An Exercise in Data Analytics on Bibliographic Data

Andreas Harth

BDVA Data Architectures Session, Valencia, 2016-12-01

Institute of Applied Informatics and Formal Description Methods (AIFB)

Method

- Assumption
  - Researchers stay for research topics and continue a line of research over time
- Goal
  - Identification of topics of a group of researchers
  - Ranked list of important publications and persons in a group
- Groups
  - Authors: The Clinical Data Intelligence Project - A smart data initiative
  - Authors: Data intelligence on the Internet of Things
  - ISWC 2015 Senior PC
  - ESWC 2016 Area Chairs
  - Stream Reasoning Workshop 2016 Programme Committee
  - Internet Architecture Board Semantic Interop in IoT Workshop Chairs
  - BDVA Officials

Semantic Web

- „The Semantic Web is a web of data.“ – http://www.w3.org/2001/sw/

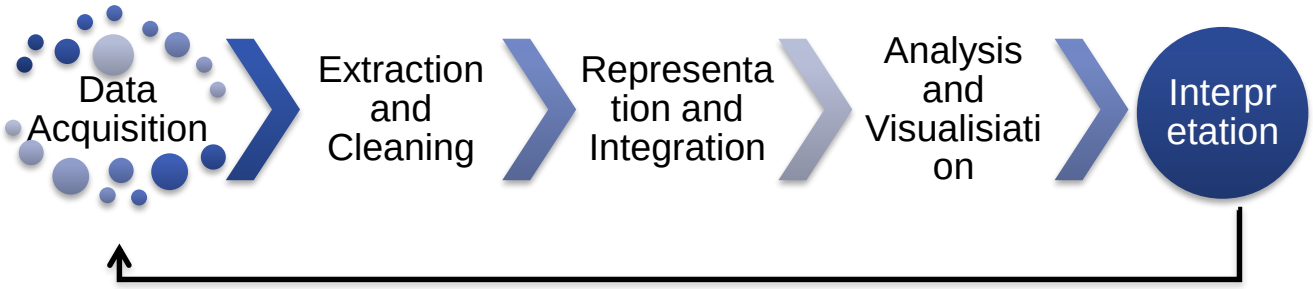
- Resource Description Framework (RDF) is a data model based on RDF triples, consisting of subject-predicate-object:

```
<http://dblp.org/rec/journals/puc/ZhouTZG16>
        dblp:title "Data intelligence on the Internet of Things.“ .
<http://dblp.org/rec/journals/puc/ZhouTZG16>
        dblp:authoredBy <http://dblp.org/pers/z/Zhou:Zhangbing> .
```

- Query language SPARQL for RDF

Can we use data from the Semantic Web to get an overview of the topics in a group of researchers?

Pipeline for data analytics



Step 1: Data acquisition

- DBLP is a database about computer science publications
  - Instance data about publications (3.4m) and researchers (1.7m)
  - Available as RDF (unofficially since 2006, officially since 2015)
- AMiner („Mining Deep Knowledge from Scientific Networks“)
  - Citation network extracted from PDFs
  - Own data format, publications are identified via DBLP title

