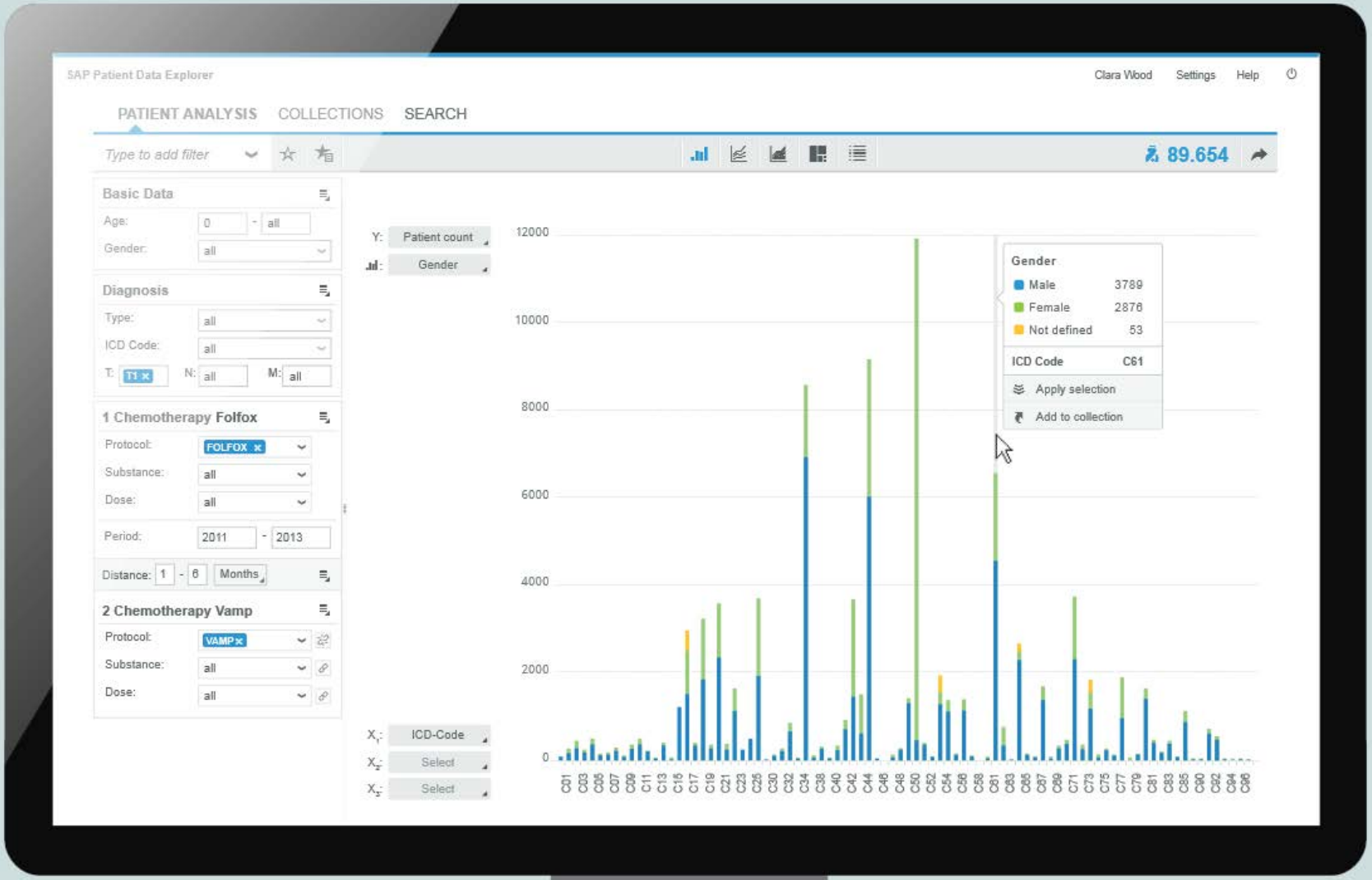
Towards a Semantic Clinical Data Warehouse: A Case Study of Discovering Similar Genes

