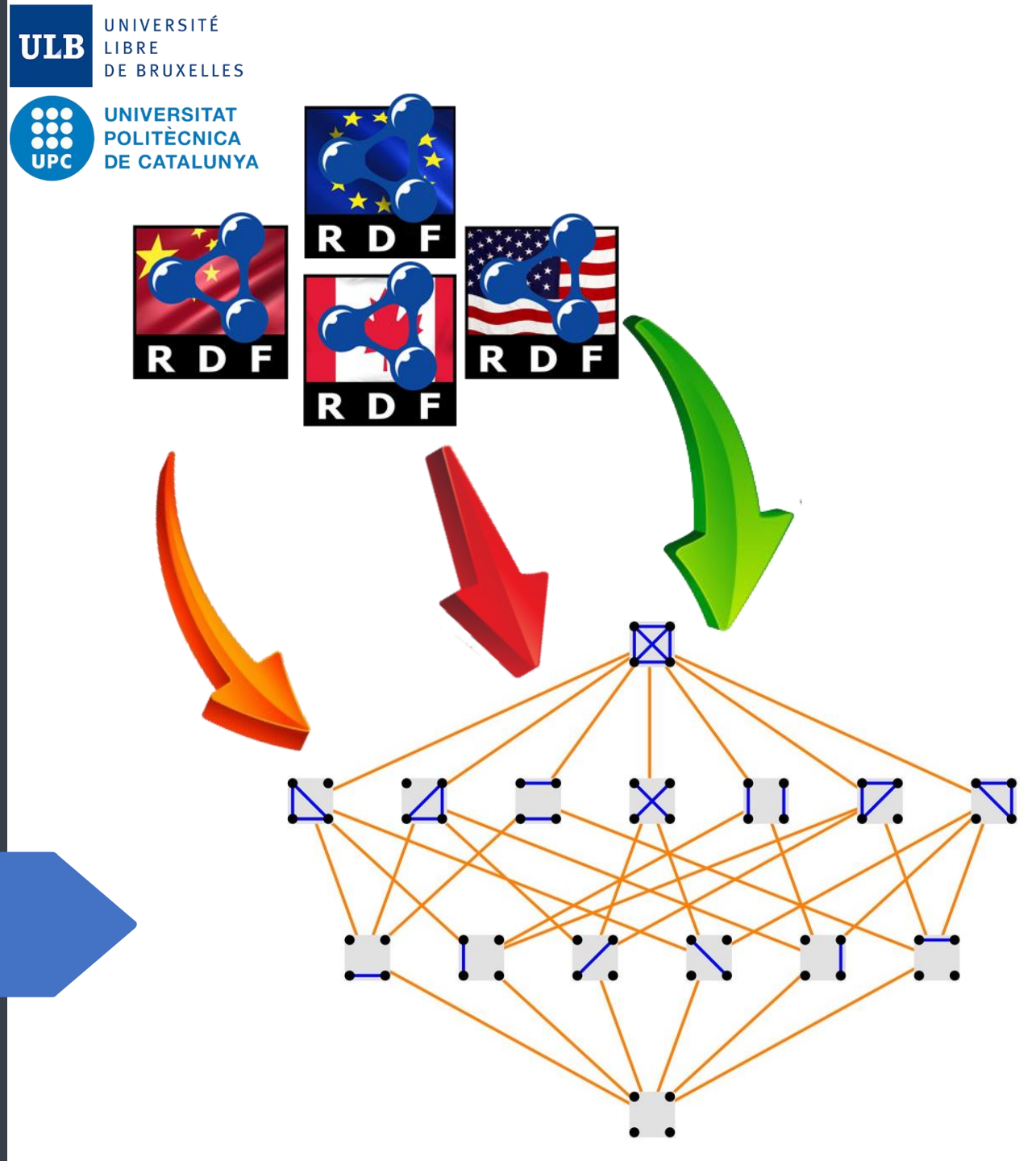
Statistical Multidimensional Data Modeling based on Linked Open Data