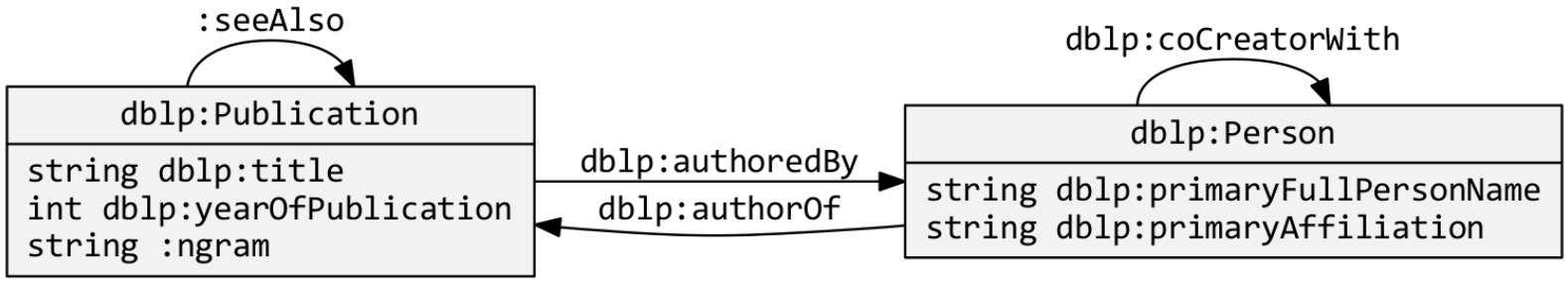
Step 2: Extraction and cleaning

- Analysis would benefit from topics of publications
- But we only have the title („Data intelligence on the Internet of Things”)
- Therefore: extraction of topics from titles via heuristic
  - Query title and year, filter for publication year > 2006
  - Remove stopwords
  - Porter stemming
  - Bigrams of titles leads to topics („data_intellig“, „intellig_internet“, „internet_thing“)

- Conversion of AMiner dataset to RDF (including DBLP identifiers)
- Initial run reveled problems with AMiner data (wrong links) -> manual filtering

