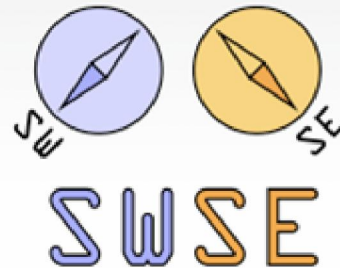


SWSE/YARS@SemData Sofia 2010

Andreas Harth

(Joint work with A. Hogan, S. Kinsella, A. Polleres, J. Umbrich, S. Decker)

Institute AIFB



Objects before documents!

Semantic Web Vision

- “creating technologies for supporting machine-readable web content sustaining autonomous interactive agents”
- Lack of single ontology on the web
- Lack of global agreement, e.g. UMBL says India is an organisation, ISO makes distinction between organisations (EU) and countries
- Lack of a priori knowledge about schema/vocabulary used when developing „intelligent agents“ (and if you decide to code against a schema you limit your application’s scope)
- Developing schema-agnostic systems is tricky but necessary when operating on web scale

Billion Triple Challenge 2009 Dataset

■ Top-10 properties (of 136,188 properties total)

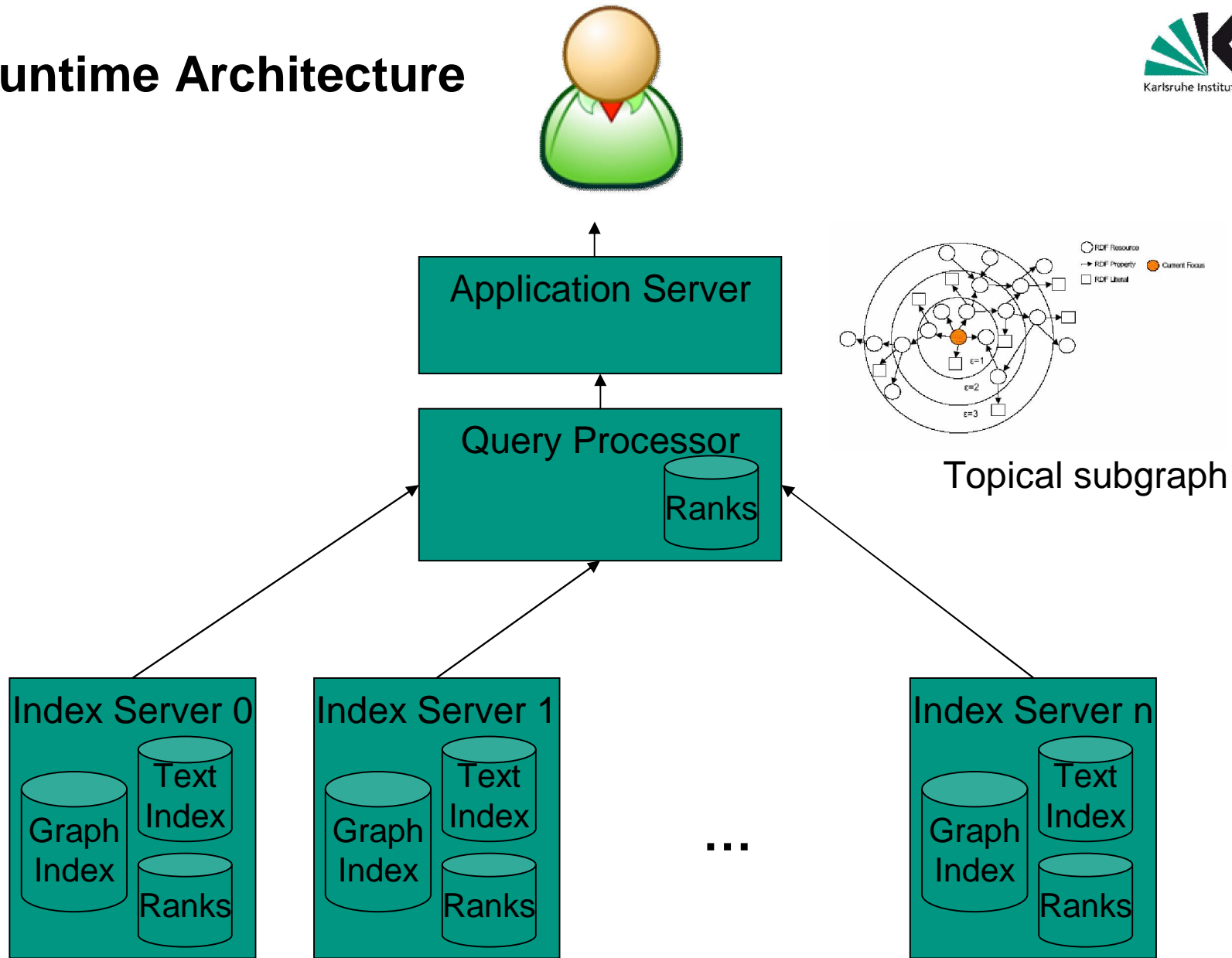
http://dbpedia.org/property/wikilink	156,434,900
rdf:type	143,479,200
rdfs:seeAlso	53,852,300
foaf:knows	35,786,400
foaf:nick	32,979,500
foaf:weblog	23,239,200
dc:title	22,356,700
akt:has-author	19,541,900
sioc:links_to	19,228,400
skos:subject	18,280,600