Jam Jahanzeb Khan Behan
jam.behan@ulb.ac.be

Esteban Zimányi (Supervisor)
Université Libre de Bruxelles

Òscar Romero (Co-supervisor)
Universitat Politècnica de Catalunya

Ninth European Business Intelligence and Big Data Summer School (eBISS) 2019



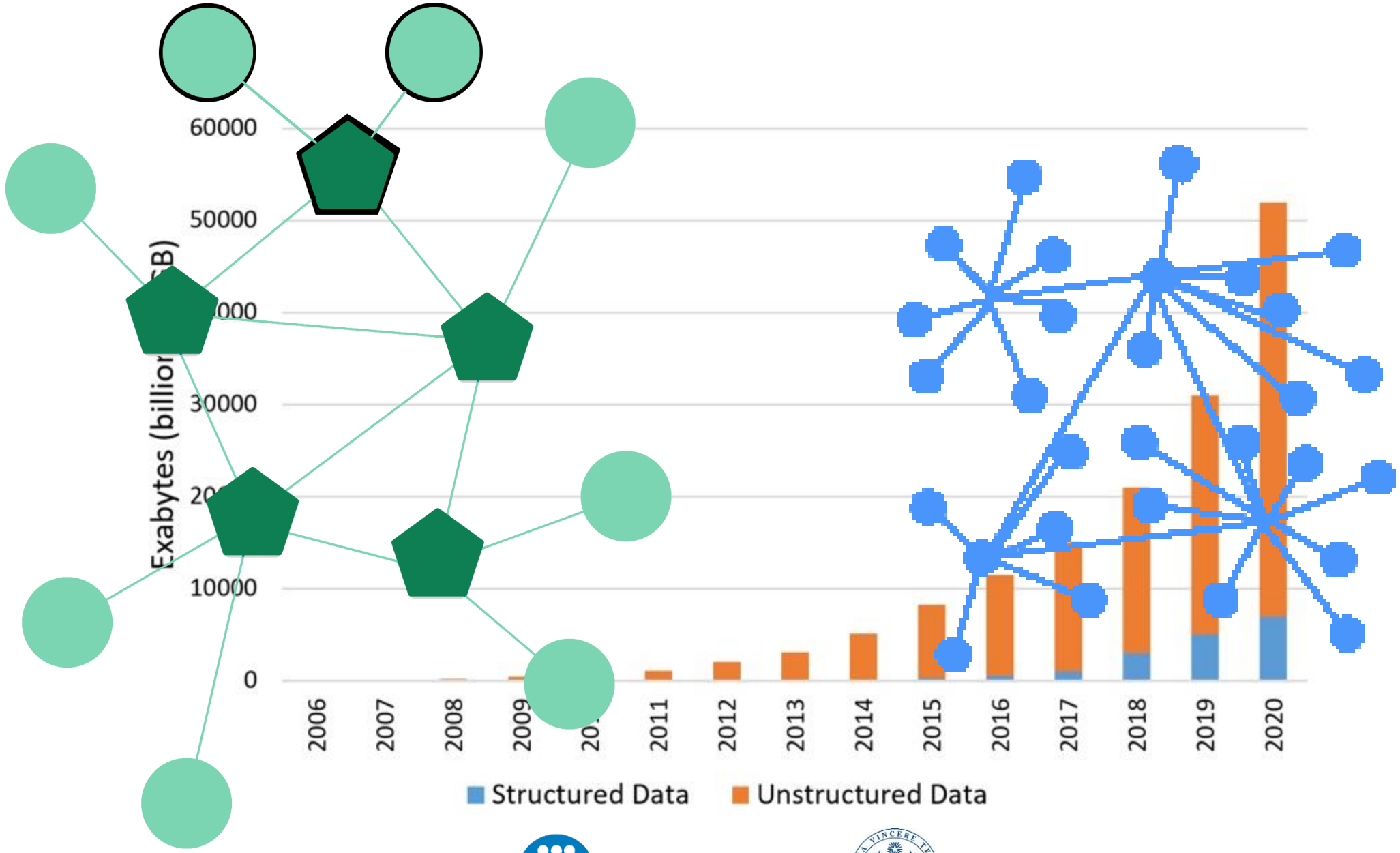
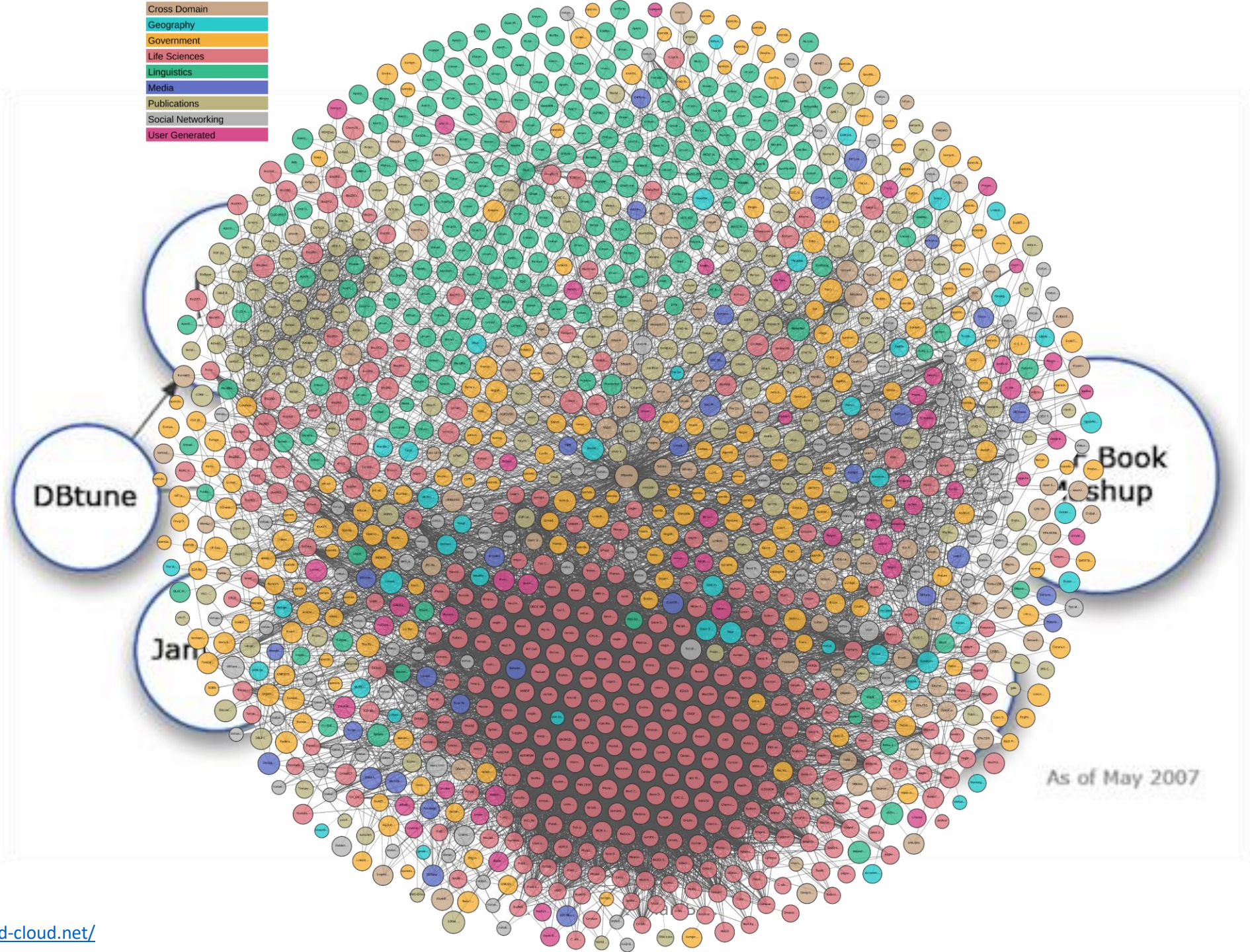


