors:
Richard Cyganiak, DERI, NUI Galway
David Wood, 3 Round Stones
Markus Lanthaler, Graz University of Technology

DaSe Lab

Languages based on formal logic allow for automated (deductive) reasoning.
Corresponding algorithms are mathematically sophisticated and require formal correctness and complexity assessments.

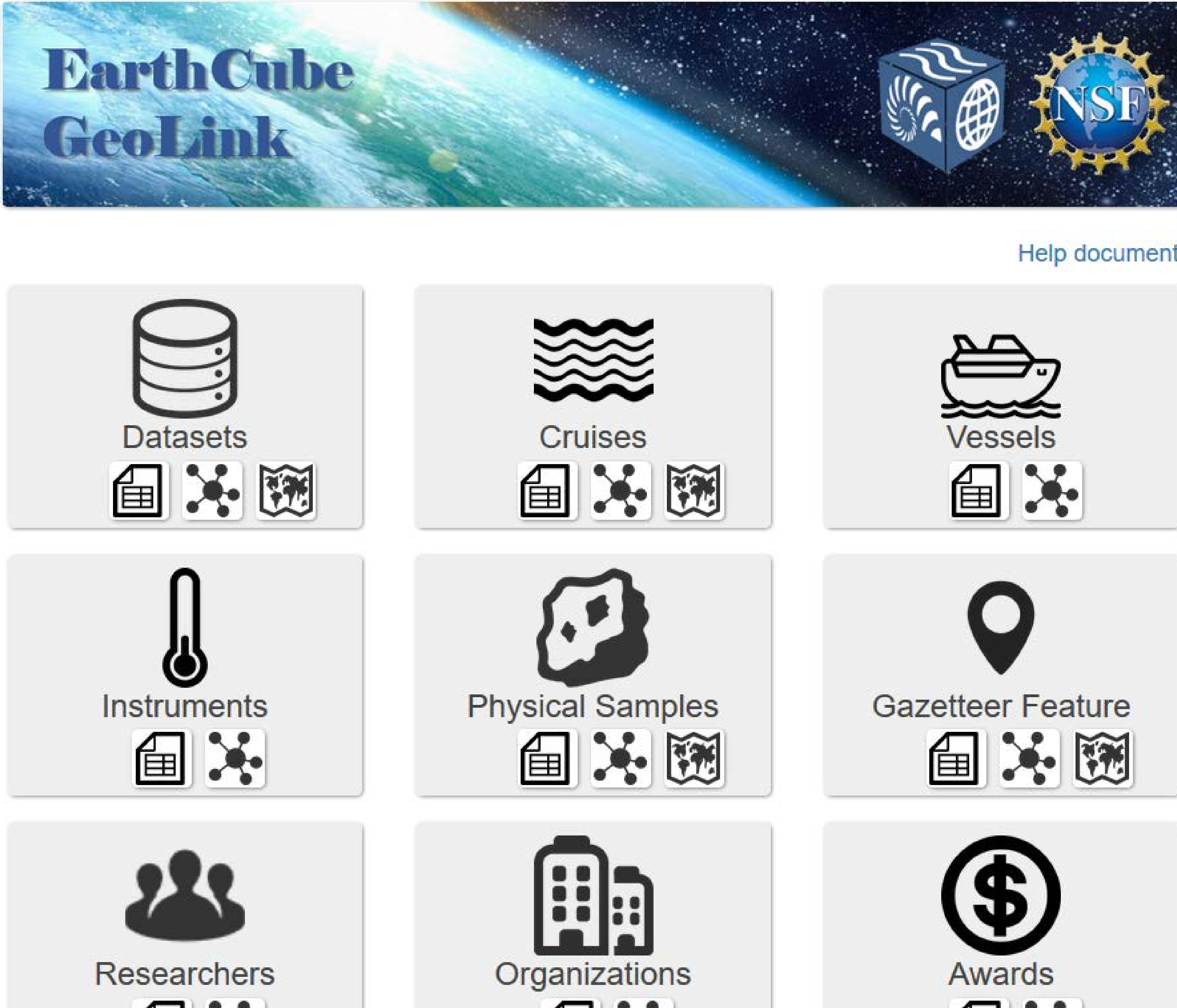
Also:
The Standards need improvements!

W3C Recommendation

OWL 2 Web Ontology Language Primer (Second Edition)