2015-05-31, Know@LOD – ESWC

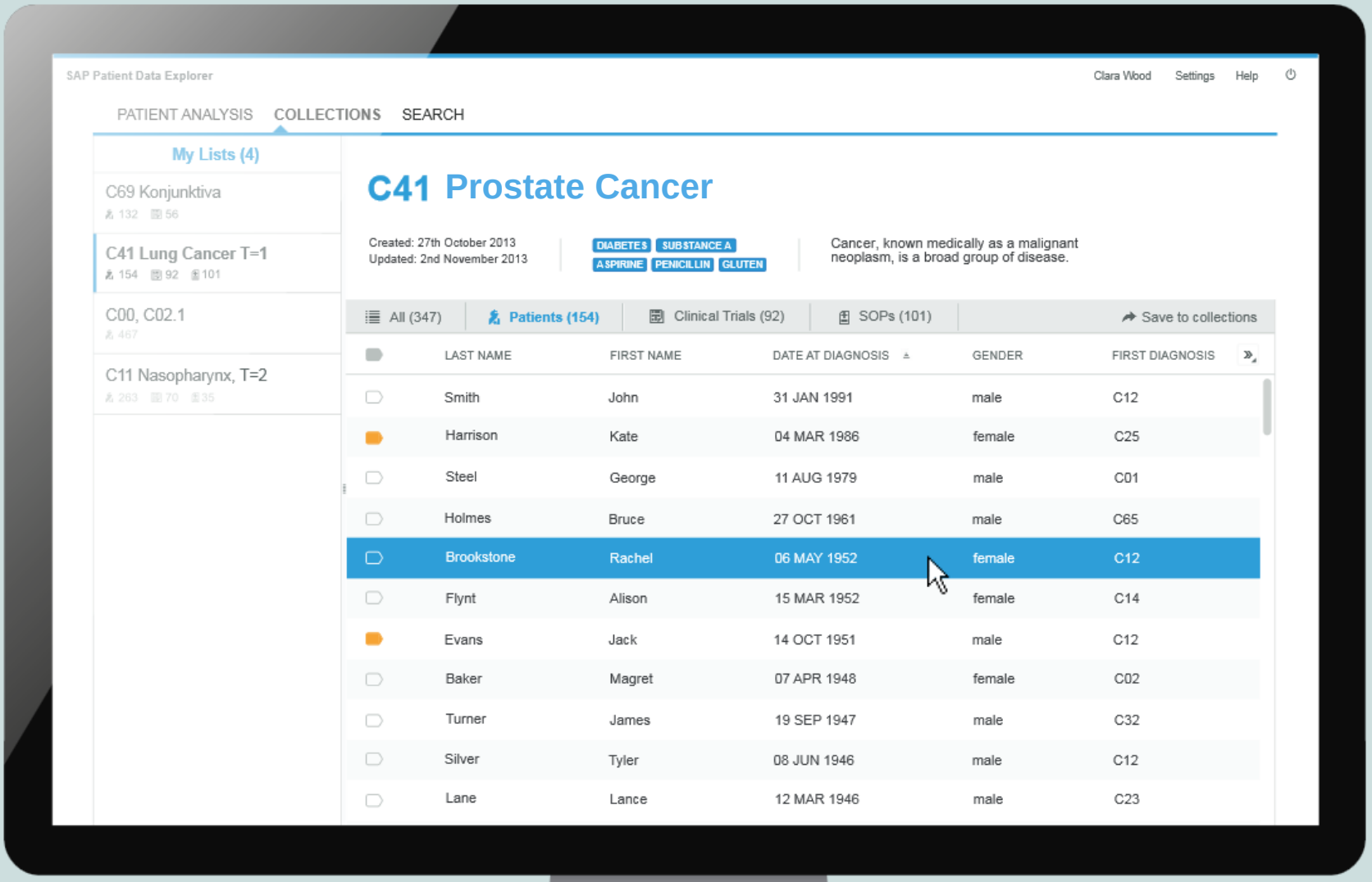
Benedikt Kämpgen, Horst Werner,
Radwan Deeb, and Christof Bornhövd