Data Preparation



Sorting and scanning as basic operations

Runtime Architecture



Reasoning (Authoritative Reasoning in SAOR)

- Traditional reasoning: one knowledge base from fixed group of contributors
- Web reasoning: anyone can say anything
- Question: how to deal with explosion of number of inferred statements?
- Answer: authoritative reasoning

Authoritative Reasoning

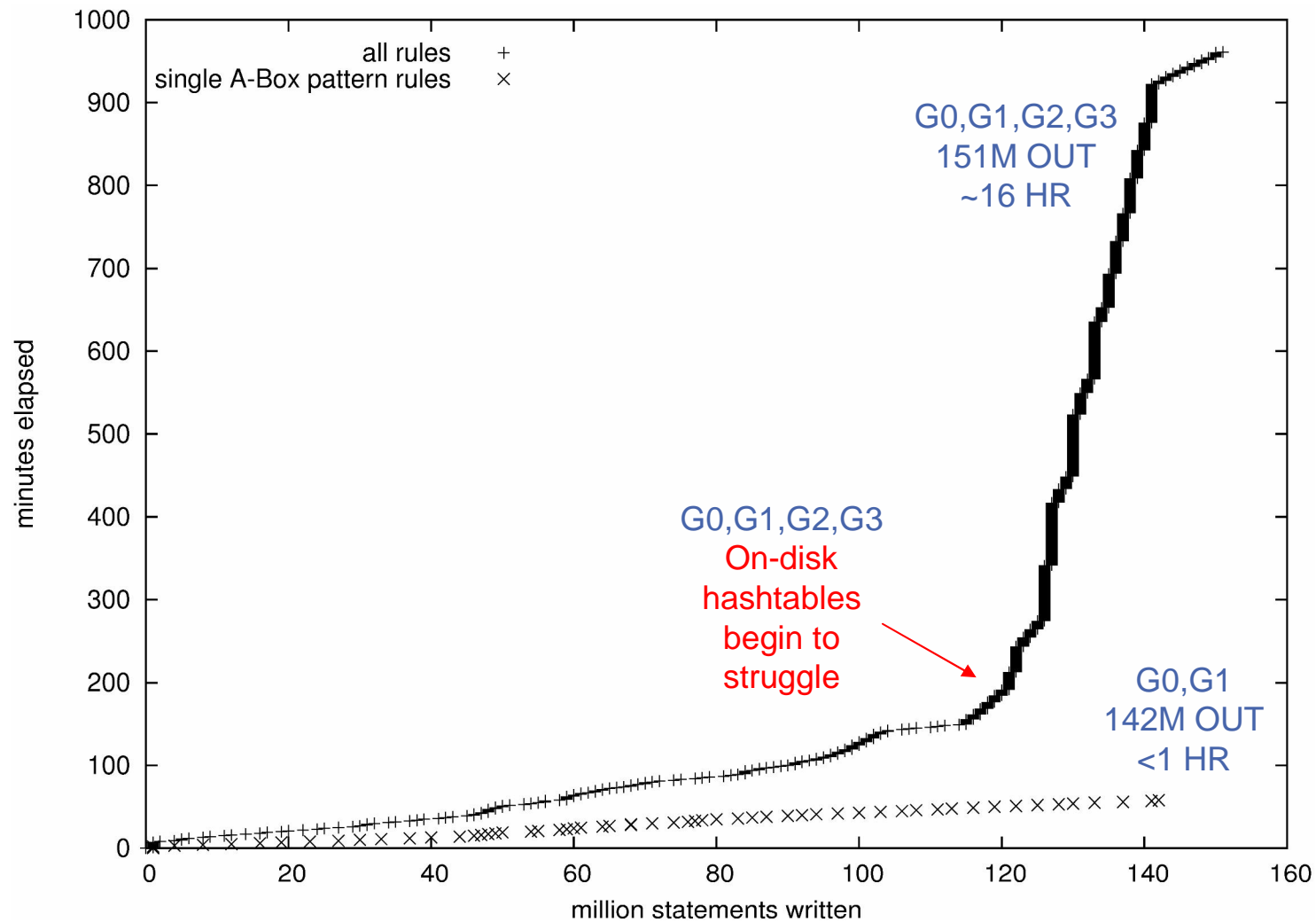
- T-Box only!
 - Document **D** authoritative for concept **C** iff:
 - **C** not identified by URI
 - OR
 - De-referenced URI of **C** coincides with or redirects to **D**
 - FOAF spec authoritative for `foaf:Person`
 - MY spec not authoritative for `foaf:Person`
-
- Only allow extension in authoritative documents
 - `my:Person rdfs:subClassOf foaf:Person . (MY spec)`
 - **BUT:** Reduce obscure memberships
 - `foaf:Person rdfs:subClassOf my:Person . (MY spec)`
 - **ALSO:** Protect specifications
 - `foaf:mbox rdf:type owl:SymmetricProperty . (MY spec)`
 - Similarly for other T-Box statements.