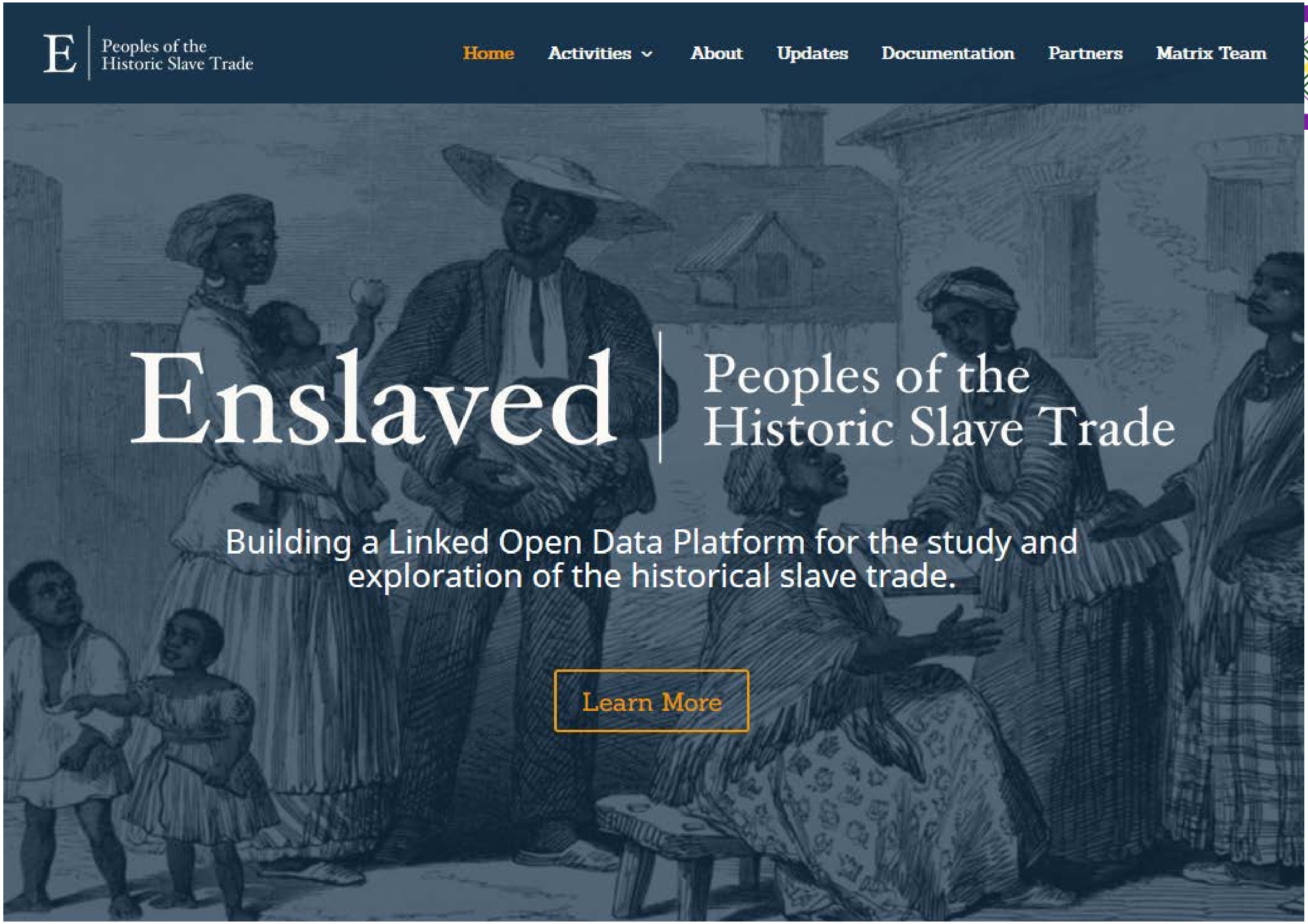
W3C Recommendation 11 December 2012

This version:
http://www.w3.org/TR/2012/REC-owl2-primer-20121211/
Latest version (series 2):
http://www.w3.org/TR/owl2-primer/
Latest Recommendation:
http://www.w3.org/TR/owl-primer
Previous version:
http://www.w3.org/TR/2012/PER-owl2-primer-20121018/
Editors:
Pascal Hitzler, Wright State University
Markus Krötzsch, University of Oxford
Bijan Parsia, University of Manchester
Peter F. Patel-Schneider, Nuance Communications
Sebastian Rudolph, FZI Research Center for Information


EarthCube
GeoLink
NSF
Help document
Datasets
Cruises
Vessels
Instruments
Physical Samples
Gazetteer Feature
Researchers
Organizations
Awards

e Lab

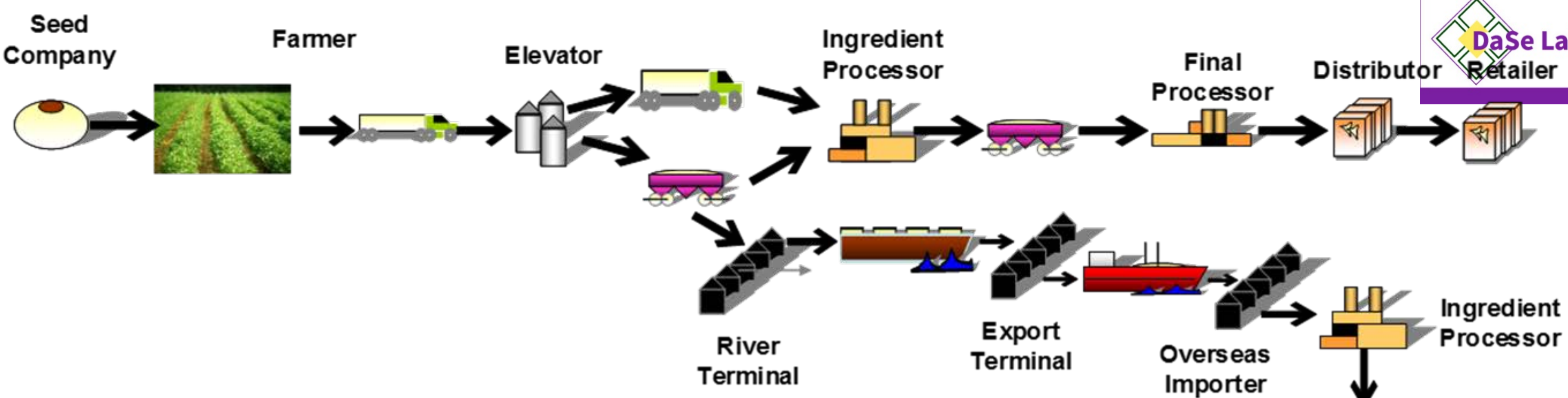
enslaved.org



Recently started project

- National Institute of Standards and Technology (NIST)
- Data Integration for Food Supply Chains
- Focus on grains
- Development of a data model (schema/ontology) and software tool support for integrating data relevant to the traceability of food supply chains.
- Working in close collaboration with NIST.