FZI FORSCHUNGSZENTRUM INFORMATIK

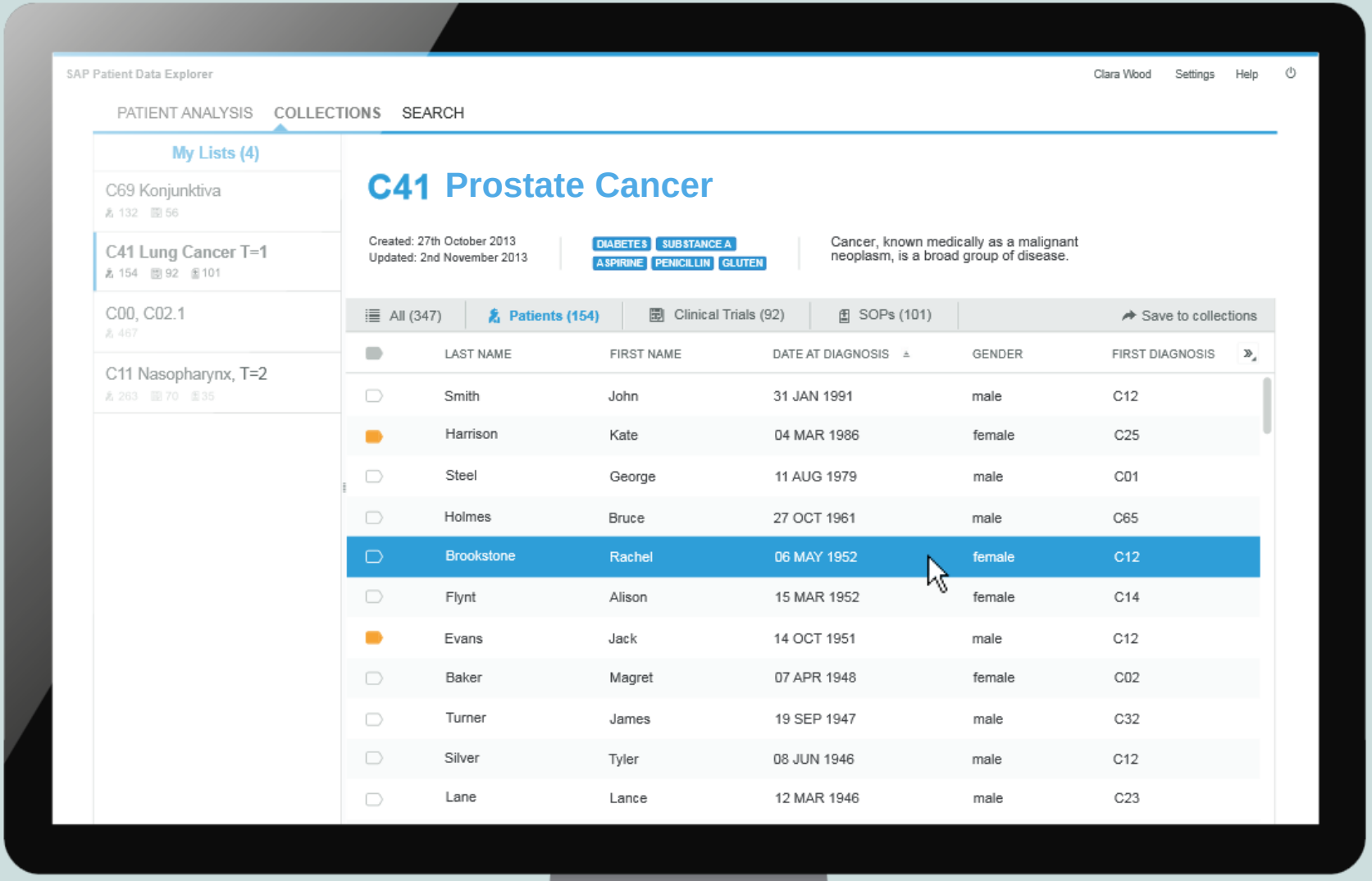
Patient Data Explorer



Patient Data Explorer (2)



Oh, yeah?