*Image from https://www.eetimes.com/author.asp?section_id=36&doc_id=1330462#



*Images from https://lod-cloud.net/

RDF Triples

Associate Professor at

Associate Professor at

Host university supervisor

Home university supervisor

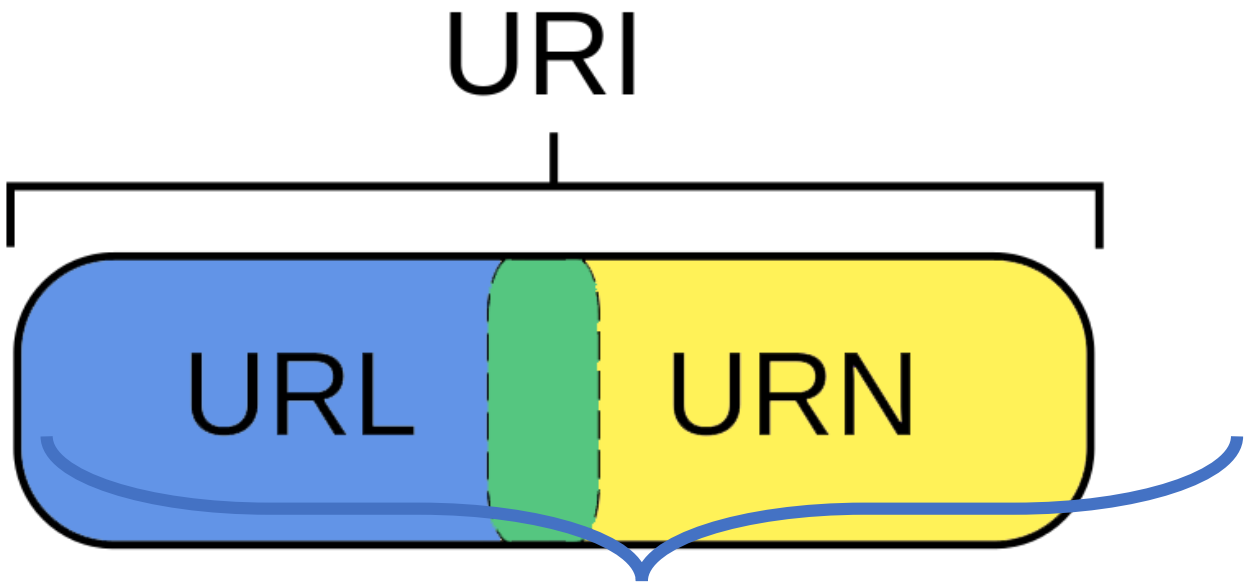
Professor at

subject

predicate

object

*Images from https://www.linkedin.com/



https://opendata.camden.gov.uk/resource/4txj-pb2i/carbon_emissions_kgco2e