NIST project



- Grain Traceability
- Data model is central for lowering cost in all parts of the pipeline.

- Tracing along splits and merge.
- Elevators as black boxes.
- Containers may carry contaminants
- …

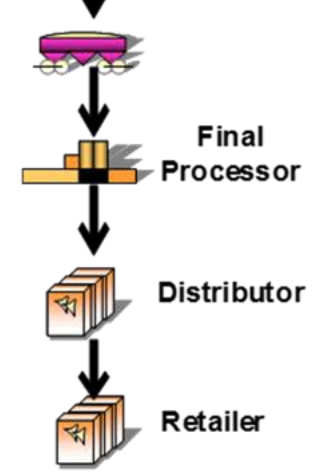


Figure acknowledgement: NIST / Evan Wallace

News Release 19-016

NSF Convergence Accelerator awards bring together scientists, businesses, nonprofits to benefit workers

New projects address some of the most promising areas of research

Convergence Accelerator awards are focused on three areas:

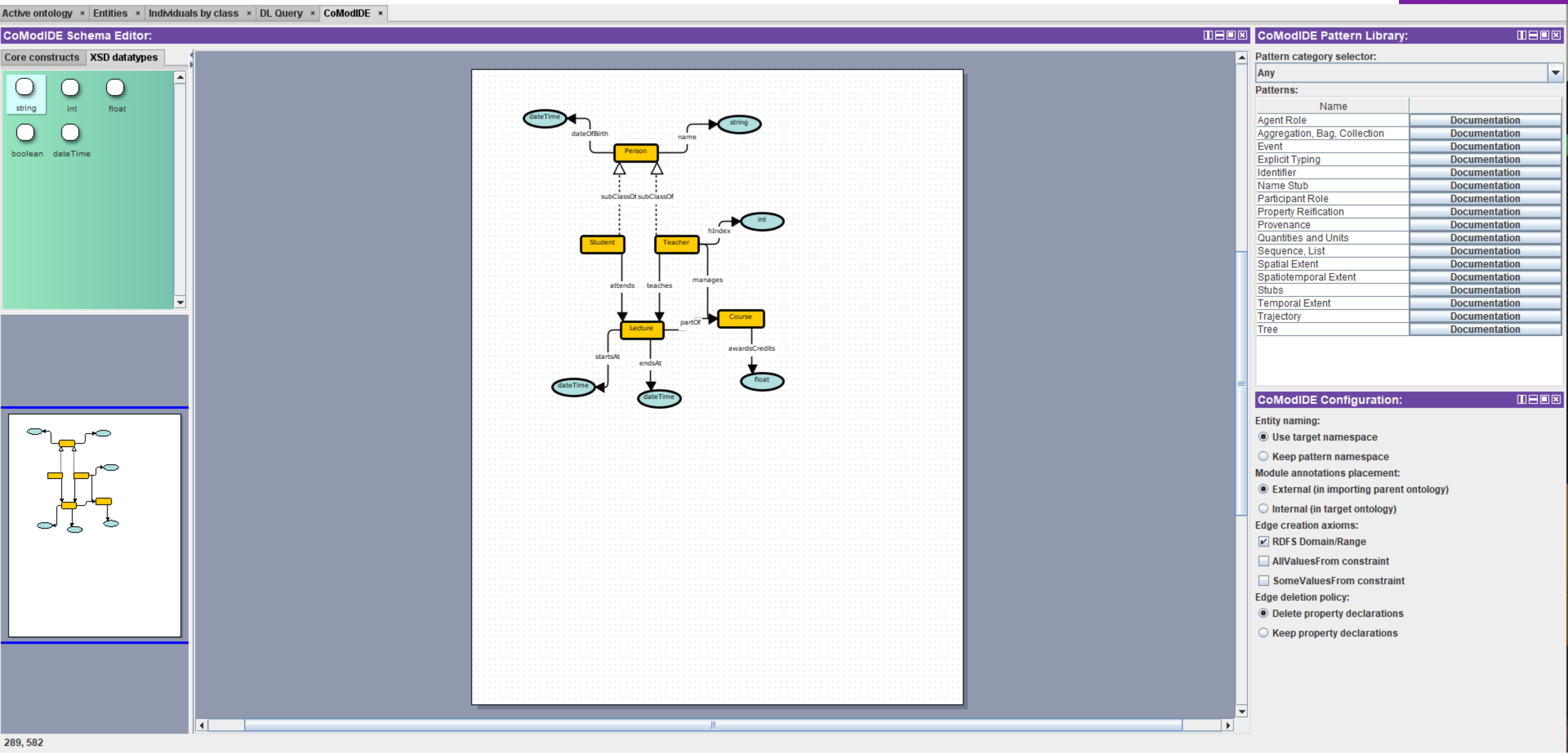
- Open Knowledge Network - Knowledge networks pool together many types of information and ideas so that they can be accessed and leveraged to create new understanding. These networks have become important tools for many large organizations that are taking advantage of the current Big Data revolution. However, these vast information networks are often unavailable to many in government, academia, small businesses and nonprofits. The Convergence Accelerator's new awards will fund the creation of a nonproprietary infrastructure for building an Open Knowledge Network. Some of the teams supported by the new awards will build tools that will identify, harvest, and incorporate datasets for the network. Others will build elements of the open knowledge network that address specific challenges, such as manufacturing, urban infrastructure, geosciences, biomedicine and much more. Yet others will provide key aspects of the technical infrastructure needed to facilitate the creation and use of such networks.

