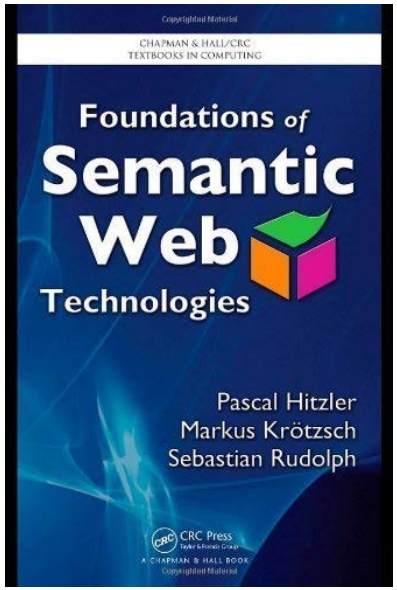
Managing Data for Integration and Reuse

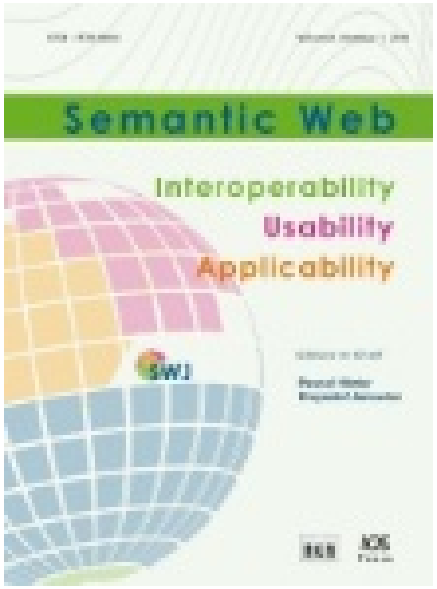
Pascal Hitzler
Data Semantics Laboratory (DaSe Lab)
Kansas State University

http://www.daselab.org

About me

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





The Data Pipeline

Acuisition (GPS, UAVs, senors) is fun and flashy as is Reuse (all these cool data analytics / machine learning results)

But a lot of time is sunk in the data management piece.

Managing data for

- **Sharing**
- **Discovery**
- **Integration**
- **Reuse**

Management is most expensive when

- **Integrating from independent sources**
- **Repurposing**

Data Management

DaSe Lab

- **Driving question: How do you lower data management cost?**
  - **How do you make it easier to share and discover data?**
  - **How do you make it easier to integrate data?**
  - **How do you make it easier to repurpose data?**

- **Three key aspects to addressing these issues:**
  - **Use of leading standards.**
  - **Self-explanatory data (which carries a *formal semantics*).**
  - **Data organization which resonates with human experts' conceptualizations.**