URN

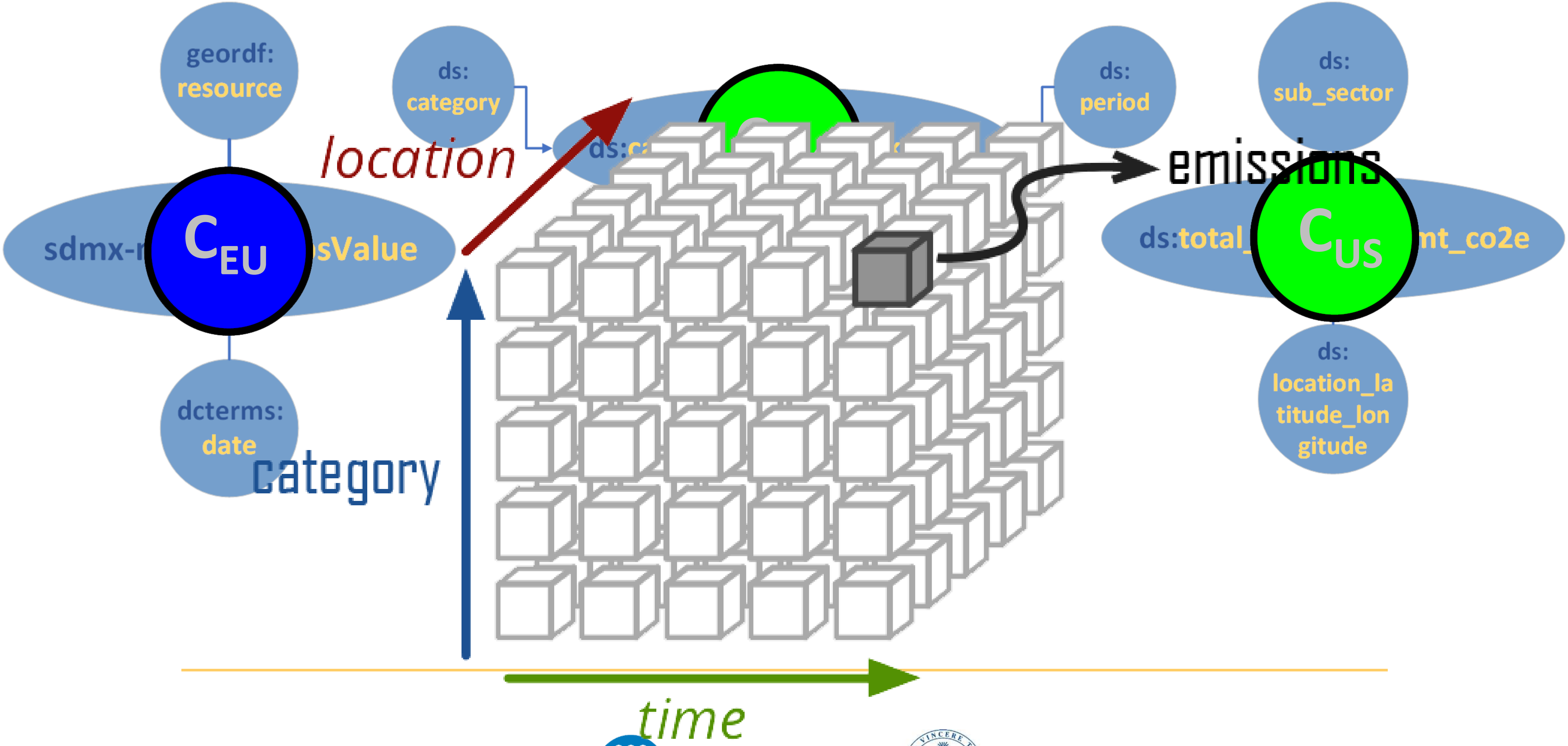
ds:carbon_emissions_kgco2e

Use case: Carbon Emission

sdmx-measure:obsValue

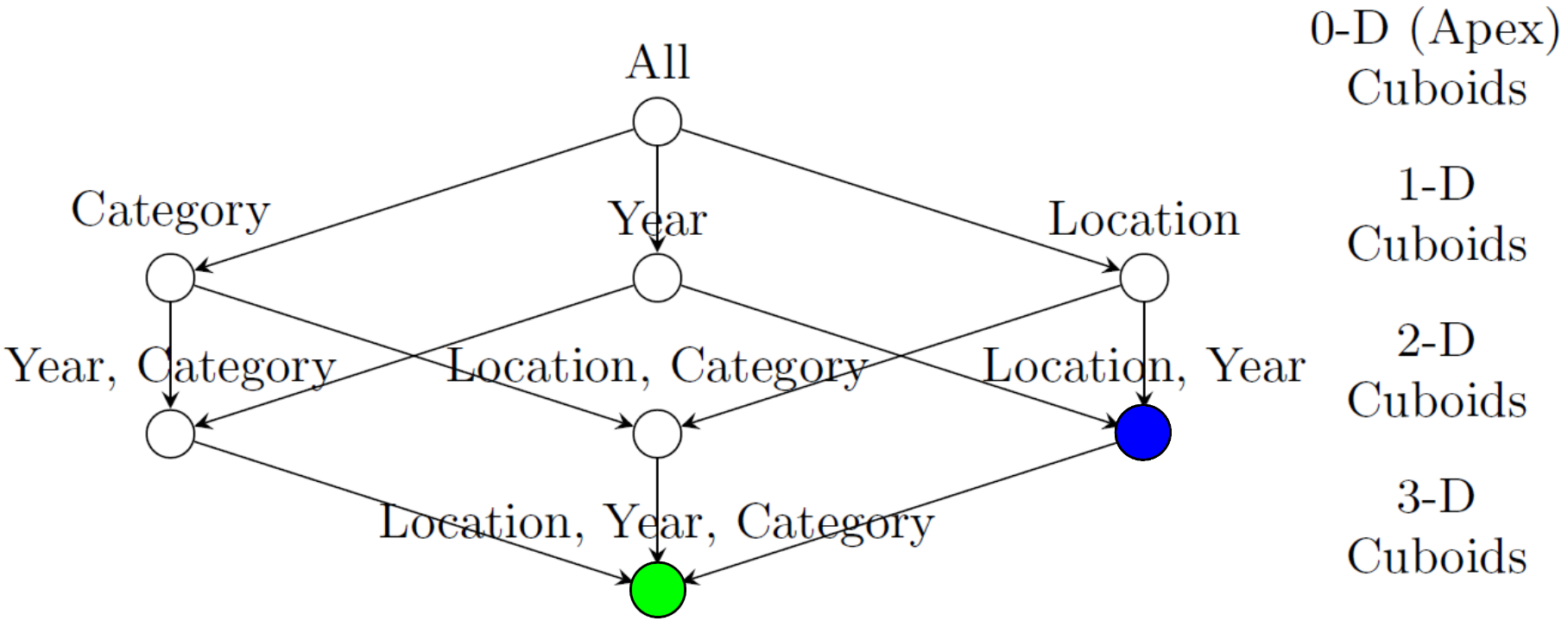
ds:carbon_emissions_kgco2e

ds:total_emissions_mt_co2e



geordf:
resource
ds:
category
location
ds:c
ds:
period
ds:
sub_sector
emissions
sdmx-r
C_{EU}
sValue
ds:total_
C_{US}
mt_co2e
dcterms:
date
category
ds:
location_la
titude_lon
gitude
time