OKN project

Convergence Accelerator Phase I (RAISE): Spatially-Explicit Models, Methods, and Services for Open Knowledge Networks

NSF Org:	OIA Office of Integrative Activities
Initial Amendment Date:	September 10, 2019
Latest Amendment Date:	September 10, 2019
Award Number:	1936677
Award Instrument:	Standard Grant
Program Manager:	Lara Campbell OIA Office of Integrative Activities O/D Office Of The Director
Start Date:	September 1, 2019
End Date:	May 31, 2020 (Estimated)
Awarded Amount to Date:	$999,547.00
Investigator(s):	Krzysztof Janowicz jano@geog.ucsb.edu (Principal Investigator) Mark Schildhauer (Co-Principal Investigator) Dean Rehberger (Co-Principal Investigator) Pascal Hitzler (Co-Principal Investigator) Wenwen Li (Co-Principal Investigator)

CoMODIDE modeling interface



Methods

- We develop and apply a whole range of techniques to problems around knowledge graphs, including
  - Deep learning
  - Natural language processing
  - Logic-based knowledge representation
  - Computational logic and automated reasoning
- We apply our methods to other fields
  - Intelligence data integration and analysis (DARPA)
  - Cognitive Agents (AFOSR)
  - Humanities (Mellon Foundation)
  - Explainable Deep Learning (OBOR)
  - Food Systems data (NIST / Department of Commerce)
  - Scientific data (NSF GEO)
  - Industry (several)

Artificial Intelligence:
Bridging between AI paradigms

RDF deductive reasoning

- [Note: RDF is one of the simplest useful knowledge representation languages beyond propositional logic.]

- Think knowledge graph.
- Think node-edge-node triples such as

BarackObama	rdf:type	President
President	rdfs:subClassOf	Human

- Then there is a (fixed, small) set of inference rules, such as
  rdf:type(x,y) AND rdfs:subClassOf(y,z)THEN rdf:type(x,z)

- Logical consequence:

BarackObama	rdf:type	Human

Deductive (logical) reasoning

- Given a set of logical axioms K.
- Given another logical axiom A.

- Is A a logical consequence of K? (yes/no)

- This is a classification problem.
- Very complicated but provably correct algorithms exist for many logics.
  - These algorithms often take a long time.
  - They can rarely be distributed.
  - They are brittle with respect to noisy input data.

- Since this is a classification problem, can we use machine learning (deep learning) to solve it?

Representation

- **Goal is to be able to reason over unseen knowledge graphs. I.e. the out-of-vocabulary problem needs addressing.**

- **Normalization of vocabulary (i.e., it becomes shared vocabulary across all input knowledge graphs.**
- **One vocabulary item becomes a one-hot vector (dimension d, number of normalized vocabulary terms)**
- **One triple becomes a 3 x d matrix.**
- **The knowledge graph becomes an n x 3 x d tensor. (n is the number of knowledge graph triples)**

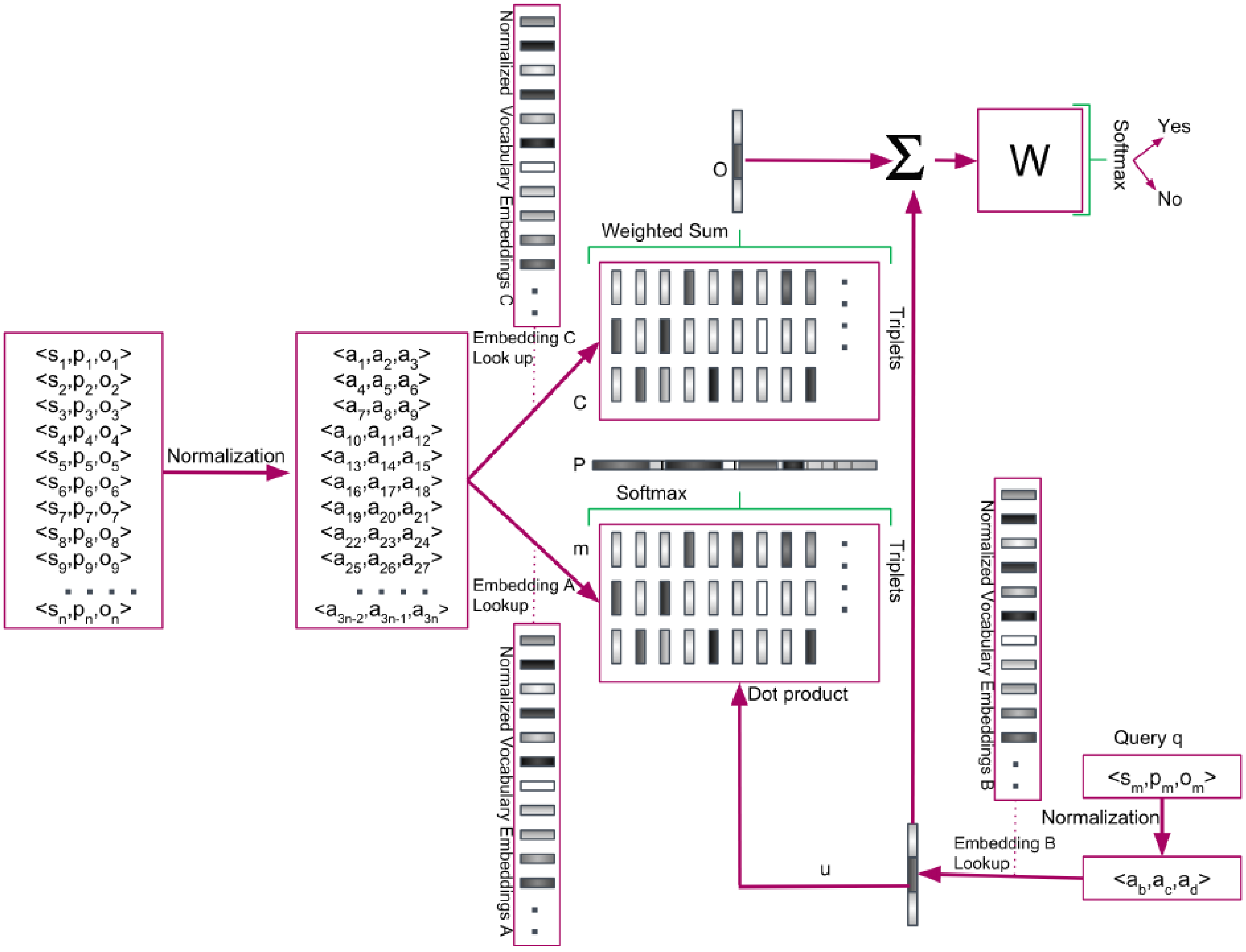
- **Knowledge graph is stored in "memory."**

