INSCI 2017
4th International Conference on Internet Science

Evaluation of Linked, Open Data Sources for Mining Adverse Drug Reaction Signals

Pantelis Natsiavas[1,2], Nicos Maglaveras[1,2] and Vassilis Koutkias[2]

[1] Lab of Computing, Medical Informatics and Biomedical Imaging Technologies, Department of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

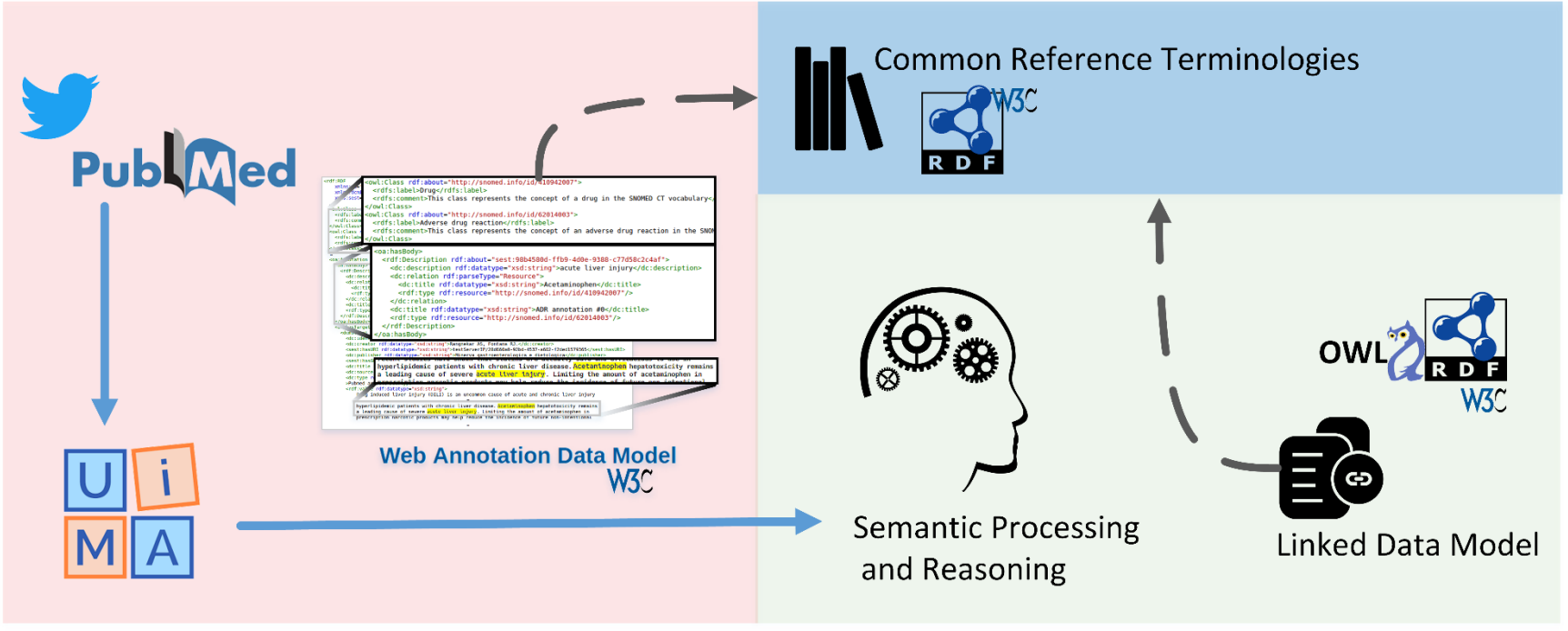
[2] Institute of Applied Biosciences, Centre for Research & Technology Hellas, Thermi, Thessaloniki, Greece

pnatsiavas@certh.gr

Pharmacovigilance Signals: Definition

" … information that arises from one or multiple sources (including observations and experiments), which suggests a new potentially causal association OR a new aspect of a known association, between an intervention and an event …"