Multidimensional Integration



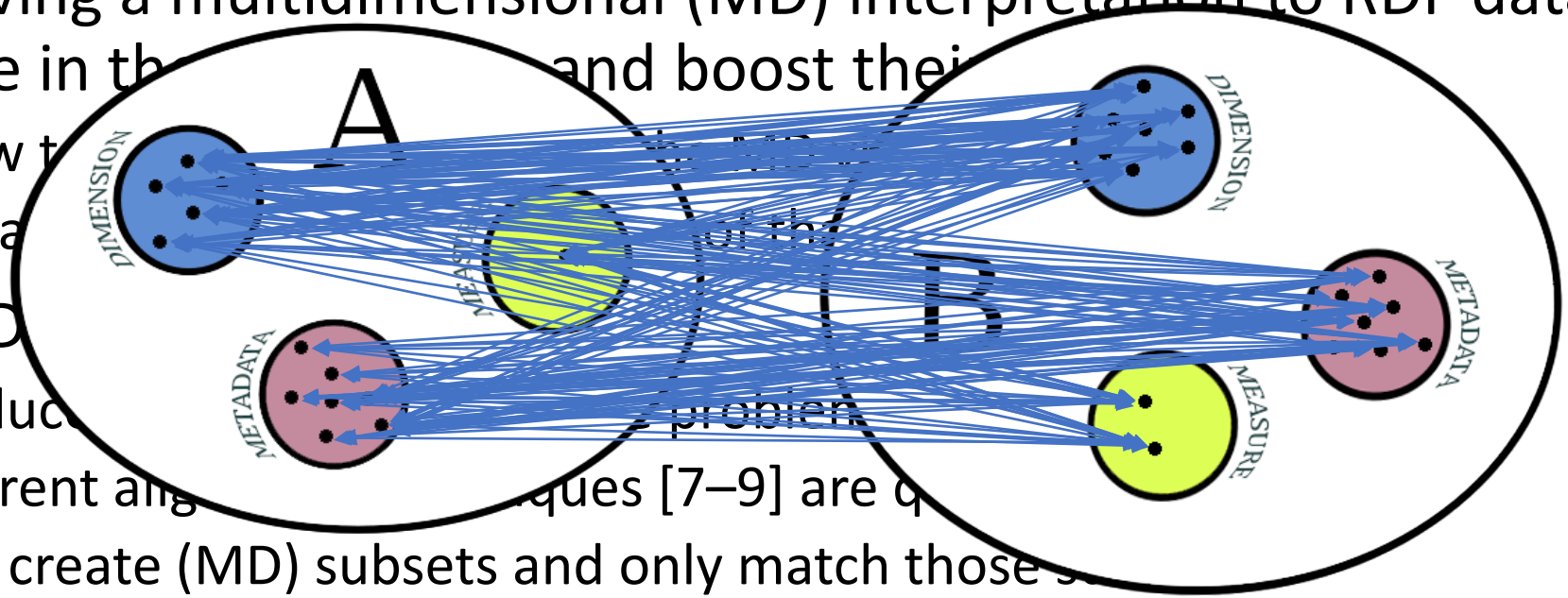
Research Question

Does giving a multidimensional (MD) interpretation to RDF datasets facilitate in th and boost thei

- How t
- Wha

Why MD

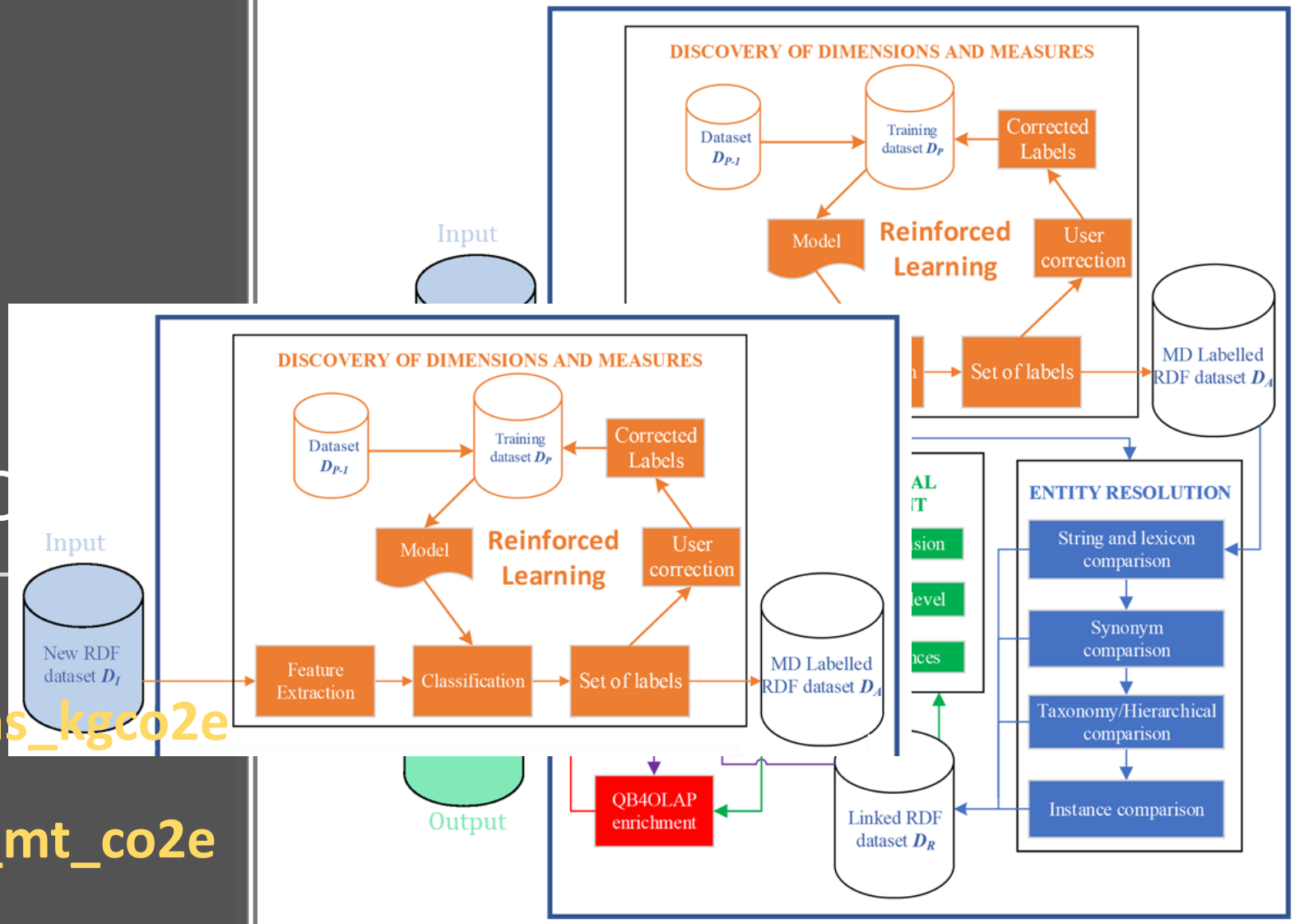
- Reduc
- Current alig ques [7–9] are q
- We create (MD) subsets and only match those s
- Label each resource as a Dimension, Measure or Metadata