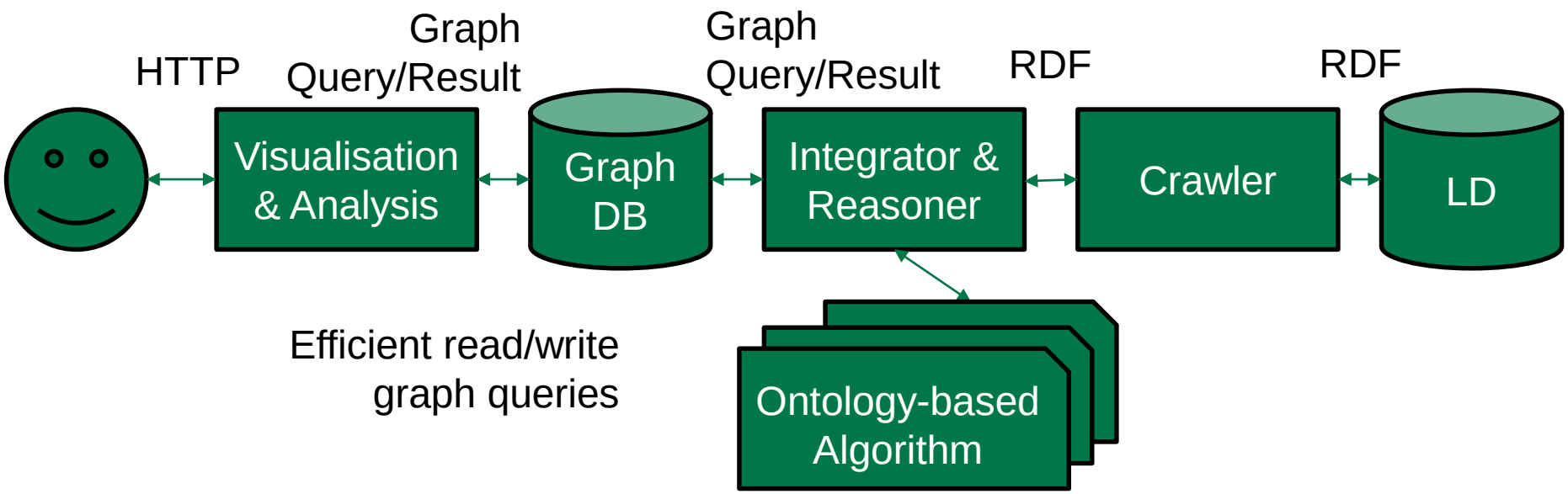
Problems

- ETL + RDBMS need adaptation if additional background information to be considered (e.g., PubMed references)
- No automatic interpretation and knowledge discovery (e.g., possibly wrong information)
- Analysis results not written back to data warehouse (e.g., for provenance tracking and information sharing)

Semantic Clinical Data Warehouse

Flexible visualisation of graph data

Easy access to self-descriptive data (Linked Data)



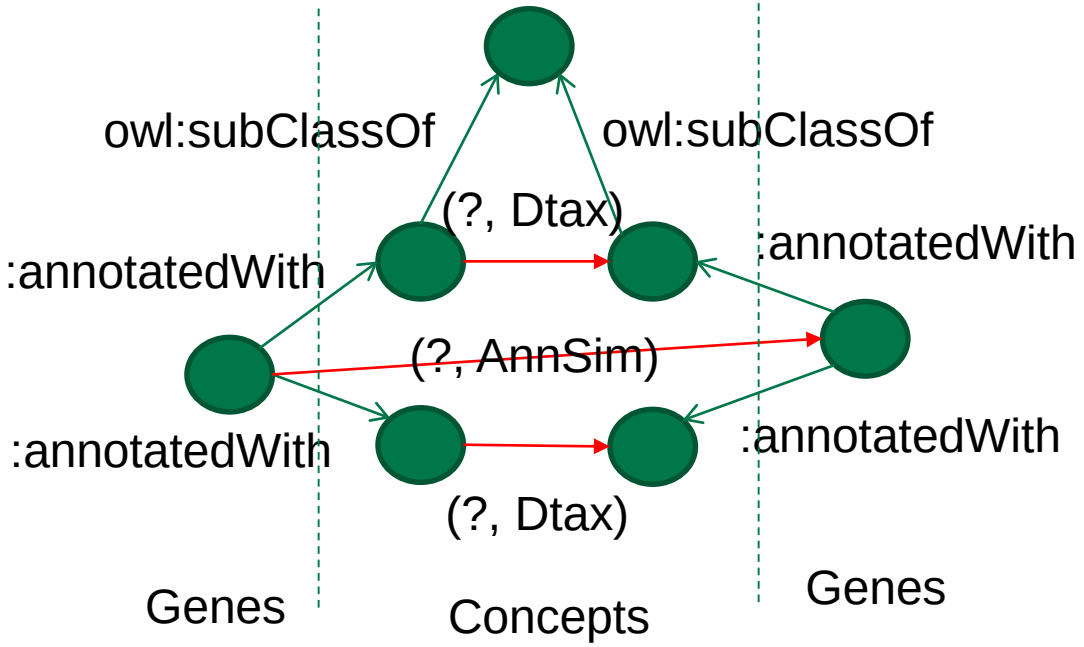
Efficient read/write graph queries

Automatic discovery of new information (Ontology-based Algorithm)