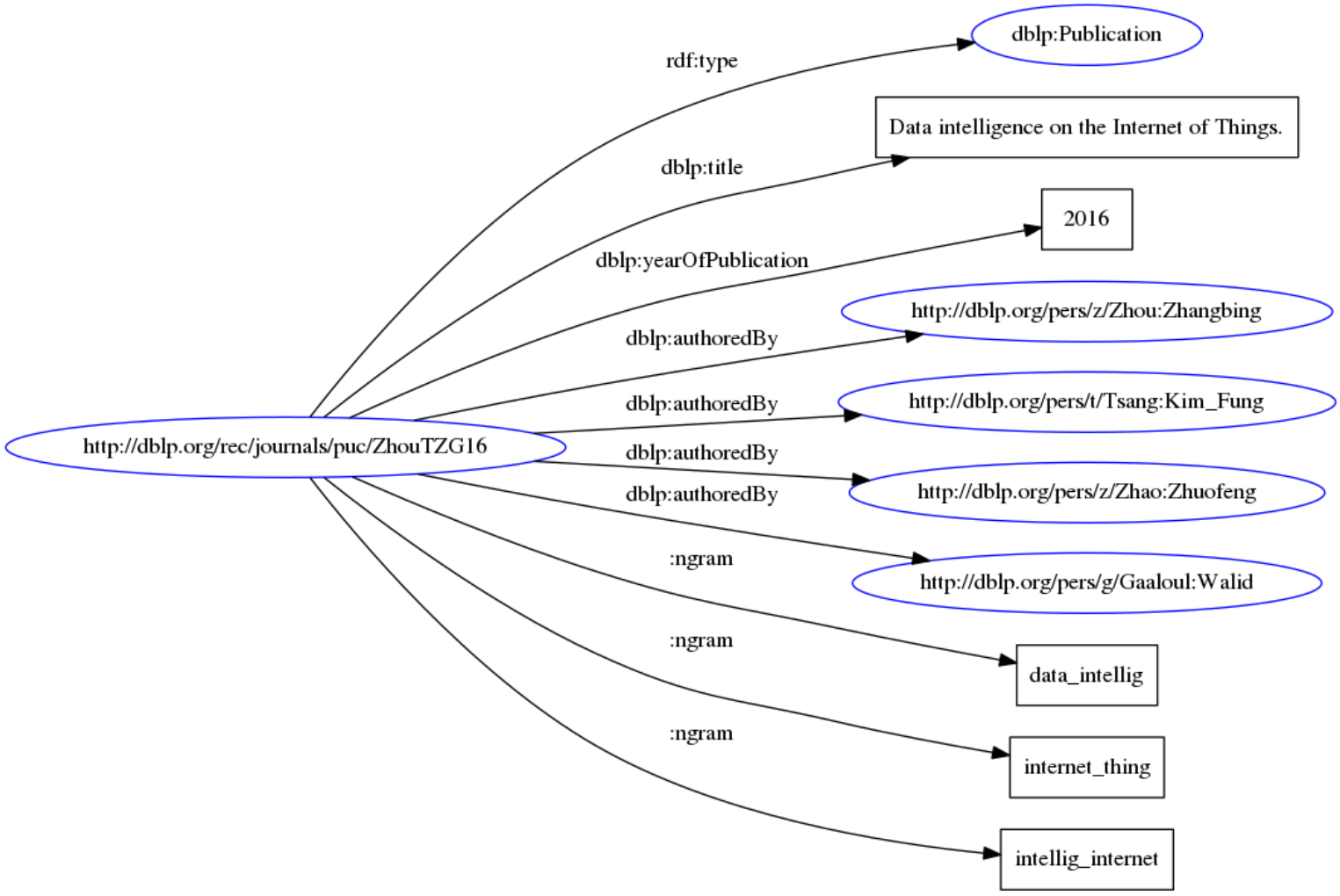
Step 3: Representation and integration

- Combination of DBLP and AMiner data (RDF)
  - DBLP data are available as RDF archive (beta 2016-07-03, 7.5 GB, 58m triple)
  - Own RDF version of AMiner data (2016-04-02, 762 MB, 5.7m triple)



- Indexing of RDF data for queries in a SPARQL repository
- Extraction of subgraphs

RDF representation



Extraction of subgraphs

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dblp: <http://dblp.org/rdf/schema-2015-01-26#>