Mechanics

- An attention mechanism retrieves momory slots useful for finding the correct answer to a query.
- These are combined with the query and run through a (learned) matrix to retrieve a new (processed) query.

- This is repeated (in our experiment with 10 “hops”).
- The final out put is a yes/no answer to the query.

Memory Network based on MemN2N



Experiments: Performance

Test Dataset	#KG	Base						Inferred						Invalid
		#Facts	#Ent.	%Class	%Indv	%R.	%Axiom.	#Facts	#Ent.	%Class	%Indv	%R.	%Axiom.	#Facts
OWL-Centric	2464	996	832	14	19	3	0	494	832	14	0.01	1	20	462
Linked Data	20527	999	787	3	22	5	0	124	787	3	0.006	1	85	124
OWL-Centric Test Set	21	622	400	36	41	3	0	837	400	36	3	1	12	476
Synthetic Data	2	752	506	52	0	1	0	126356	506	52	0	1	0.07	700

Table 2: Statistics of various datasets used in experiments

Baseline: non-normalized embeddings, same architecture

Training Dataset	Test Dataset	Valid Triples Class			Invalid Triples Class			Accuracy
		Precision	Recall /Sensitivity	F-measure	Precision	Recall /Specificity	F-measure	
OWL-Centric Dataset	Linked Data	93	98	96	98	93	95	**96**
OWL-Centric Dataset (90%)	OWL-Centric Dataset (10%)	88	91	89	90	88	89	**90**
OWL-Centric Dataset	OWL-Centric Test Set [b]	79	62	68	70	84	76	**69**
OWL-Centric Dataset	Synthetic Data	65	49	40	52	54	42	**52**
OWL-Centric Dataset	Linked Data [a]	54	98	70	91	16	27	86
OWL-Centric Dataset [a]	Linked Data [a]	62	72	67	67	56	61	91
OWL-Centric Dataset(90%) [a]	OWL-Centric Dataset(10%) [a]	79	72	75	74	81	77	80
OWL-Centric Dataset	OWL-Centric Test Set [ab]	58	68	62	62	50	54	58
OWL-Centric Dataset [a]	OWL-Centric Test Set [ab]	77	57	65	66	82	73	73
OWL-Centric Dataset	Synthetic Data [a]	70	51	40	47	52	38	51
OWL-Centric Dataset [a]	Synthetic Data [a]	67	23	25	52	80	62	50
Baseline								
OWL-Centric Dataset	Linked Data	73	98	83	94	46	61	43
OWL-Centric Dataset (90%)	OWL-Centric Dataset (10%)	84	83	84	84	84	84	82
OWL-Centric Dataset	OWL-Centric Test Set [b]	62	84	70	80	40	48	61
OWL-Centric Dataset	Synthetic Data	35	41	32	48	55	45	48

[a] More Tricky Nos & Balanced Dataset
[b] Completely Different Domain.