Application: HANA Linked Data AnnSim (HLA)

- Use Case
  - Discovering similar genes of plant *Arabidopsis Thaliana*
  - Similarity is an important basis for other relationships

- Implementation uses
  - SAP HANA Graph
  - AnnSim (Palma, '13)



Application: Setup

	HLA (Java)	AnnSim 1.0 (C/C++)
Objects	20 genes (1-aaap)	20 genes (1-aaap)
Hardware		
Client Workstation	Ubuntu 14.04 VM on W7 Intel Core i5-3360M CPU 2.80GHz, 16 GB RAM	Ubuntu 14.04 VM on W7 Intel Core i5-3360M CPU 2.80GHz, 16 GB RAM
HANA Instance	SUSE Linux Enterprise Server 11.1 500 GB RAM, 80 cores	-
Data	[1,2,3] (2014)	[1,2,3] (2013)
Triples	7,337,447	-
Size of data	537 MB	2.80 MB
Vertices	601,519	39,209
Edges	1,658,322	74,123

[1] ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_functional_descriptions
[2] http://purl.obolibrary.org/obo/go.owl
[3] ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt

Disclaimer: A direct comparison between implementations is not possible.

Application: Lessons Learned

- Correct Computation of Similarities?

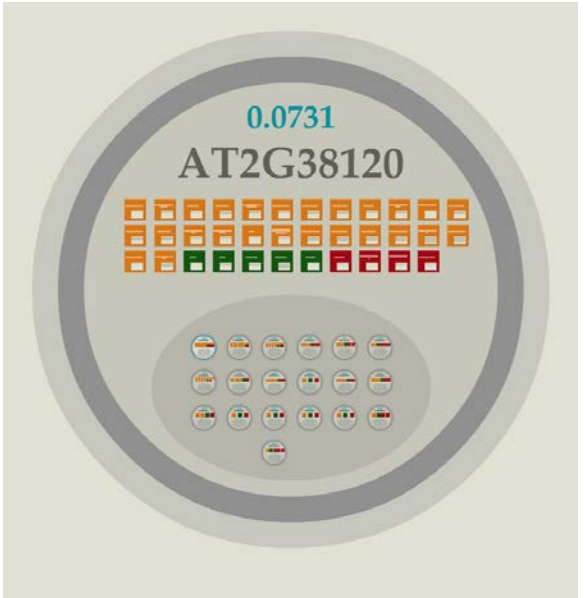
	Mean Squared Error
HLA – AnnSim 1.0	0.09
HLA – SeqSim	0.19
AnnSim 1.0 – SeqSim	0.36

- Efficient Computation of Similarities?

Approach	Load Graph	Compute AnnSim	Read Queries	Write Queries
HLA	370s	2,667s	230s	2,202s
AnnSim 1.0	0s	408s	-	-

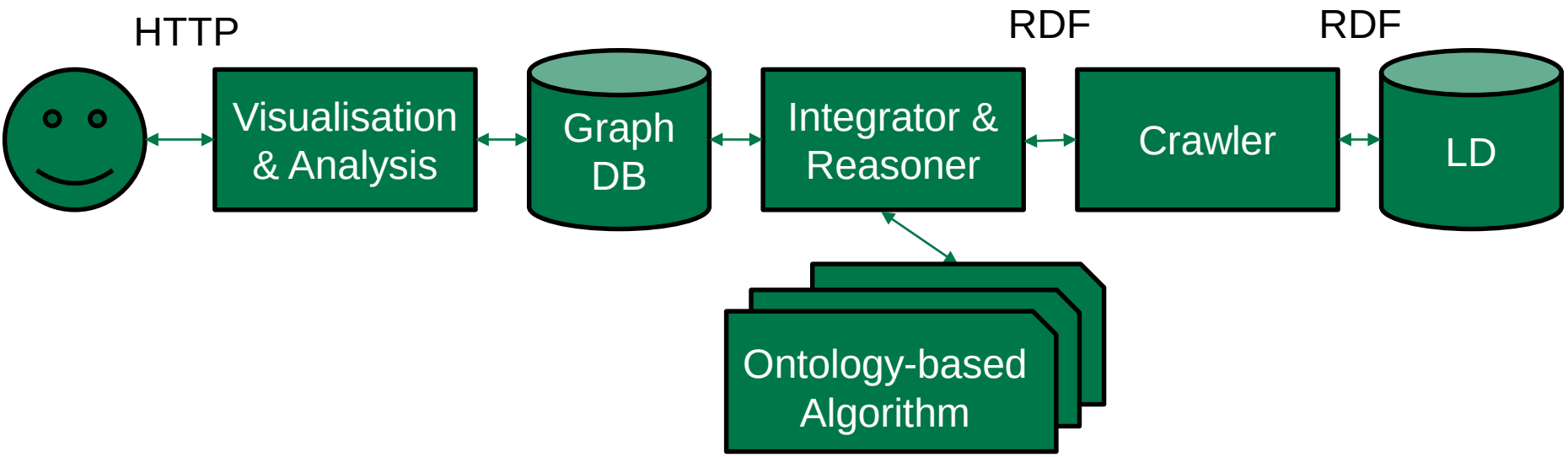
- Flexible Visualisation of Similarities?

