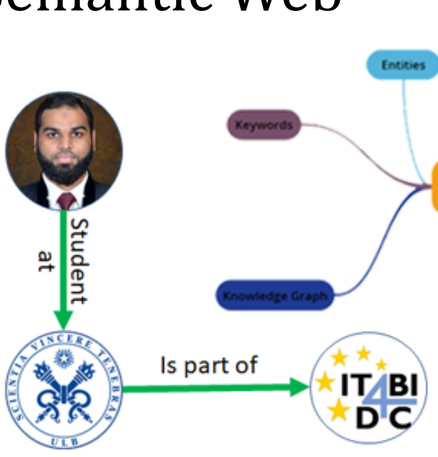
Statistical Multidimensional Data Modeling based on Linked Open Data

Jam Jahanzeb K. Behan *jam.behan@ulb.ac.be*
Esteban Zimányi *Université libre de Bruxelles*
Oscar Romero *Universitat Politècnica de Catalunya*

Semantic Web

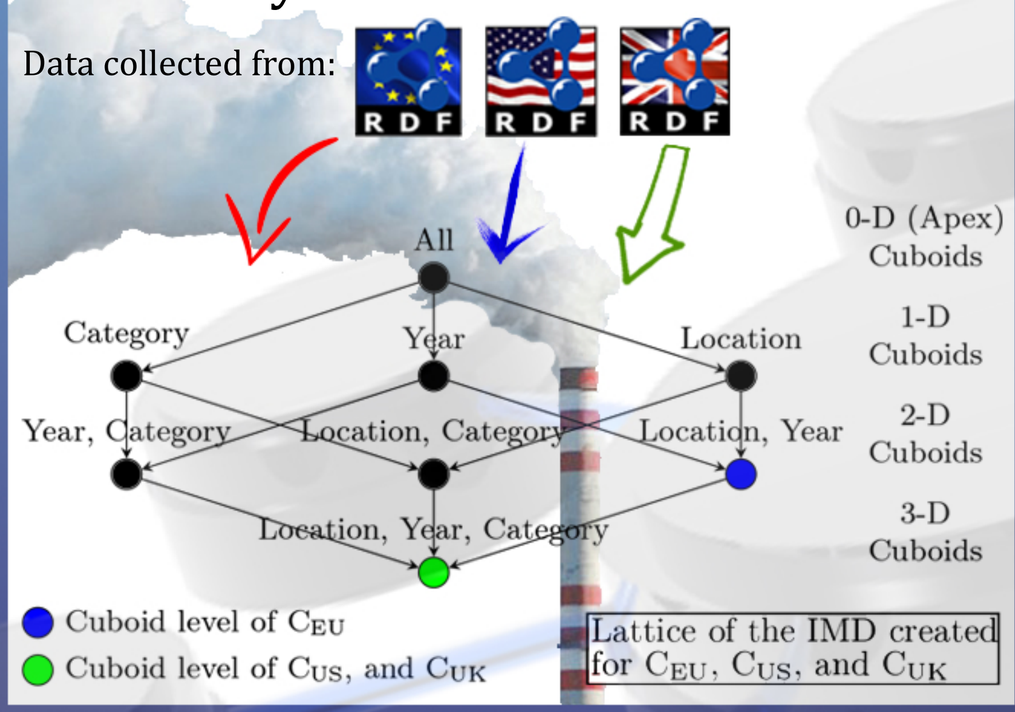


The current focus is to perform integration of RDF triples

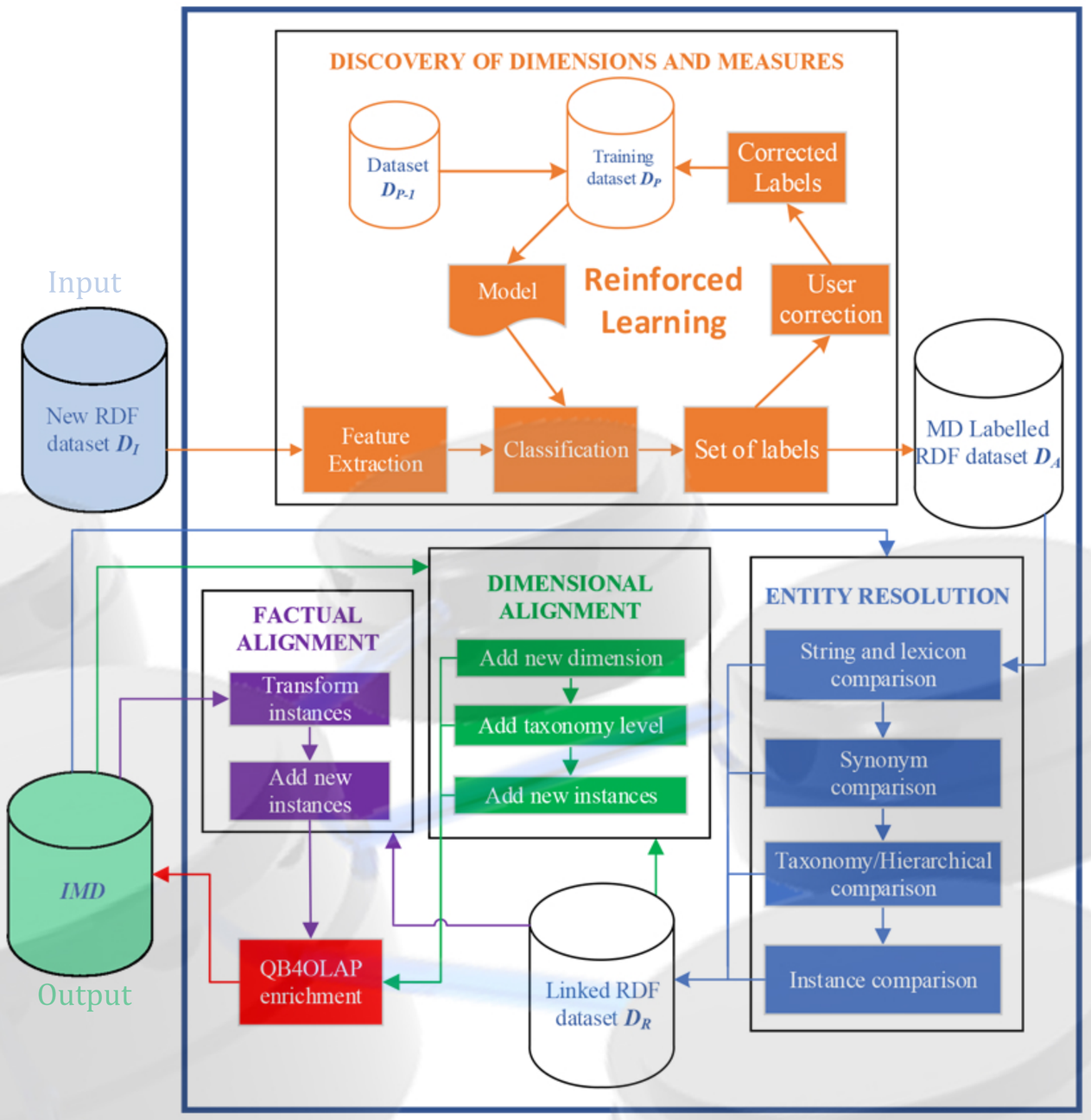
Mesh of (unlinked) information available and the aim is to integrate it by adding multidimensional concepts to the entities

Case Study: Carbon Emissions

Data collected from:



Framework



Experimentation

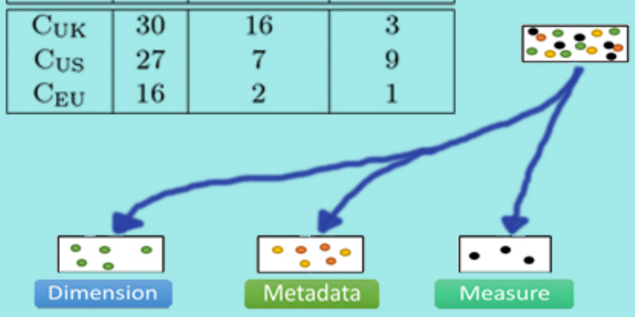
DISCOVERY OF DIMENSIONS AND MEASURES

We add Multidimensional concepts to resources that have been extracted from the RDF files. To add these concepts, we add features to the RDF triples as defined in the adjacent diagram and then label each resurce as either a Dimension, Measure or Metadata