Google Knowledge Graph

DaSe Lab

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

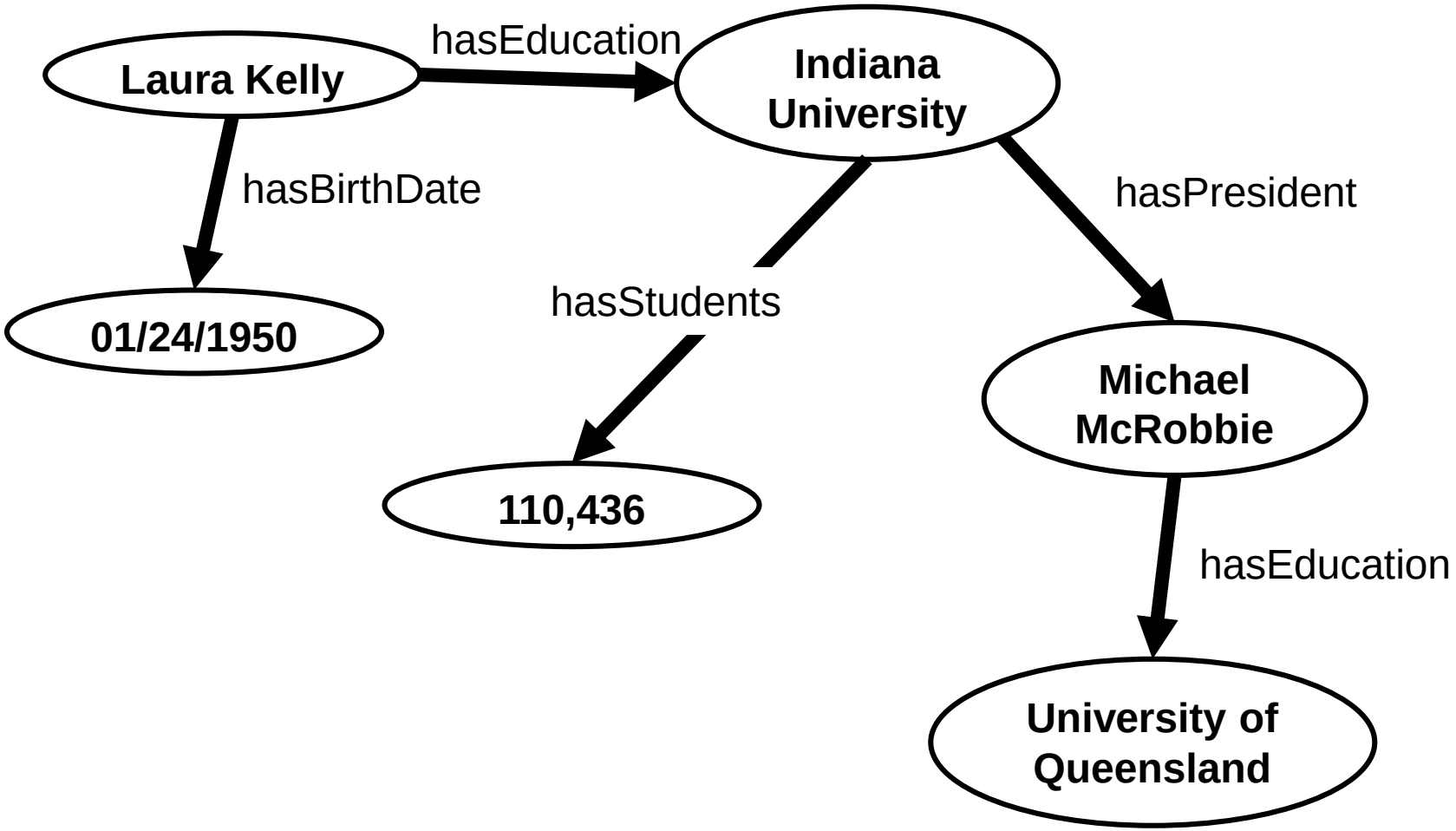
Born: October 11, 1950 (age 69 years), Melbourne, Australia

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



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

DaSe Lab

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)

Knowledge Graph Standards

DaSe Lab

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

W3C Recommendation

OWL 2 Web Ontology Language Primer (Second Edition)