Table 3: Experimental results of proposed model

Experiments: Reasoning Depth

Test Dataset	Hop 0			Hop 1			Hop 2			Hop 3			Hop 4			Hop 5			Hop 6			Hop 7			Hop 8			Hop 9			Hop 10		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Linked Data[a]	0	0	0	80	99	88	89	97	93	77	98	86	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Linked Data[b]	2	0	0	82	91	86	89	98	93	79	100	88	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OWL-Centric[c]	19	5	9	31	75	42	78	80	78	48	47	44	4	34	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Synthetic	32	46	33	31	87	38	66	55	44	25	45	32	29	46	33	26	46	33	25	46	33	25	46	33	24	43	31	25	43	31	22	36	28

[a] LemonUby Ontology
[b] Agrovoc Ontology
[c] Completely Different Domain

Table 4: Experimental results over each reasoning hop

Dataset	Hop 1	Hop 2	Hop 3	Hop 4	Hop 5	Hop 6	Hop 7	Hop 8	Hop 9	Hop 10
OWL-Centric[a]	8%	67%	24%	0.01%	0%	0%	0%	0%	0%	0%
Linked Data[b]	31%	50%	19%	0%	0%	0%	0%	0%	0%	0%
Linked Data[c]	34%	46%	20%	0%	0%	0%	0%	0%	0%	0%
OWL-Centric[d]	5%	64%	30%	1%	0%	0%	0%	0%	0%	0%
Synthetic Data	0.03%	1.42%	1%	1.56%	3.09%	6.03%	11.46%	20.48%	31.25%	23.65%

[a] *Training Set*
[b] LemonUby Ontology
[c] Agrovoc Ontology
[d] Completely Different Domain

Table 5: Data distribution per knowledge graph over each reasoning hop

Training time: just over a full day

Artificial Intelligence

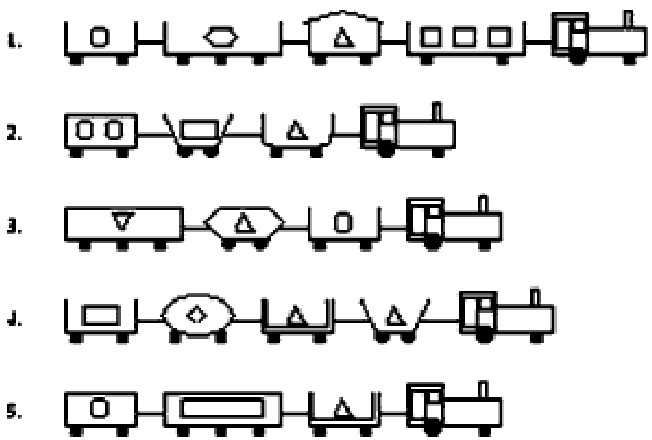
- **Work like this is of fundamental importance as it bridges between two of the major subfields of Artificial Intelligence:**
  - **Machine Learning (including deep learning)**
  - **Knowledge Representation and Reasoning**

Artificial Intelligence: Concept Induction

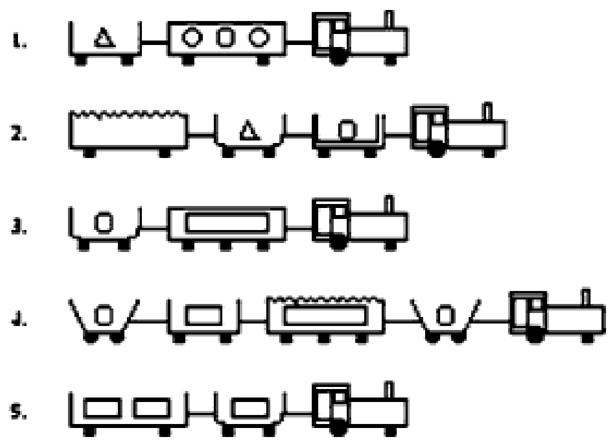
DL-Learner [Lehmann, Hitzler]

Approach similar to inductive logic programming, but using Description Logics (the logic underlying OWL).

Positive examples:



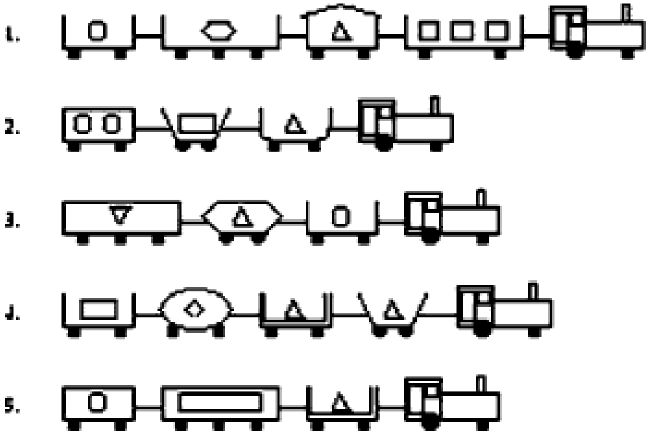
negative examples:



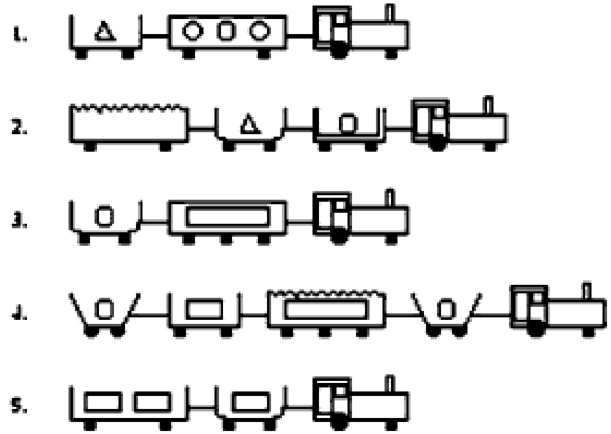
Task: find a class description (logical formula) which separates positive and negative examples.