Unique Values	The ratio of unique values based on the total occurrences
Data Types	Such as float, integer, string, Boolean, categorical, date, geolocation, a resource (i.e., a URI) or description (containing metadata information)
URI Prefix and URN	The URI is parsed to obtain these features: <ds:location> is parsed as <ds> and <location>
URI Resource Name Length	The total number of characters in a URI
Additive Property	Identifies numerical type resources as additive or non-additive

Total & labeled resources

Dataset	Total	Dimension	Measures
CUK	30	16	3
CUS	27	7	9
CEU	16	2	1



ENTITY RESOLUTION

After adding the Multidimensional concepts to resources we performed entity resolution operations using state-of-the-art tools LOGMAP and IUT to test the our hypothesis We see that after performing ER operations on DM labelled resources reduces the number of comparisons by 88% and the runtime by 81%

ER with & without labels

Using labels	Comparisons	Run-time (s)
Yes	201	31
No	1658	165

Rule 1 Given two resources, d_1 and d_2, d_1 is the same resource as d_2 if there is an equivalence when considering the lemmas in the names of both resources:

$$\{d_1,d_2\} \mapsto \{d_2\} \text{ iff } d_1 \equiv d_2 \text{ where } d_1 \in D_A \wedge d_2 \in IMD$$

Rule 2 Given two resources, d_1 and d_2, d_1 is the same resource as d_2 if there is an equivalence (i.e., same lemma) when considering the synonym map (S_d) of both resources:

$$\{d_1,d_2\} \mapsto \{d_2\} \text{ iff } d_3 \equiv d_4 \text{ where } d_1 \in D_A \wedge d_2 \in IMD \wedge d_3 \in S_{d1} \wedge d_4 \in S_{d2}$$

Rule 3 Given two resources, d_1 and d_2, d_2 subsumes or is subsumed by d_1 if d_1 and d_2 (or their synonyms) participate in the same hierarchical-map (H_d) either directly or through their synonym map:

$$d_3 \equiv d_4 \wedge d_4 \sqsubseteq d_5 \text{ where } (d_3 \in (d_1 \cup S_{d1}) \wedge d_4 \in (d_2 \cup S_{d2}) \wedge d_5 \in H_{d2}) \vee (d_3 \in (d_2 \cup S_{d2}) \wedge d_4 \in (d_1 \cup S_{d1}) \wedge d_5 \in H_{d1})$$

Rule 4 Given two resources, d_1 and d_2, there is an equivalence if I_{d1}, the instance space of d_1, has an intersection with I_{d2}, the instance space of d_2, that is greater than an input parameter θ, which is the required level of equal instances in both resources:

$$d_1 \equiv d_2 \text{ iff } I_{d1} \cap I_{d2} > \theta \text{ where } I_{d1} = \text{instances of } d_1 \wedge I_{d2} = \text{instances of } d_2 \wedge \theta \in \mathbb{R}$$

References

Bizer, C., Heath, T., Berners-Lee, T.: Linked Data: The Story so Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)

Etcheverry, L., Vaisman, A.A.: QB4OLAP: A New Vocabulary for OLAP Cubes on the Semantic Web. In: Proceedings of the 3rd International Conference on Consuming Linked Data. vol. 905, pp. 27–38. CEUR-WS.org (Nov 2012)

Jam Jahanzeb Khan Behan, Oscar Romero, Esteban Zimányi: Multidimensional Integration of RDF datasets. 21st International Conference on Big Data Analytics and Knowledge Discovery – DaWaK 2019 (accepted)

Jiménez-Ruiz, E., Grau, B.C.: LogMap: Logic-Based and Scalable Ontology Matching. In: Proceedings of the 10th International Semantic Web Conference. pp. 273–288. Springer (Oct 2011)

Zong, N.: Instance-based Hierarchical Schema Alignment in Linked Data. Ph.D. thesis, Seoul National University Graduate School, Seoul, South Korea (2015)

https://catalog.data.gov/dataset/2015-greenhouse-gas-report-data

https://opendata.camden.gov.uk/resource/4txj-pb2i

http://estatwrap.ontologycentral.com/page/t2020_rd300

Ninth European Business Intelligence and Big Data Summer School - 2019

Erasmus+