W3C Recommendation 11 December 2012

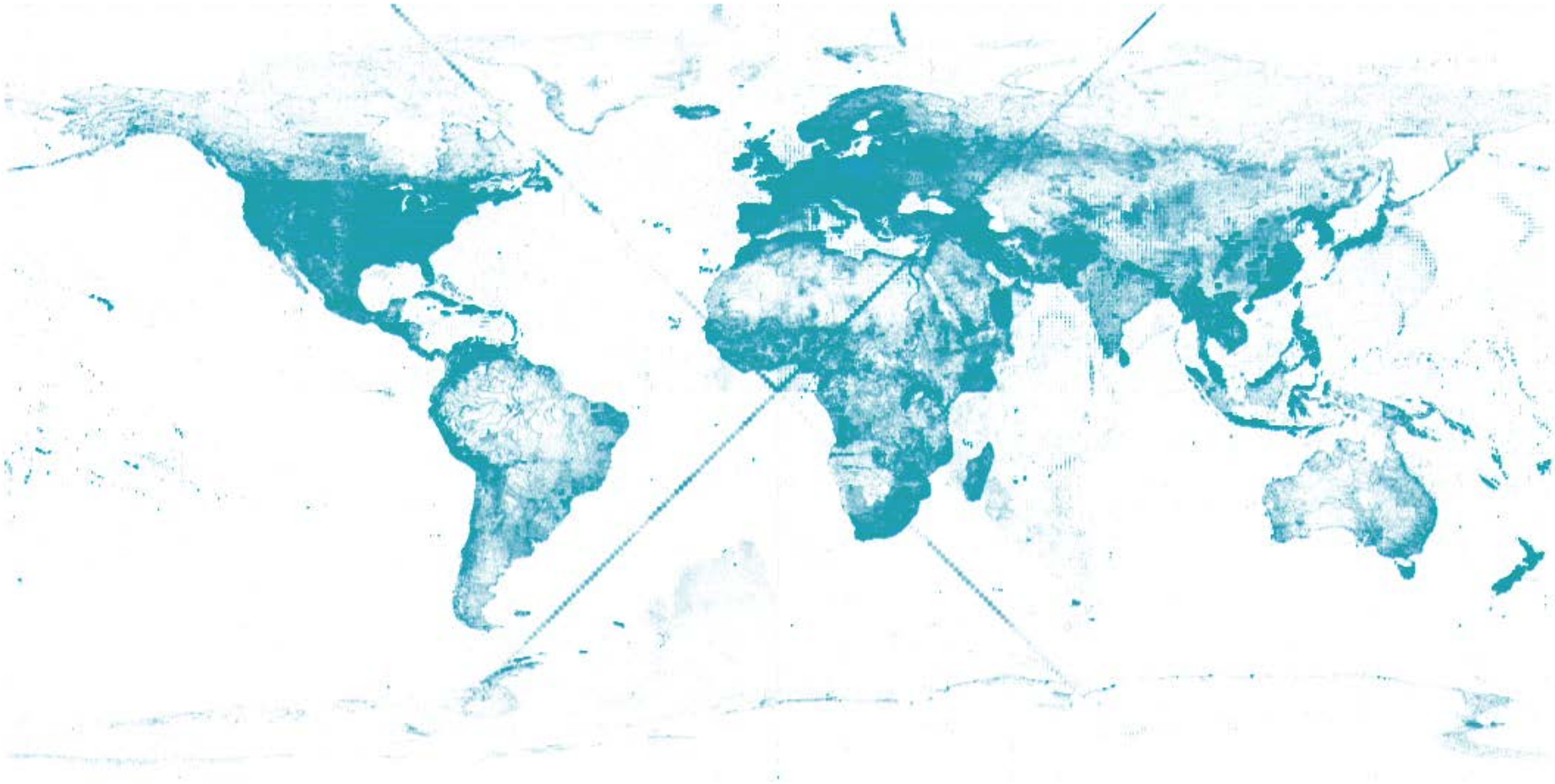
This version:
http://www.w3.org/TR/2012/REC-owl2-primer-20121211/
Latest version (series 2):
http://www.w3.org/TR/owl2-primer/
Latest Recommendation:
http://www.w3.org/TR/owl-primer
Previous version:
http://www.w3.org/TR/2012/PER-owl2-primer-20121018/
Editors:
Pascal Hitzler, Wright State University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