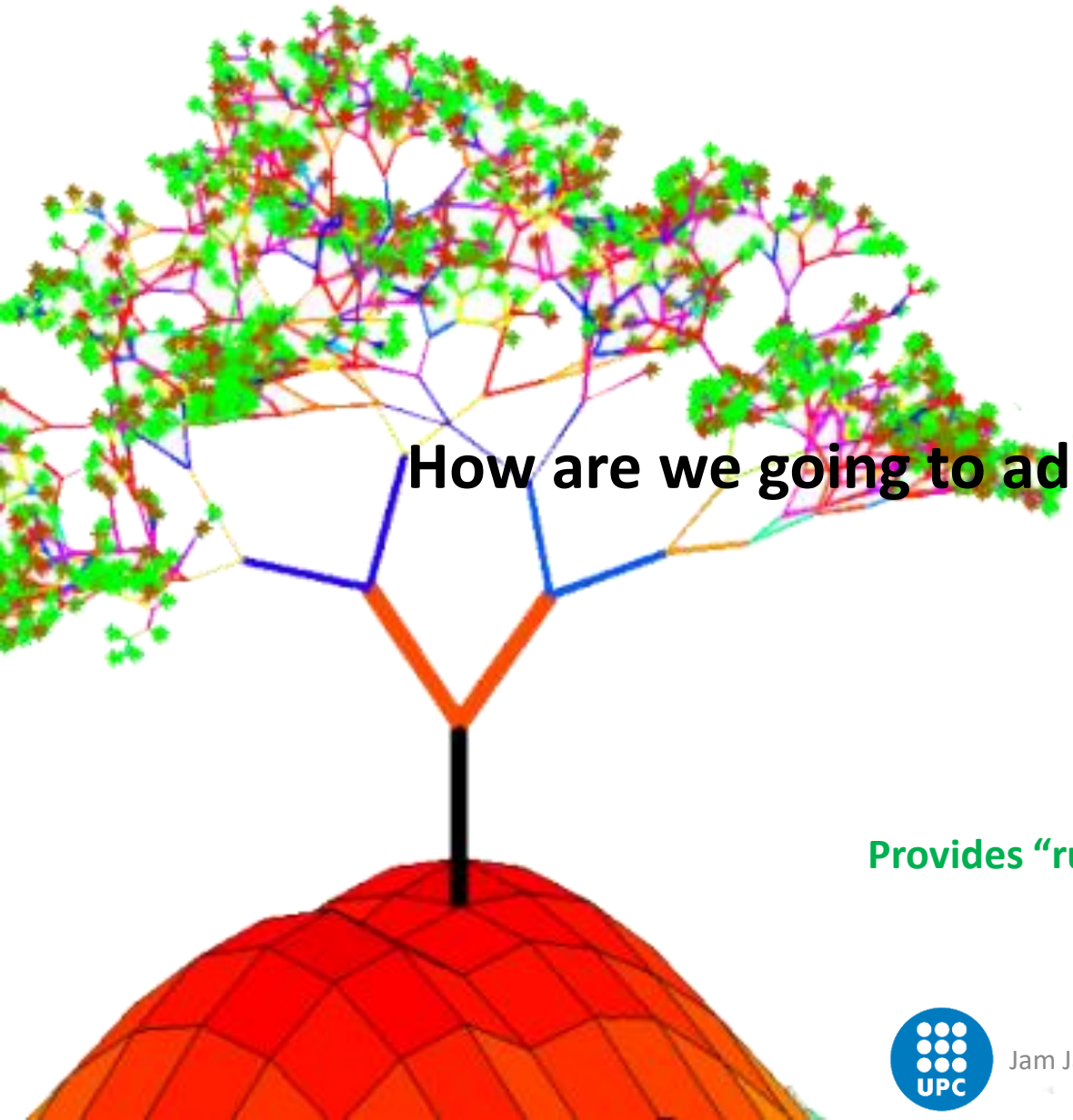


Framewo

ds:carbon_emissions_kgco2e

ds:total_emissions_mt_co2e





How are we going to add multidimensional concepts?

Decision Trees!

Why?