DL-Learner

Positive examples:



negative examples:



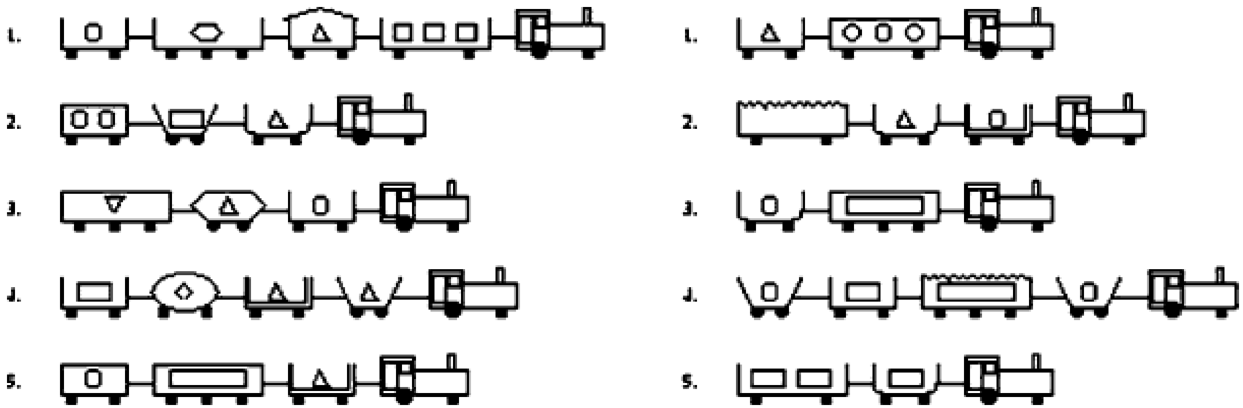
DL-Learner result: $\exists \mathtt{hasCar.(Closed} \sqcap \mathtt{Short)}$

In FOL:

$$\{x \mid \exists y(\mathrm{hasCar}(x, y) \land \mathrm{Closed}(y) \land \mathrm{Short}(y))\}$$

DL-Learner

DL-Learner uses refinement operators to construct ever better approximations of a solution.



$\top$

Train – covers all examples.

$\exists$hasCar.$\top$

$\exists$hasCar.Closed – covers all positives, two negatives

$\exists$hasCar(Closed $\sqcap$ Short) – solution

Scalability Issues with DL-Learner

DaSe Lab

- For large-scale experiments, DL-Learner took 2 hours or more for one run.
- We knew we needed at least thousands of runs.

- So we needed a more scalable solution.
- The provably correct algorithms have very high complexity.

- Hence we had to develop a heuristic which trades (some) correctness for speed.

ECII algorithm and system

- **We thus implemented our own system, ECII (Efficient Concept Induction from Instances) which trades some correctness for speed. [Sarker, Hitzler, AAAI-19]**

Experiment Name	Number of Logical Axioms	Runtime (sec)					Accuracy (α_3)		Accuracy α_2			
		DL[a]	DL FIC(1)[b]	DL FIC(2)[c]	ECII DF[d]	ECII KCT[e]	DL[a]	ECII DF[d]	DL FIC(1)[b]	DL FIC(2)[c]	ECII DF[d]	ECII KCT[e]
Yinyang_examples	157	0.065	0.0131	0.019	0.089	0.143	1.000	0.610	1.000	1.000	0.799	1.000
Trains	273	0.01	0.020	0.047	0.05	0.095	1.000	1.000	1.000	1.000	1.000	1.000
Forte	341	2.5	1.169	6.145	0.95	0.331	0.965	0.642	0.875	0.875	0.733	1.000
Poker	1,368	0.066	0.714	0.817	1	0.281	1.000	1.000	0.981	0.984	1.000	1.000
Moral Reasoner	4,666	0.1	3.106	4.154	5.47	6.873	1.000	0.785	1.000	1.000	1.000	1.000
ADE20k I	4,714	577.3[f]	4.268	31.887	1.966	23.775	0.926	0.416	0.263	0.814	0.744	1.000
ADE20k II	7,300	983.4[f]	16.187	307.65	20.8	293.44	1.000	0.673	0.413	0.413	0.846	0.900
ADE20k III	12,193	4,500[g]	13.202	263.217	51	238.8	0.375	0.937	0.375	0.375	0.930	0.937
ADE20k IV	47,468	4,500[g]	93.658	523.673	116	423.349	0.375	NA	0.608	0.608	0.660	0.608

[a] DL : DL-Learner
[b] DL FIC (1) : DL-Learner fast instance check with runtime capped at execution time of ECII DF
[c] DL FIC (2) : DL-Learner fast instance check with runtime capped at execution time of ECII KCT
[d] ECII DF : ECII default parameters
[e] ECII KCT : ECII keep common types and other default parameters
[f] Runtimes for DL-Learner were capped at 600 seconds.
[g] Runtimes for DL-Learner were capped at 4,500 seconds.