DESCRIBE ?x ?y ?paper ?paper1
FROM <focus-${FOCUS}.nt>
FROM <dblp-2016-07-03.nt>
FROM <dblp-citation-good-links.nt>
WHERE {
  ?s foaf:focus ?x .
  ?paper dblp:authoredBy ?x .
  ?paper dblp:authoredBy ?y .
  OPTIONAL { ?paper1 dblp:authoredBy ?y . }
}
```

Step 4: Analysis and visualisation

- Visualisation of topics
  - Query n-grams/year
  - Import into Excel and visualisation as sorted table with sparklines
- Ranked list of persons and publications
  - Ranking with PageRank on extracted subgraphs
  - Query top-1000 researchers and publications
  - Representation as sorted list

Input: Focus people – BDVA officials

x	name
http://dblp.org/pers/c/Curry:Edward	Edward Curry
http://dblp.org/pers/m/Metzger:Andreas	Andreas Metzger
http://dblp.org/pers/p/Petkovic:Milan	Milan Petkovic
http://dblp.org/pers/m/M=uuml=ller_0003:J=uuml=rgen	Jürgen Müller
http://dblp.org/pers/l/Lama:Nuria_De	Nuria De Lama
http://dblp.org/pers/r/Robles:Ana_Garcia	Ana Garcia Robles

Result: top-k topics over the past ten years

web_servic	264
busi_process	190
servic_orient	173
product_line	132
softwar_engin	116
case_studi	109
servic_composit	104
softwar_product	103
self_adapt	103
model_driven	98
servic_base	86
inform_system	82
real_time	81
link_data	74
semant_web	72
cloud_comput	72
sensor_network	68
larg_scale	66
base_approach	64
model_base	64
softwar_develop	63
cloud_applic	61
social_network	61
context_awar	61
base_system	61

Result: top-k people

x	name	homepage	org	value
http://dblp.org/pers/s/Sheth:Amit_P=	Amit P. Sheth	http://knoesis.org/amit/		23.982754362527356
http://dblp.org/pers/l/Leymann:Frank	Frank Leymann	http://www.iaas.uni-stuttgart.de/institut/mitarbeiter/leymann/	University of Stuttgart, Germany	16.097017848853785
http://dblp.org/pers/k/Kossmann:Donald	Donald Kossmann	http://www.systems.ethz.ch/people/donaldk	ETH Zürich, Switzerland	14.018330680252722
http://dblp.org/pers/d/Dustdar:Schahram	Schahram Dustdar	http://www.infosys.tuwien.ac.at/Staff/sd/	Vienna University of Technology, Austria	13.956245062183461
http://dblp.org/pers/p/Pohl:Klaus	Klaus Pohl	http://www.sse.uni-due.de/de/team/leitung/prof-dr-klaus-pohl	University of Duisburg-Essen, Germany	12.226215286889335
http://dblp.org/pers/p/Papazoglou:Mike_P=	Mike P. Papazoglou	http://infolab.uvt.nl/~mikep/		11.659606273456527
http://dblp.org/pers/g/Ghezzi:Carlo	Carlo Ghezzi	http://home.dei.polimi.it/ghezzi/	Polytechnic University of Milan, Italy	10.55513894788003
http://dblp.org/pers/v/Vossen:Gottfried	Gottfried Vossen	http://dbis-group.uni-muenster.de/	University of Münster, Germany	9.541837139979409
http://dblp.org/pers/k/Kersten:Martin_L=	Martin L. Kersten	http://homepages.cwi.nl/~mk/	National Research Institute for Mathematics and Computer Science, Amsterdam, Netherlands	9.538315767254415
http://dblp.org/pers/s/Shadbolt:Nigel	Nigel Shadbolt	http://www.ecs.soton.ac.uk/~nrs/	University of Southampton, UK	9.399650684986979
http://dblp.org/pers/l/Linden_0001:Frank_van_der	Frank van der Linden		Philips Research Laboratories, Eindhoven	9.014816266969149
http://dblp.org/pers/p/Pernici:Barbara	Barbara Pernici	http://home.deib.polimi.it/pernici/	Politecnico di Milano, Italy	8.699239219842678
http://dblp.org/pers/l/Lehner:Wolfgang	Wolfgang Lehner	https://wwwdb.inf.tu-dresden.de/team/head/prof-dr-ing-wolfgang-lehner/	Dresden University of Technology, Germany	8.643431694683894