Provides “rules” used for explanatory analysis of the resource labeling

Extracting Feature from each Resource

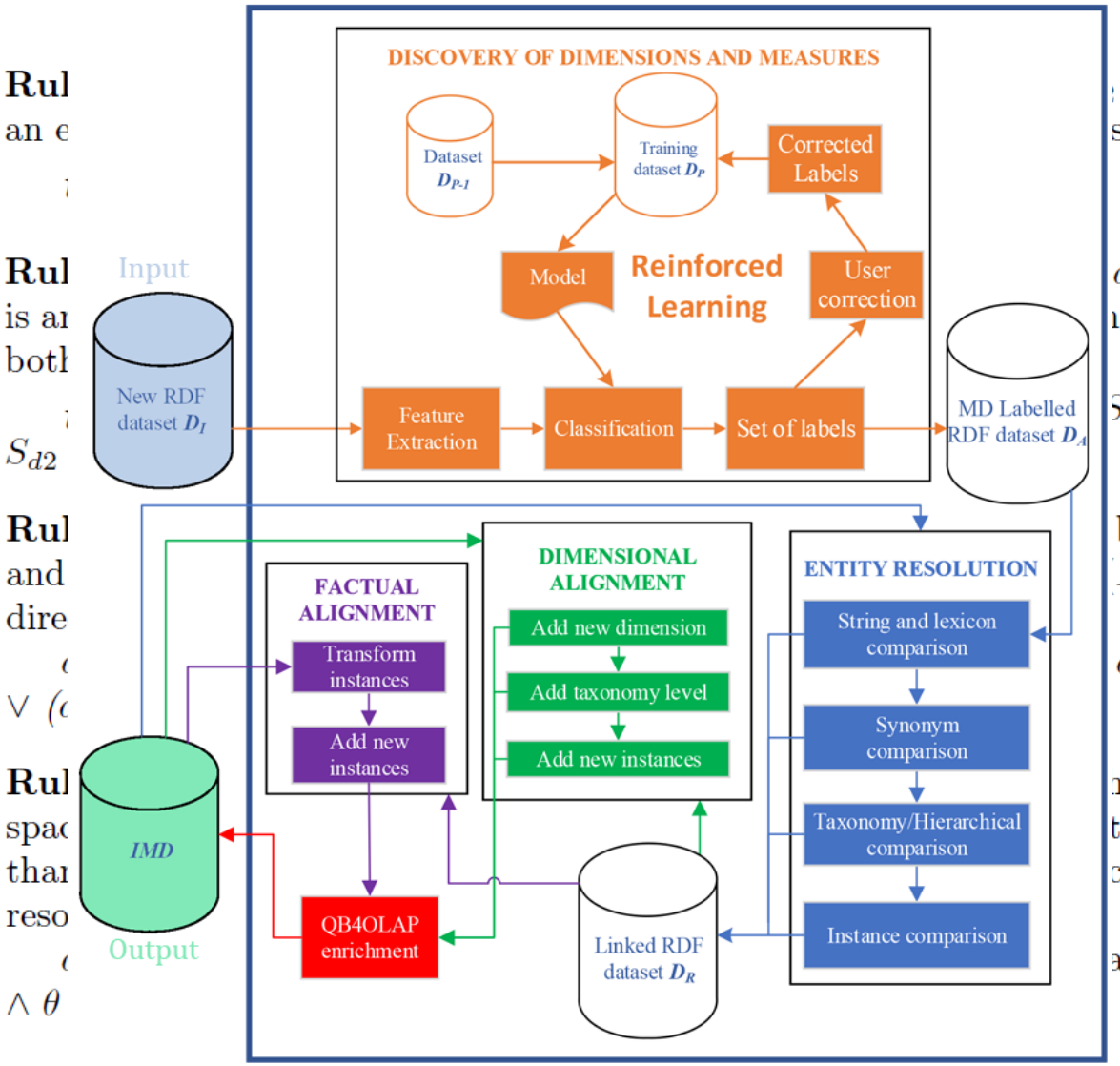
Unique Values	The ratio of unique values based on the total occurrences
Data Types	Such as float, integer, string, Boolean, categorical, date, geolocation, a resource (i.e., a URI) or description (containing metadata information)
URI Prefix and URN	The URI is parsed to obtain these features: <ds:location> 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

Dimension

Measure

Metadata

Entity Resolution



Note: We deal with equivalence and sub-sumption relationships only

Experimentation Results

- We use LogI[illegible]2] for Rules 1 [illegible]d IUT [1] for Ru[illegible]
  - WordNet[1] [illegible]T (that uses So[illegible]ging System) [6]

Total resources:	**16**	**30**	**27**

- By adding MD concepts, we reduce the number of comparisons by 88% and the runtime by 81%
  - Comparison is not a cartesian product
  - When C_{UK} and [illegible] e resources reduce
    - Without MD l[illegible]
    - With MD labe[illegible] tual) resources

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

  - Hence resources of C_{EU} are compared with fewer resources from C_{UK} and C_{US}

1. https://wordnet.princeton.edu/

Experimentation Results

1. Is the DT a good choice?

Method	Experiments repeated	Error rate (average)	Variance
LOO	230	7.39%	6.84%
LOO (random order)	230	7.39%	6.84%
Stratified Sampling (10-fold)	10	7.61%	7.24%
Random Sampling (10-fold)	10	7.83%	9.85%

- Ac

Second	70%	83.3%	C_{US}	30
Third	81.3%	100%	C_{EU}	27

3. Does the order, in which the resources are fed, matter?

2. http://qweb.cs.aau.dk/qboairbase/
3. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g

Next steps

1. Querying the IMD
   - Currently performing physical integration, motive is to move towards virtual integration like F-DW [10]
   - Re-write the SPARQL queries in terms of the schema of the sources (Federated SPARQL query)
2. Create an end-to-end tool for physical and virtual integration

Publications

- Accepted:

  Jam Jahanzeb Khan Behan, Oscar Romero and Esteban Zimányi. "**Multidimensional Integration of RDF Datasets.**" 21st International Conference on Big Data Analytics and Knowledge Discovery – DaWaK 2019

- Planned:

  Journal Paper I: **"RDF2OLAP: A Platform for Multidimensional Integrating of Linked Data"**
  - Authors: Jam Jahanzeb Khan Behan, Oscar Romero and Esteban Zimányi.
  - Target Venue and deadline: Data & Knowledge Engineering (DKE), November 2019

  Demo Paper I: "**RDF2OLAP**"
  - Authors: Jam Jahanzeb Khan Behan, Oscar Romero and Esteban Zimányi.
  - Target Venue and deadline: The 20th International Conference on Web Information Systems Engineering – WISE 2019, November 2019

  Conference Paper II: **"Multidimensional Analysis of Remote RDF Datasets using Federated Queries"**
  - Authors: Jam Jahanzeb Khan Behan, Oscar Romero and Esteban Zimányi.
  - Target Venue and deadline: 17th Extended Semantic Web Conference – ESWC, January 2020

  Demo Paper II: "**LD2OLAP**"
  - Authors: Jam Jahanzeb Khan Behan, Oscar Romero and Esteban Zimányi.
  - Target Venue and deadline: The 17th Extended Semantic Web Conference, March 2020

  Journal Paper II: "Optimization of Federated queries for querying remote datasets"