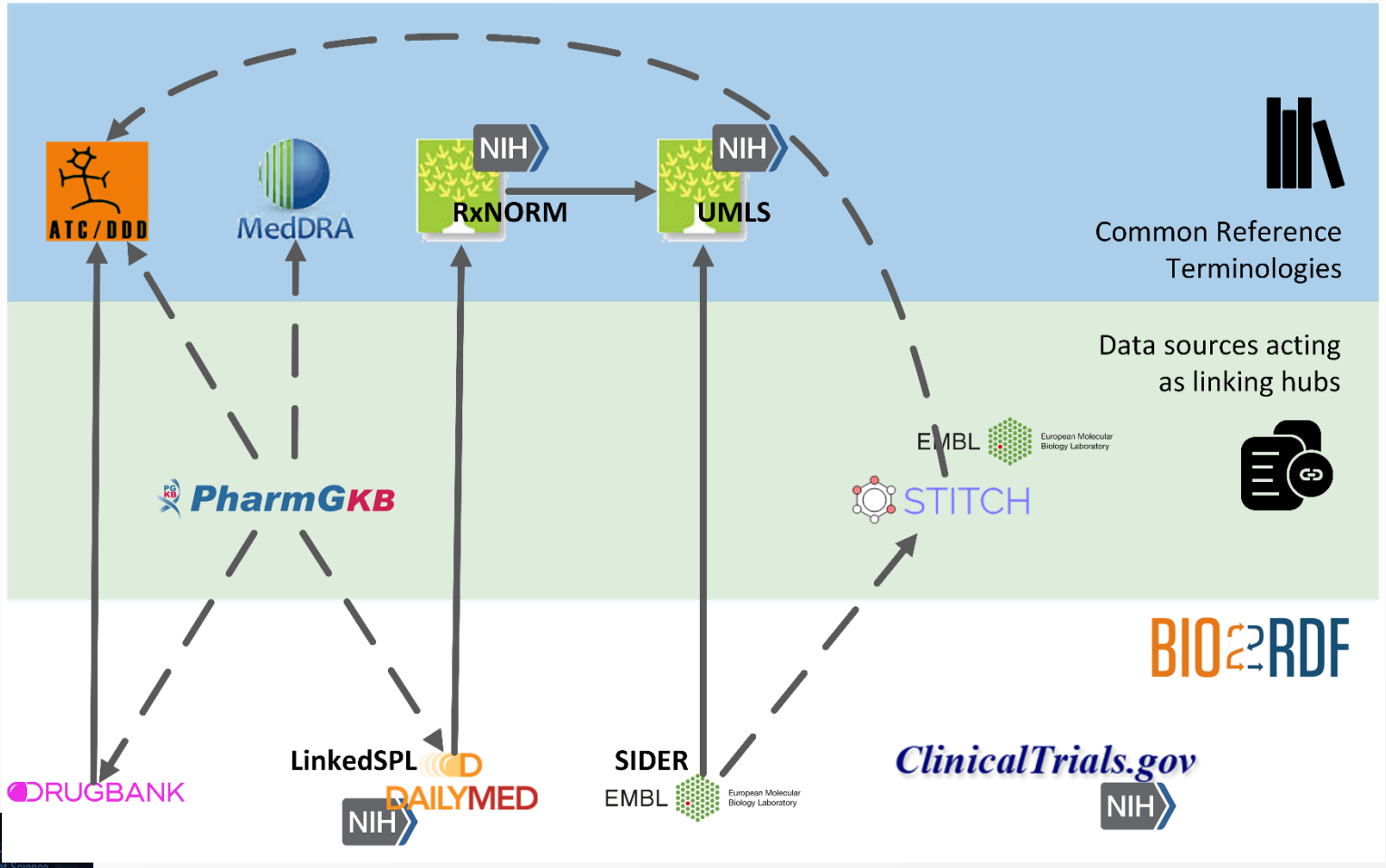
The Goal of this Work



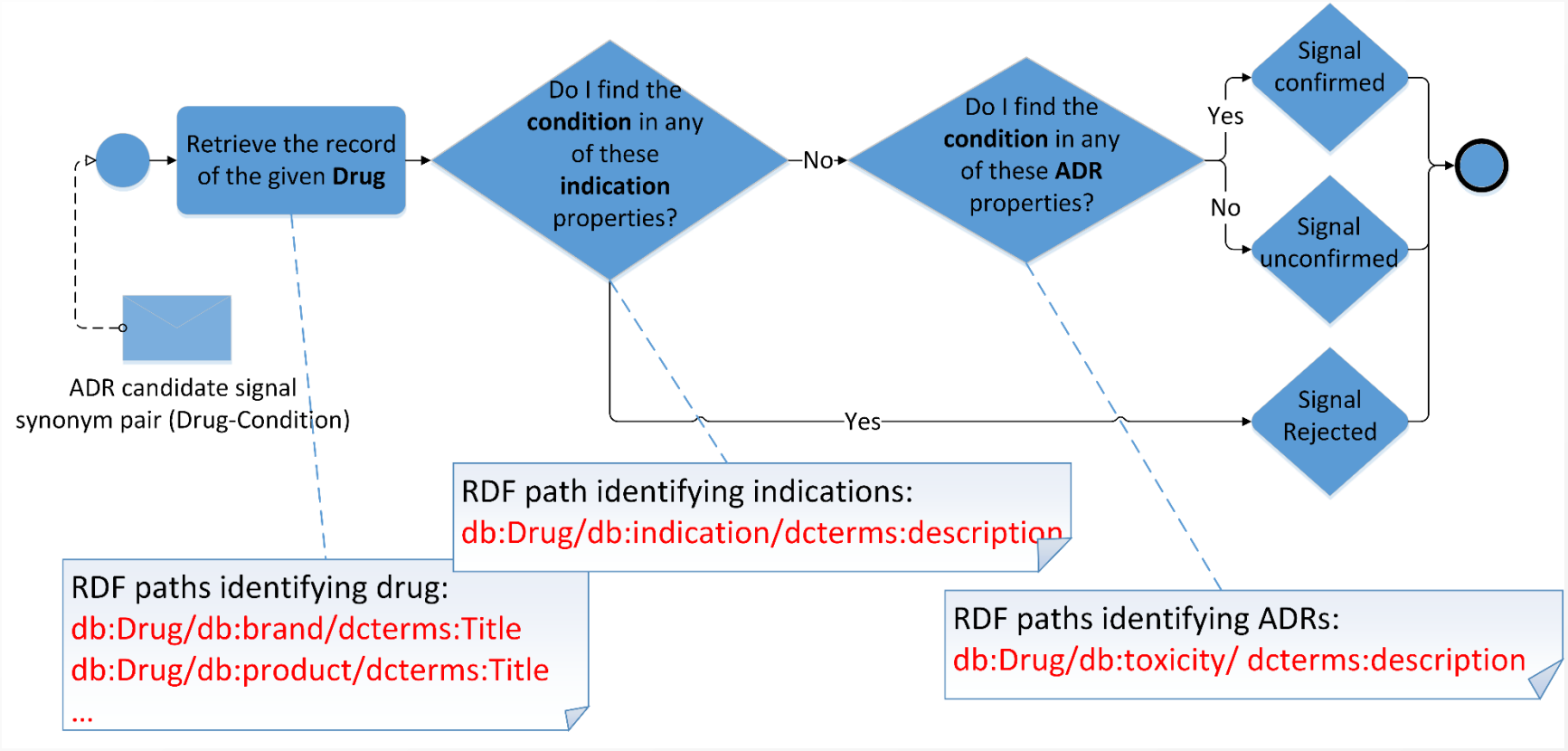
Linked Data Model Requirements

- Data richness
- Semantic richness
- Semantic normalization
- Scalability
- Up-to-date maintenance

Linked Data Model: Overview



Data Richness Evaluation Process



Model Evaluation via Reference Datasets

Dataset	Positive controls	Negative controls	Sum
Harpaz et al.[1]	62	75	137
Ryan et al.[2]	165	233	398
Coloma et al.[3]	44	45	89
Total	**271**	**353**	**624**

[1] Harpaz et al.: A time-indexed reference standard of adverse drug reactions. Sci. Data. 1, (2014).
[2] Ryan, P.B. et al.: Defining a Reference Set to Support Methodological Research in Drug Safety. Drug Saf. 36, 33–47 (2013).
[3] Coloma, P.M. et al.: A Reference Standard for Evaluation of Methods for Drug Safety Signal Detection Using Electronic Healthcare Record Databases. Drug Saf. 36, 13–23 (2012).

Data Richness: Evaluation Results

Evaluation results for the three reference datasets

		SIDER	DrugBank	LinkedSPL	ClinicalTrials.gov
1st	Sensitivity	**0.774**	0.129	0.483	0.322
	Specificity	0.213	**0.906**	0.506	0.64
2nd	Sensitivity	**0.782**	0.176	0.145	0.230
	Specificity	0.408	**0.893**	**0.725**	**0.815**
3rd	Sensitivity	**0.909**	0.295	0.545	0.409
	Specificity	0.111	**0.978**	0.444	0.689