http://dblp.org/pers/k/Katzenbeisser_0001:Stefan	Stefan Katzenbeisser	http://www.seceng.informatik.tu-darmstadt.de/	Darmstadt University of Technology, Computer Science Department	8.558399303344048
http://dblp.org/pers/s/Sheng:Quan_Z=	Quan Z. Sheng	http://www.cs.adelaide.edu.au/~qsheng/	University of Adelaide, Australia	7.937988721513815
http://dblp.org/pers/z/Zimmermann_0001:Thomas	Thomas Zimmermann	http://thomas-zimmermann.com/	Microsoft Research, Redmond, USA	7.37787838239585
http://dblp.org/pers/f/Franch:Xavier	Xavier Franch	http://www.lsi.upc.es/~franch/	Polytechnic University of Catalonia, Barcelona, Spain	7.198816042217885
http://dblp.org/pers/r/Reussner:Ralf_H=	Ralf H. Reussner	http://sdq.ipd.kit.edu/people/ralf_reussner/	Karlsruhe Institute of Technology (KIT), Institute for Program Structures and Data Organization	7.043322533023784

Result: top-k publications

x	title	year	value
http://dblp.org/rec/conf/vldb/Sheth91	Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases.	1991	47.56501825132263
http://dblp.org/rec/books/daglib/0015277	Software Product Line Engineering - Foundations, Principles, and Techniques.	2005	42.88118192736089
http://dblp.org/rec/conf/icde/BorzsonyiKS01	The Skyline Operator.	2001	39.163342717959274
http://dblp.org/rec/reference/crc/Barni010	Digital Watermarking.	2010	38.03305338123086
http://dblp.org/rec/reference/sp/PedrinaciDS11	Semantic Web Services.	2011	33.232107334262565
http://dblp.org/rec/conf/vldb/SchmidtWKCMB02	XMark: A Benchmark for XML Data Management.	2002	32.089774385069084
http://dblp.org/rec/conf/cav/CimattiCGGPRST02	NuSMV 2: An OpenSource Tool for Symbolic Model Checking.	2002	30.55117996133293
http://dblp.org/rec/journals/dpd/GeorgakopoulosHS95	An Overview of Workflow Management: From Process Modeling to Workflow Automation Infrastructure.	1995	26.96097013689502
http://dblp.org/rec/conf/www/SigurbjornssonZ08	Flickr tag recommendation based on collective knowledge.	2008	26.952155586808843
http://dblp.org/rec/journals/ijahuc/BaldaufDR07	A survey on context-aware systems.	2007	24.1919735351826
http://dblp.org/rec/reference/se/Carro10	Logic Programming.	2010	21.99333868266909
http://dblp.org/rec/journals/expert/ShadboltBH06	The Semantic Web Revisited.	2006	19.344576858826954
http://dblp.org/rec/journals/csur/Kossmann00	The State of the art in distributed query processing.	2000	18.10753574444898
http://dblp.org/rec/conf/edbt/AgrawalGL98	Mining Process Models from Workflow Logs.	1998	17.85234892190127
http://dblp.org/rec/journals/computer/OmmeringLKM00	The Koala Component Model for Consumer Electronics Software.	2000	17.545490400638016
http://dblp.org/rec/books/daglib/0067100	Fundamentals of software engineering.	1991	17.165995460563785
http://dblp.org/rec/books/daglib/0005815	Production workflow - concepts and techniques.	2000	17.00612205375769
http://dblp.org/rec/journals/tse/CugolaNF01	The JEDI Event-Based Infrastructure and Its Application to the Development of the OPSS WFMS.	2001	16.889777055101742
http://dblp.org/rec/conf/coopis/MenaKSI96	OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies.	1996	16.367929834386697
http://dblp.org/rec/conf/vldb/KossmannRR02	Shooting Stars in the Sky: An Online Algorithm for Skyline Queries.	2002	16.086253558647176
http://dblp.org/rec/journals/tois/MiddletonSR04	Ontological user profiling in recommender systems.	2004	15.919633676143127