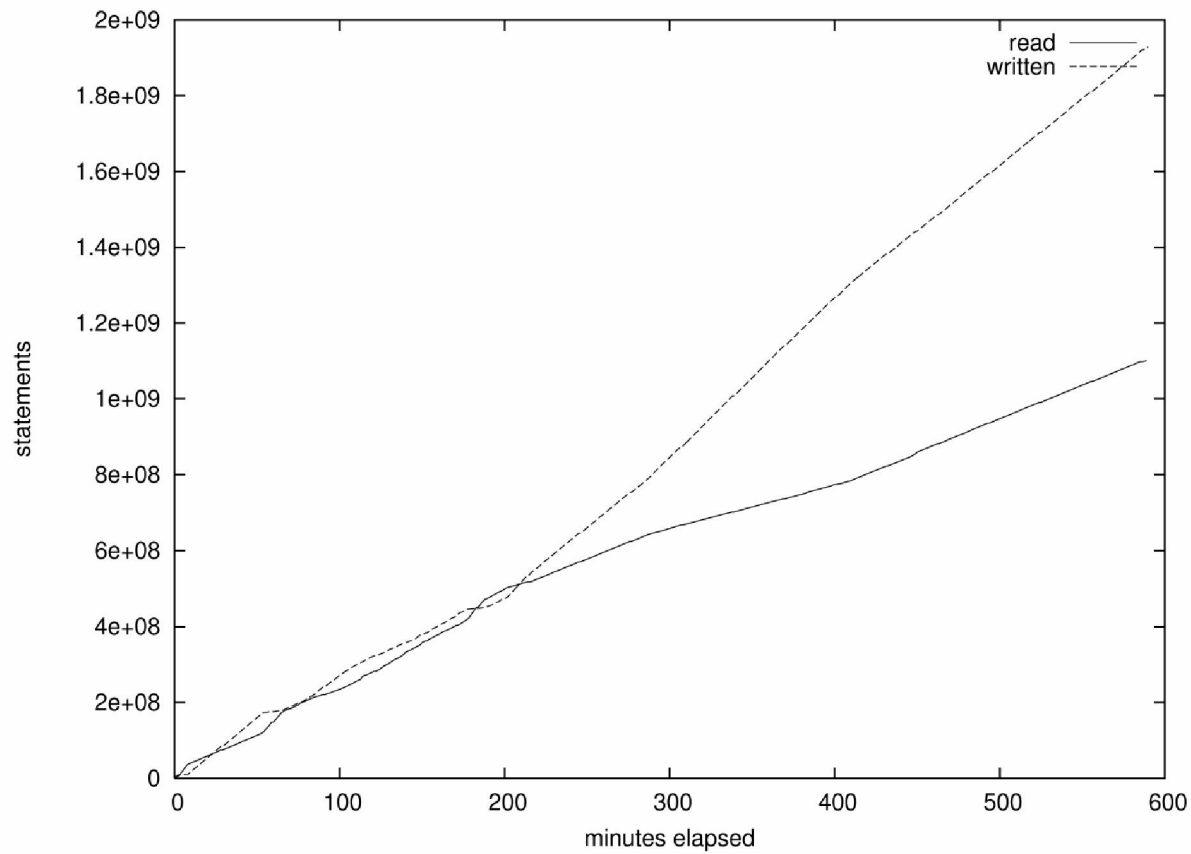
Ontology Hijacking



Evaluation: Scalable Reasoning



Performance

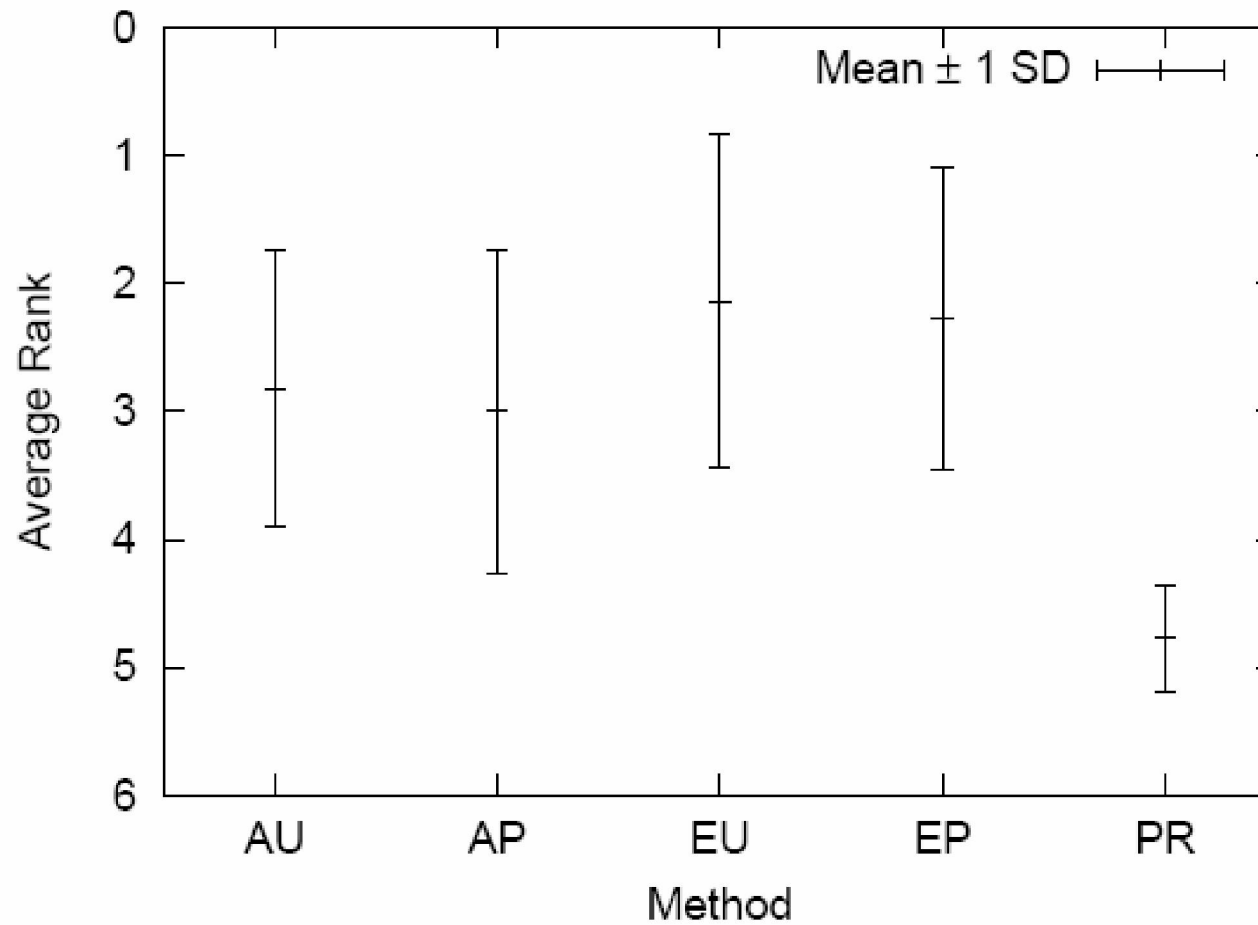


Graph showing SAOR's rate of input/output statements per minute for reasoning on 1.1b statements: reduced input rate correlates with increased output rate and vice-versa

Ranking (IdentifierRank)

- Ranking of data in relational databases (ObjectRank):
 - Domain expert provides „schema authority graph“ manually
 - Apply PageRank where schema authority graph encodes the strength of links
- Problem: in the BTC dataset there are tens of thousands of vocabulary terms
- Solution: use different level of abstraction to construct graph for PageRank calculation
 - Reuse of an URI coined by data source B in data source A counts as vote of A for B
 - Rank data sources
 - Propagate rank of data sources back to URIs of objects