ECII vs. DL-Learner

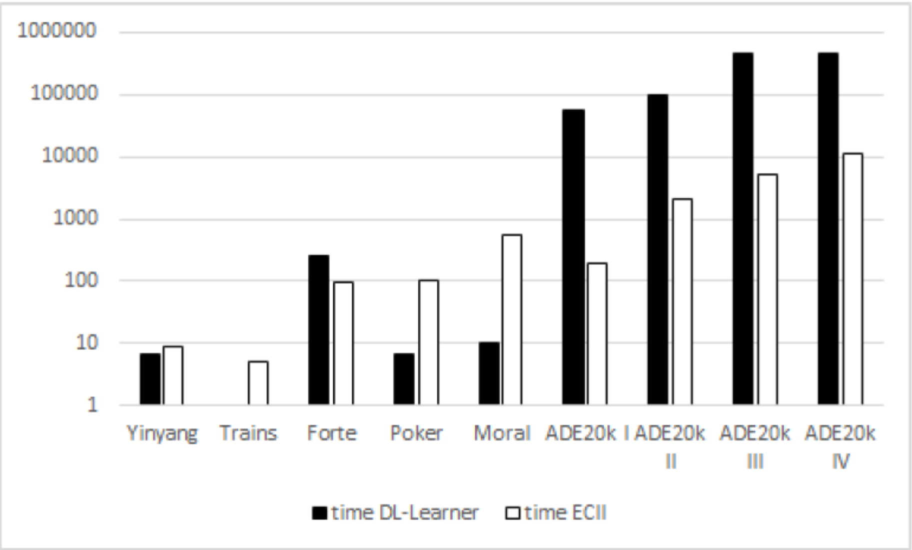


Figure 1: Runtime comparison between DL-Learner and ECII. The vertical scale is logarithmic in hundredths of seconds, and note that DL-Learner runtime has been capped at 4,500 seconds for ADE20k III and IV. For ADE20k I it was capped at each run at 600 seconds.

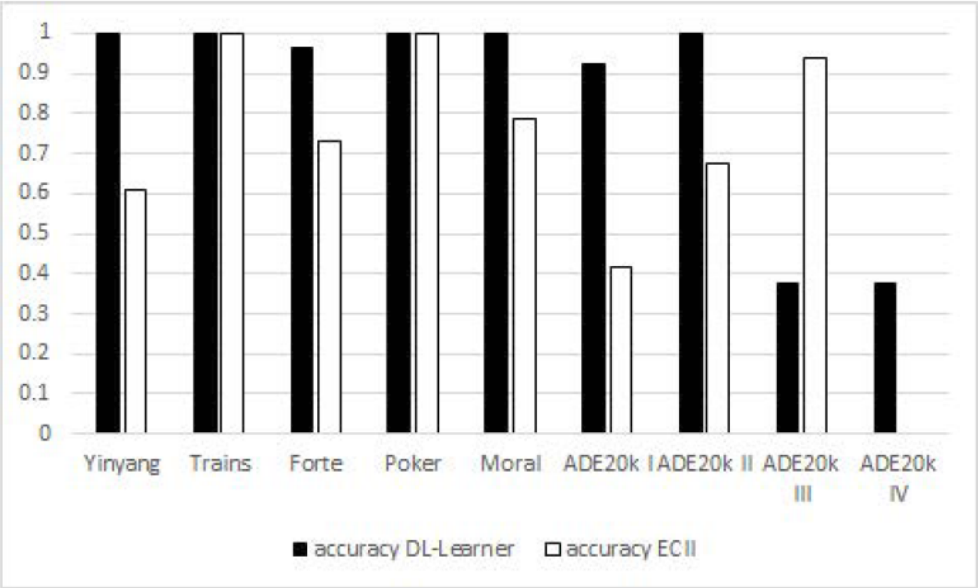


Figure 2: Accuracy (α_3) comparison between DL-Learner and ECII. For ADE20k IV it was not possible to compute an accuracy score within 3 hours for ECII as the input ontology was too large.

ECII application areas

(some of these we are working on, some of these we hope to be working on in the near future)

- Explaining black-box machine learning systems (including deep learning)
- Explaining results from data analysis such as the meaning of factors in a factor analysis.
- Explaining results from recommender systems.
- Uncovering data bias.
- Etc.

All of this will require the use of knowledge graphs as background knowledge.

Thanks!

Get in touch if interested:
hitzler@ksu.edu

Consider coming to my class in Spring or Fall

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, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, Peng Ji, The GeoLink Knowledge Graph. Big Earth Data 2 (2), 2018, 131-143.
- Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Alicia Sheill, Seila Gonzalez, Catherine Foley, Dean Rehberger, Ethan Watrall, Walter Hawthorne, Duncan Tarr, Ryan Carty, Jeff Mixter, The Enslaved Ontology 1.0: People of the Historic Slave Trade. Technical Report, enslaved.org, 23 April 2019.
- Krzysztof Janowicz, Frank van Harmelen, James A. Hendler, Pascal Hitzler, Why the Data Train Needs Semantic Rails. AI Magazine 26 (1), 2015, 5-14.

References

- Pascal Hitzler, Cogan Shimizu, Modular Ontologies as a Bridge Between Human Conceptualizations and Data. In: Peter Chapman, Dominik Endres, Nathalie Pernelle: Graph-Based Representation and Reasoning - 23rd International Conference on Conceptual Structures, ICCS 2018, Edinburgh, UK, June 20-22, 2018, Proceedings. Lecture Notes in Computer Science 10872, Springer 2018, pp. 3-6.
- Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, Valentina Presutti (eds.), Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web Vol. 25, IOS Press/AKA Verlag, 2016.
- Adila Krisnadhi, Pascal Hitzler, Modeling With Ontology Design Patterns: Chess Games As a Worked Example. In: Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, Valentina Presutti (eds.), Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web Vol. 25, IOS Press/AKA Verlag, pp. 3-22.
- Cogan Shimizu, Karl Hammar, CoModIDE – The Comprehensive Modular Ontology IDE. In: 18th International Semantic Web Conference: Satellite Events, 2019, to appear.