Semantic Richness: Evaluation Results

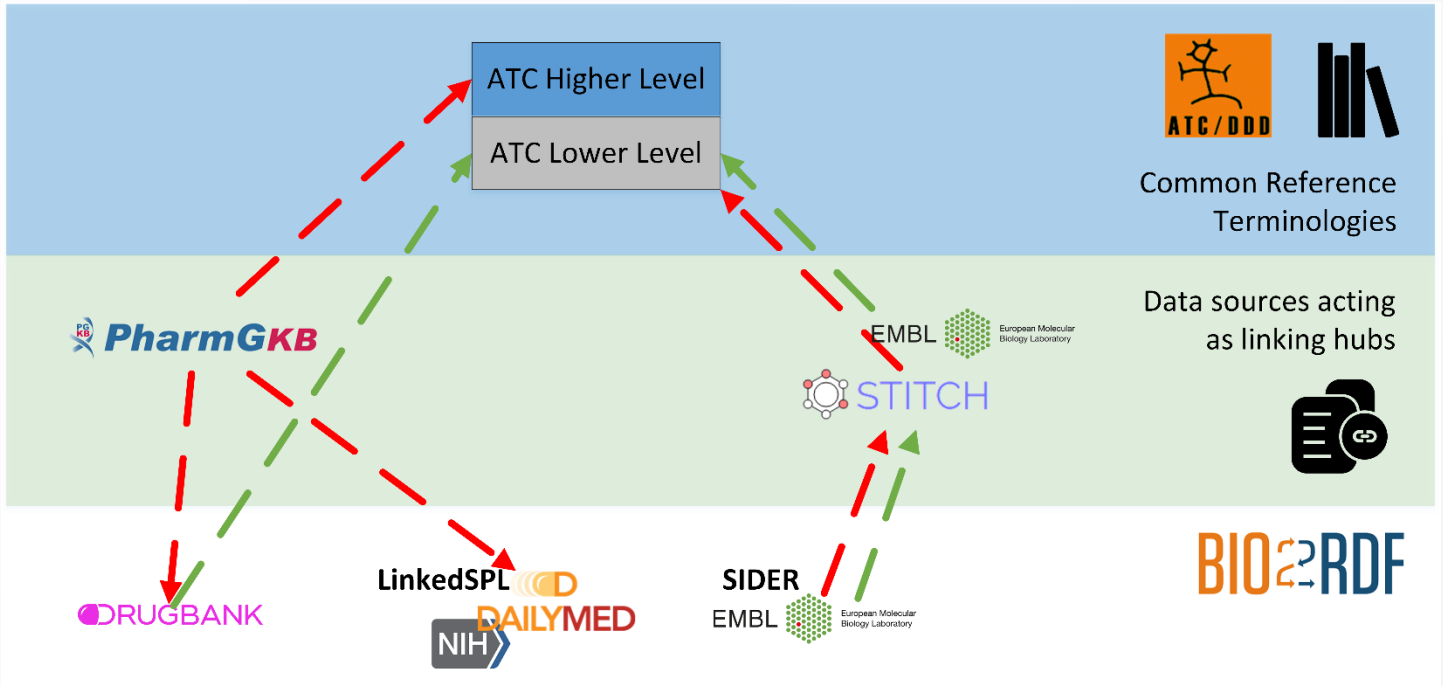
	Classes	Properties	External reference properties
SIDER	15	22	2
DrugBank	104	114	25
LinkedSPL	5	104	3
ClinicalTrials.gov	63	157	0
Totals	**187**	**397**	**30**

Semantic Normalization: Evaluation Results

	Total number of drugs	PharmGKB references	Percentage
SIDER	1,593	1,593	**100%**
DrugBank	8,054	1,625	20.17%
LinkedSPL	51,305	886	1.72%

	Total number of drugs	ATC references	Percentage of ATC references	Common ATC codes	Linking percentage
SIDER	1,593	1,593	**100%**	1049	**65.85%**
DrugBank	8,054	1,739	21.59%		13.02%

Semantic Normalization: Evaluation Results



Scalability: Evaluation Results

Average execution time (in seconds)

	1st query	2nd query	3rd query
SIDER	2.79	22.12	21.53
DrugBank	1.37	1.42	1.55
LinkedSPL	23.58	21.73	22.86
ClinicalTrials.gov	1.70	3.33	3.26

Challenges and Future Work

- Semantic normalization and interlinking
- Semantic enrichment
- Maintenance
- Future work
  - Integrate the proposed model in unstructured data analysis focusing on ADRs
  - Expand the model to accommodate biological data

Discussion

Pantelis Natsiavas

pnatsiavas@certh.gr

INSCI 2017
4th International Conference on Internet Science