User study (11 users, keyword query for own's name)



Querying/Benchmarks

- Traditional: devise schema and data and queries
 - TPC-H, BSBM, LUBM, ...
- But: web data does not follow any fixed schema

- Solution
 - Create synthetic graphs (e.g. using preferential attachment) that mimic the shape of web data
 - Use abstract query shapes (e.g. star-shaped and path-shaped) and the generated data to create synthetic queries

Faceted Browsing/Navigation

- Traditional: limited variety in vocabulary
 - Possible to pre-compute data guides
 - Test and tweak the UI and the data (taxonomies)

- Web data: huge variety in schema
 - Data guides become very large because data is irregular
 - No clear taxonomic organisation of data
 - Long property chains

- Question: what operations are required? What operations are still possible?

Faceted Browsing: Relational vs Web

■ Ten properties

vs

tens of thousands

▼ Categories

Cell Phones & PDAs (9,996)

Cell Phone & PDA Accessories (10,448)

Cell Phones & Smartphones (341)

Wholesale Lots (125)

Bluetooth Accessories (35)

Other (3)

More ▼

[See all categories](#)

In Cell Phone & PDA Accessories

▶ Type

▶ Price

▶ Compatible Brand

▶ Condition

▶ Seller

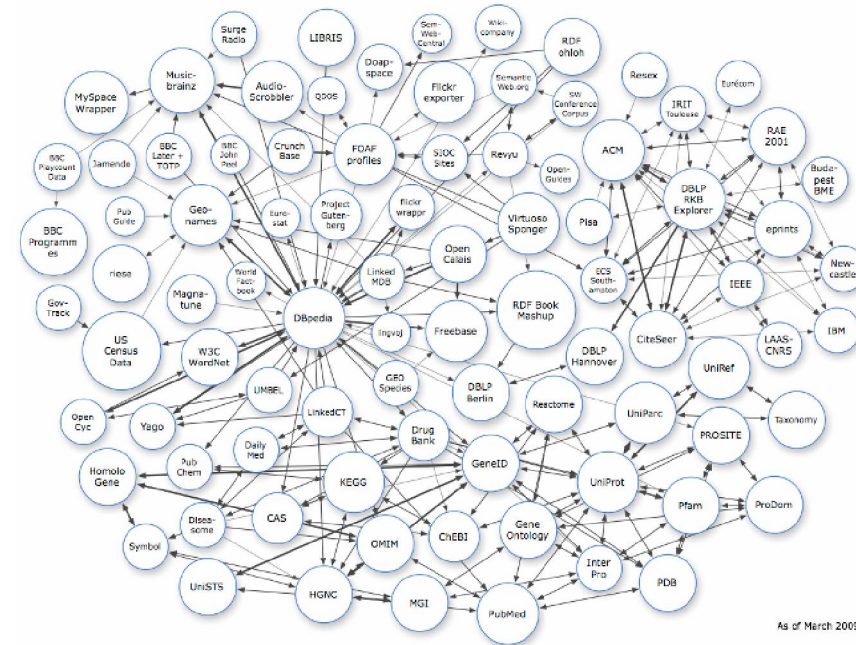
Preferences

▶ Buying formats

▶ Show only

▶ Location

▶ Distance



Faceted Browsing over Relational Data vs Web Data

- Previously: mostly facets because there's only one set of data items of interest (e.g. products)
- Web data: many data items of interest (people, organisations, locations, events, drugs, proteins ...) with vastly different schemas
- Web data: much more interconnections
- Solution: add path traversal operation (from people to companies to locations to events...)

- Previously: aggregate facets and compute possible number of results per facet
- Web data: not feasible (too expensive) due to data irregularity
- Solution: don't aggregate facets but use objects as prototypes to select parameters for subsequent operations

Conclusion

- „semi-structured“ web data is of huge variety
- Systems exploiting full potential of web data have to be schema agnostic
- Ranking as substitute for schema
- New challenges involved, but it's fun to tackle them and to devise more flexible algorithms and systems

References

- 6 indices over quads: Harth and Decker, LA-WEB 2005
- SOAR: Hogan, Harth, Polleres, ASWC 2008
- IdRank: Harth, Kinsella, Decker, ISWC 2009
- VisiNav: Harth, DEXA 2009, ISWC Semantic Web Challenge 2009