http://dblp.org/rec/journals/tse/BalsamoMIS04	Model-Based Performance Prediction in Software Development: A Survey.	2004	15.882470893229653
http://dblp.org/rec/conf/vldb/HaasKWY97	Optimizing Queries Across Diverse Data Sources.	1997	15.43908723936281
http://dblp.org/rec/journals/ao/RomanKLBLSPFBF05	Web Service Modeling Ontology.	2005	14.993721508298266

Challenges

- Data acquisition: availability, rights, freshness
- Extraction and cleaning: syntax, data formats, access protocols
- Representation and integration: modelling, inconsistencies, object equality/identity

Source: "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights" in The New York Times. 2014.

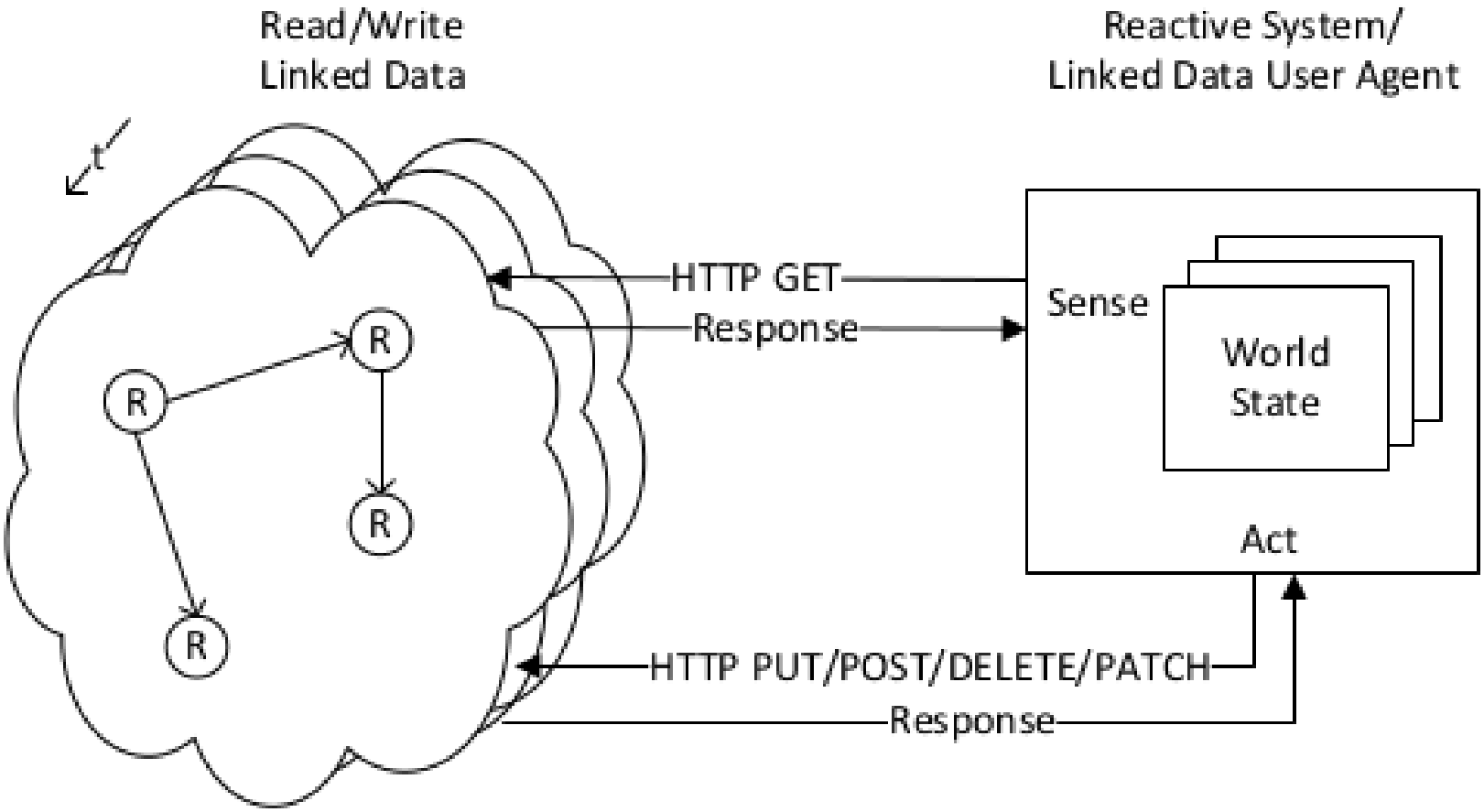
Data Scientists
"Janitor Work"

50%-80% of data scientists' time spent in "data wrangling"

- Analysis and visualisation: identification of questions, selection of visualisation

- Can we make decisions based on the analysis results? Can we carry out actions?
- Can we decrease the runtime of the pipeline from hours/minutes to seconds/milliseconds?

Sense-act cycle



Linked Data-Fu language and system

- Rule-based language for specifying interaction and composition
- Allows developers to state their intentions and execute desired interactions with Linked APIs and datasets
  - **Request rules** specify how and when to interact with APIs, i.ie., retrieve the state of resources (sense), trigger actions (act)
  - **Deduction rules** support reasoning constructs, e.g., transitivity, reflexivity of properties

OWL/Deduction rules

- authorOf/authoredBy
  - dblp:authorOf owl:inverseOf dblp:authoredBy

  - { ?x dblp:authorOf ?y . } => { ?y dblp:authoredBy ?x . } .
  - { ?x dblp:authoredBy ?y . } => { ?y dblp:authorOf ?x . } .

- Like CoCreatorWith: but without self-link (uncle relation?)
  - { ?x dblp:authoredBy ?person ; ?x dblp:authoredBy ?coauthor . } =>
  - { ?person dblp:coCreatorWith ?coauthor . }

Scenario: On-the-fly Data Integration for Los Angeles

POIs (Crunchbase, OSM, Wikimapia)

Venues/Events (Eventful, LastFM)

Buses/Stops (LA Metro)

Vehicles (Campus Cruisers)

Marine Vessels (AIS)

USC University of Southern California

Cooperation with USC Information Sciences Institute

Andreas Harth, Craig Knoblock, Steffen Stadtmüller, Rudi Studer and Pedro Szekely. "On-the-fly Integration of Static and Dynamic Linked Data". Fourth International Workshop on Consuming Linked Data (COLD 2013).

Training

Sarah Brauns, Tobias Käfer, Dirk Koriath, Andreas Harth. "Individualisiertes Gruppentraining mit Datenbrillen für die Produktion". GI-Jahrestagung 2016

Summary and conclusion

- Linked Data provides uniform interface to diverse set of data sources
- Web provides decentralised open platform

- Processing real-time heterogeneous data
- Both at rest (batch) and at motion (real-time)
- Read and write

- Rule-based language provides specification of semantics and dynamics
- Parallel rule engine provides scalable low-latency execution environment (http://linked-data-fu.github.io/)

- Existing industrial prototypes for applications in Industrie 4.0 and Internet of Things

Acknowledgements

- Zicari, Roberto. *Big Data: A data-driven society?*. Talk at Stanford EE Computer Systems Colloquium. 2014.
  - *http://web.stanford.edu/class/ee380/Abstracts/141029.html*
- Grobelnik, Marko. *Big Data Tutorial*. Presented at the European Data Forum (EDF). 2013.
  - http://www.slideshare.net/EUDataForum/edf2013-big-datatutorialmarkogrobelnik