Linked Data: Volume

DaSe Lab

Geoindexed Linked Data – courtesy of Krzysztof Janowicz, 2012

http://stko.geog.ucsb.edu/location_linked_data



Help document

Datasets

Cruises

Vessels

Instruments

Physical Samples

Gazetteer Feature

Researchers

Organizations

Awards

enslaved.org

DaSe Lab

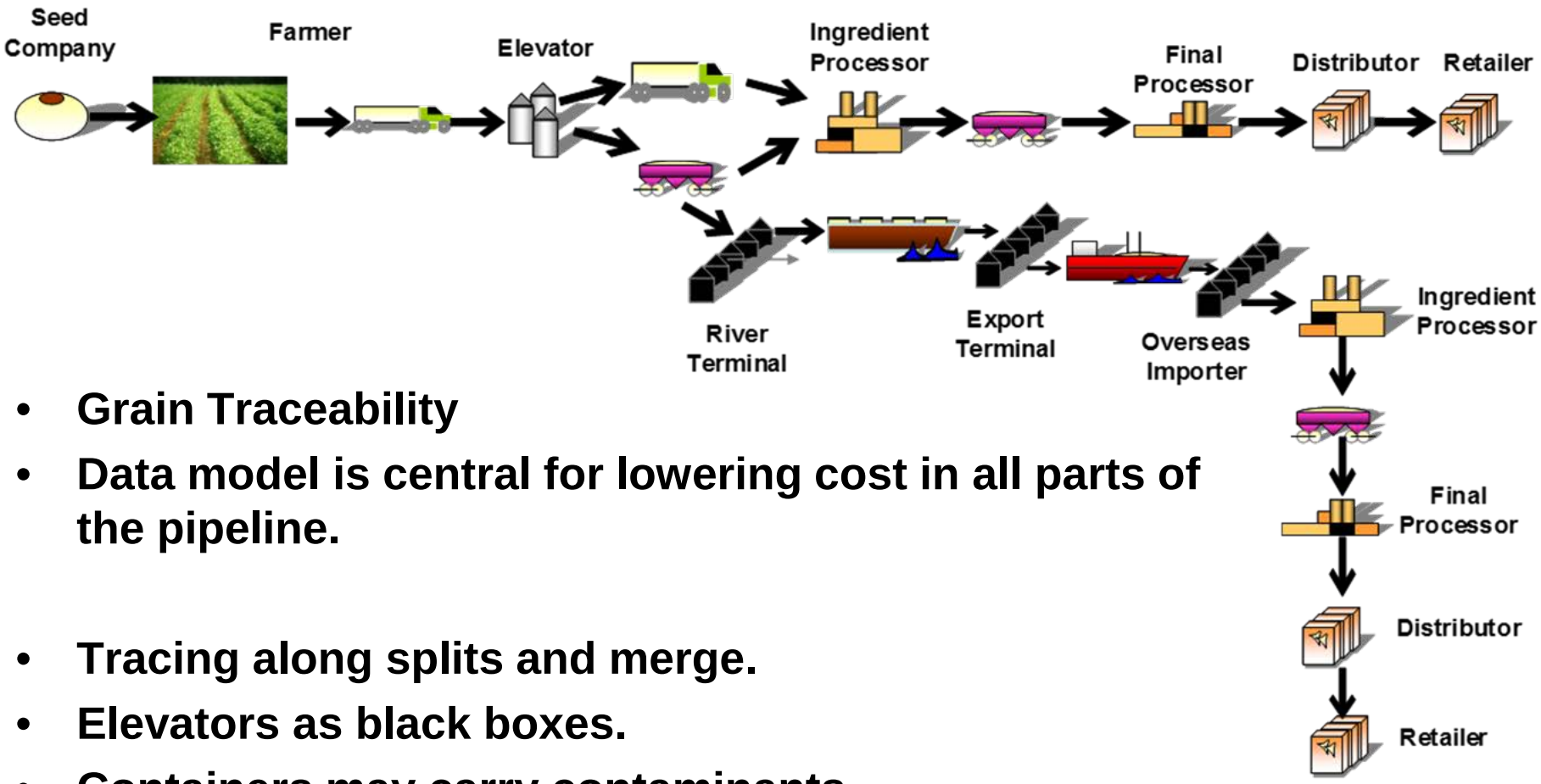
Recently started project

DaSe Lab

- 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

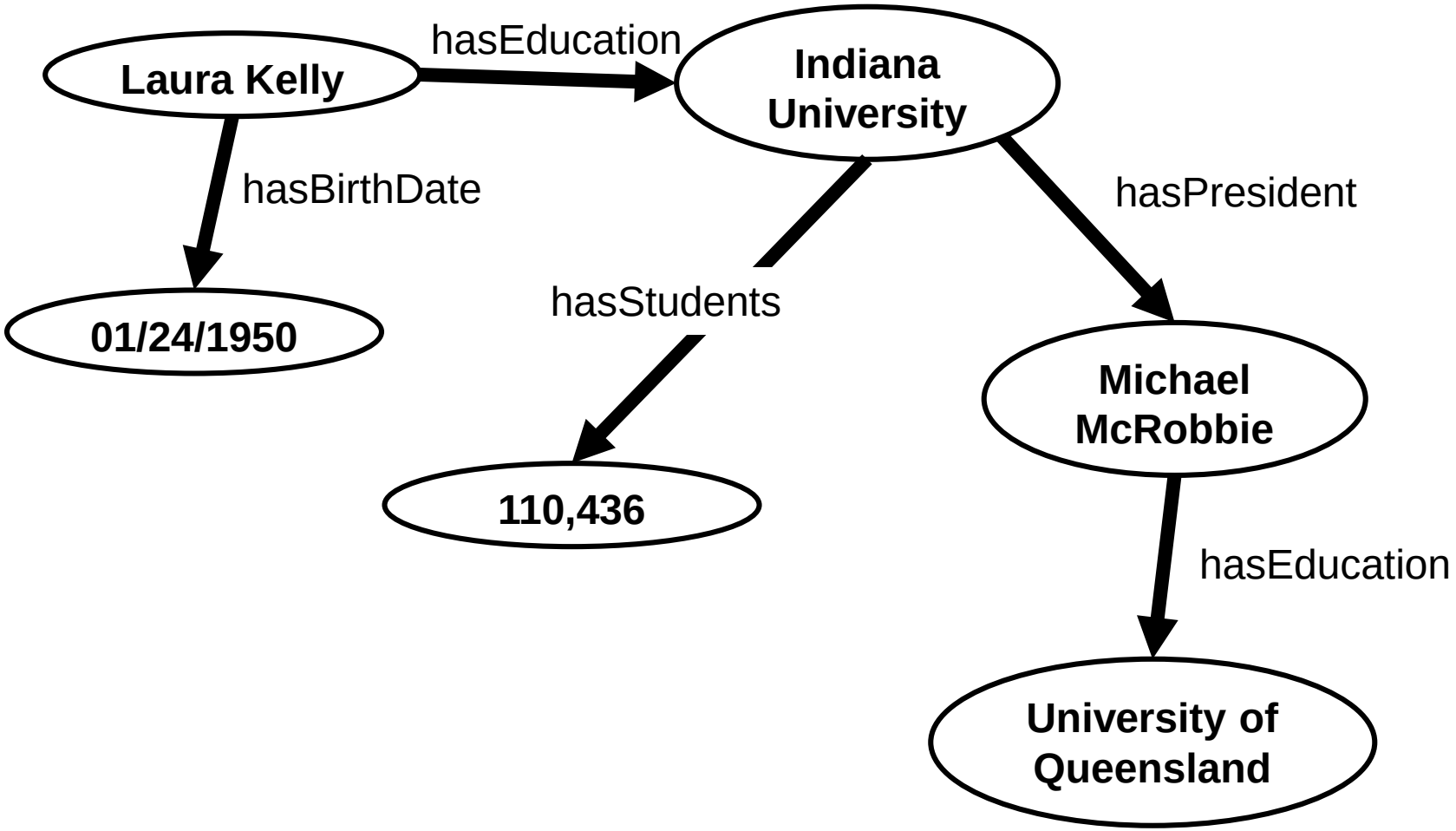


- **Grain Traceability**
- **Data model is central for lowering cost in all parts of the pipeline.**

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

Figure acknowledgement: NIST / Evan Wallace

This is not a good Knowledge Graph!



What makes a good data model?

DaSe Lab

- Structure resonates with both
  - human expert conceptualizations
  - data and use case requirements
- Generally low maintenance cost
  - Sustainable: robust for future use and re-use
  - Extendable without high management costs

- Ease of use with software and tools
- Machine processable (standards)

- Meets technical, legal, societal requirements
- Stakeholder buy-in

Some of our research

DaSe Lab

Lead Question:

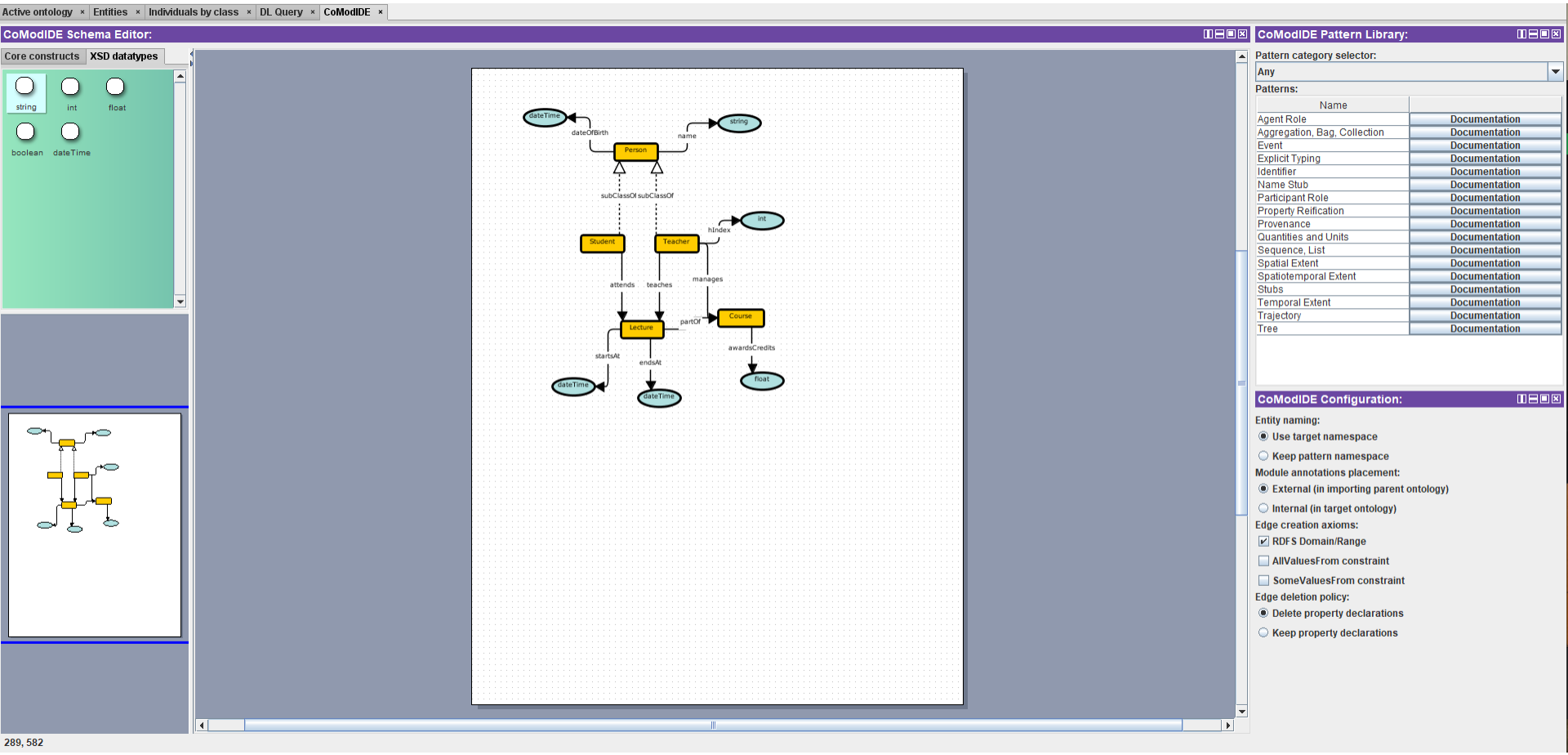
How to lower knowledge graph management cost while meeting requirements.

Principles:

Our design and development process

- **bridges interdisciplinary barriers,**
- **produces artefacts which resonate with human expert understanding,**
- **is fully compatible with leading standards,**
- **is made to save on development and management costs.**

CoMODIDE modeling interface



Partnering

DaSe Lab

- **Verticals need data management**
- **In particular if data is complex, heterogeneous, long-tem sustainable, reuseable.**

- **Our methods are domain-independent**
- **We do not need prior knowledge about the field to apply them.**

Thanks!

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.