http://people.aifb.kit.edu/bka/hla/



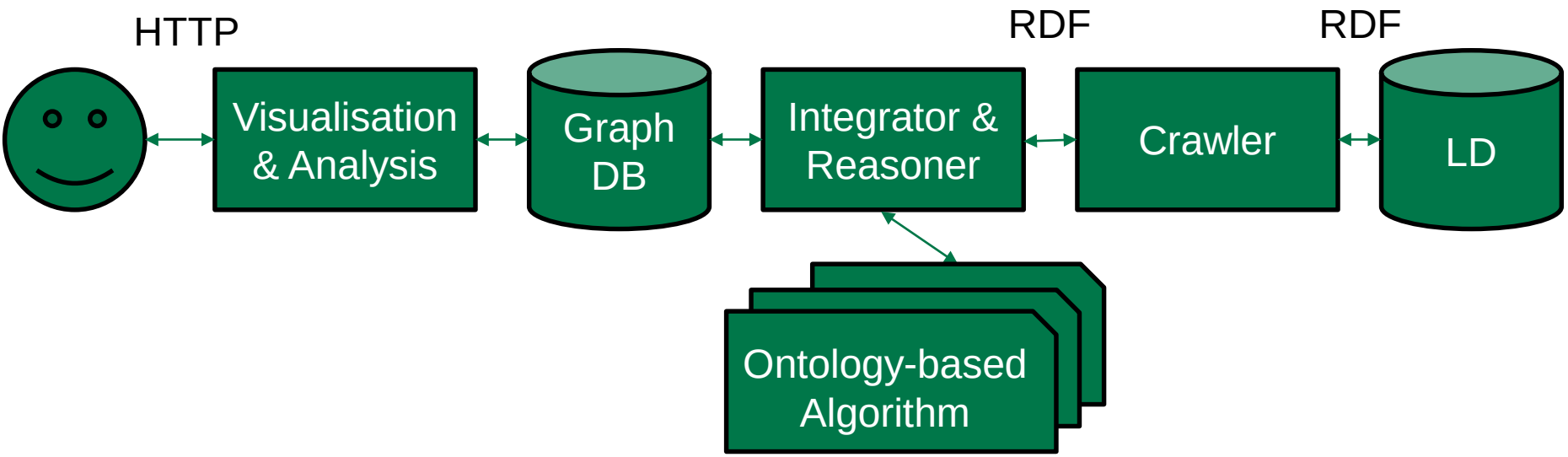
Conclusion

- A Semantic Clinical Data Warehouse should allow
  - Better results by providing more links (e.g., Bio2RDF)
  - Similarities between any objects (Pesquita, '09)
  - New algorithms (Callahan, '12)



Thanks!

- A Semantic Clinical Data Warehouse should allow
  - Better results by providing more links (e.g., Bio2RDF)
  - Similarities between any objects (Pesquita, '09)
  - New algorithms (Callahan, '12)



References

Palma, G., Vidal, M.E., Haag, E., Raschid, L., Thor, A.: Measuring Relatedness Between Scientific Entities in Annotation Datasets. In: International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (2013)

Callahan, A., Dumontier, M.: Evaluating Scientic Hypotheses Using the SPARQL Inferencing Notation. In: The Semantic Web: Research and Applications (2012)

Pesquita, C., Faria, D., Falco, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology 5(7) (2009)