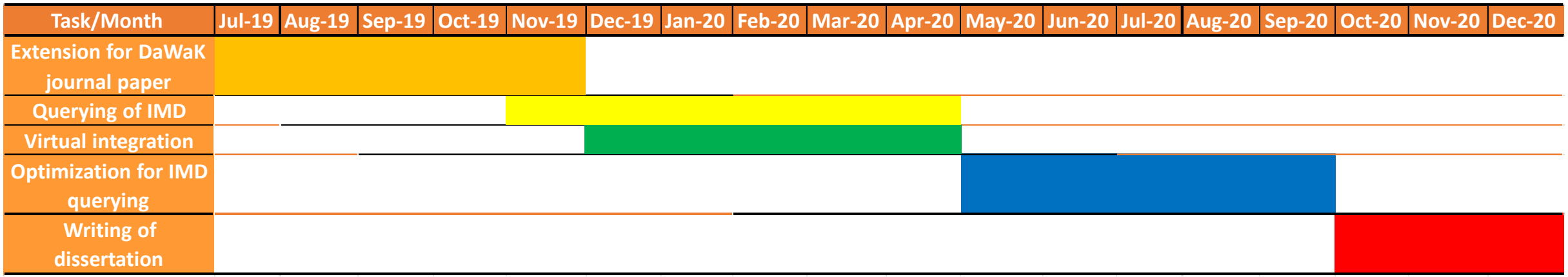
ECTS

Category	Total ECTS
General	17.5
Informal	1
Project	22
Total	40.5

Category	ECTS Earned	Requirement
General	15.5	Complete
Informal	1	Complete
Project	17	Complete
Total	**33.5**	**Complete**

Activity	Place/Organised by	ECTS	Course Category	Status
Academic Writing	ULB	2.5	General	Completed
French Course	ULB (F9 Languages)	2.5	General	Completed
Semantic Web Course	Linköping University	6	Project	Completed
Project Management	ULB	1	General	Completed
Doctoral Seminar "How to publish an article in a scientific journal?"	ULB	1	General	Completed
Doctoral Training "Career path planning? Not without building a network!"	ULB	1	General	Completed
European Business Intelligence and Big Data Summer School	Technische Universiteit Eindhoven	2	Project	Completed
Seminar	ULB	1	Project	Completed
Dutch-Belgian Database Day 2019	Hasselt University	1	Informal	Completed
Scientific Presentation Skills and Career Planning	ULB	5	General	Completed
5th International Winter School on Big Data (BigDat 2019)	University of Cambridge	3	Project	Completed
Seminar	UPC	1	Project	Completed
Spanish Course	UPC	2.5	General	Completed
Semantic Data Management	UPC	2	Project	Completed
Big Data Architecture and Design	UPC	1	Project	Completed
Seminar	UPC	1	Project	Completed
Seminar	UPC	1	Project	Completed
IT4BI Doctoral Colloquium	Berlin, Germany	3	Project	On-going
EDBT Summer School	Lyon, France	2	Project	Planned
Open Access and Bibliometrics	ULB	1	General	Fall 2020
Intellectual Property, Copyright and Knowledge Transfer	ULB	1	General	Fall 2020

Timeline


Task/Month
Jul-19
Aug-19
Sep-19
Oct-19
Nov-19
Dec-19
Jan-20
Feb-20
Mar-20
Apr-20
May-20
Jun-20
Jul-20
Aug-20
Sep-20
Oct-20
Nov-20
Dec-20
Extension for DaWaK journal paper
Querying of IMD
Virtual integration
Optimization for IMD querying
Writing of dissertation

Thank You!

References

1. Zong, N.: Instance-based Hierarchical Schema Alignment in Linked Data. Ph.D. thesis, Seoul National University Graduate School, Seoul, South Korea (2015)
2. Jiménez-Ruiz, E., & Grau, B. C. (2011, October). Logmap: Logic-based and scalable ontology matching. In *International Semantic Web Conference* (pp. 273-288). Springer, Berlin, Heidelberg.
3. Motik, B., Shearer, R., Horrocks, I.: Hypertableau Reasoning for Description Logics. Journal of Artificial Intelligence Research 36, 165–228 (2009)
4. Simancik, F., Kazakov, Y., Horrocks, I.: Consequence-based reasoning beyond Horn ontologies. In: IJCAI (2011)
5. Stoilos, G., Stamou, G.B., Kollias, S.D.: A String Metric for Ontology Alignment. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 624–637. Springer, Heidelberg (2005)
6. Heymann, P., & Garcia-Molina, H. (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems*. Stanford.
7. Schultz, A., Matteini, A., Isele, R., Bizer, C., Becker, C.: LDIF - Linked Data Integration Framework. In: Proceedings of the 2nd International Conference on Consuming Linked Data. vol. 782, pp. 125–130. CEUR-WS.org (Oct 2011)

References

8. Kämpgen, B., O'Riain, S., Harth, A.: Interacting with Statistical Linked Data via OLAP Operations. In: Proceedings of the 9th Extended Semantic Web Conference. pp. 87–101. Springer (May 2012)
9. Moaawad, M.R., Mokhtar, H.M.O., Al Feel, H.T.: On-The-Fly Academic Linked Data Integration. In: Proceedings of the International Conference on Compute and Data Analysis. pp. 114–122. ACM (May 2017)
10. Jindal, R., Acharya, A.: Federated Data Warehouse Architecture. Wipro Technologies White Paper (2004)
11. Diamantini, C., Potena, D., Storti, E.: Multidimensional query reformulation with measure decomposition. Information Systems 78, 23–39 (2018)
12. Estrada-Torres, B., Richetti, P.H.P., del-Río-Ortega, A., Baião, F.A., Resinas, M., Santoro, F.M., Ruiz-Cortés, A.: Measuring Performance in Knowledge-intensive Processes. ACM Transactions on Internet Technology 19(1), 15:1–15:26 (2019)
13. Popova, V., Sharpanskykh, A.: Formal modelling of organisational goals based on performance indicators. Data & Knowledge Engineering 70(4), 335–364 (2011)