# Knowledge Graph Completion using Embeddings


**Mehwish Alam, Ph.D.**

**Telecom Paris, France**


**September 15, 2022**


Karlsruher Institut für Technologie

FIZ Karlsruhe

Leibniz Institute for Information Infrastructure

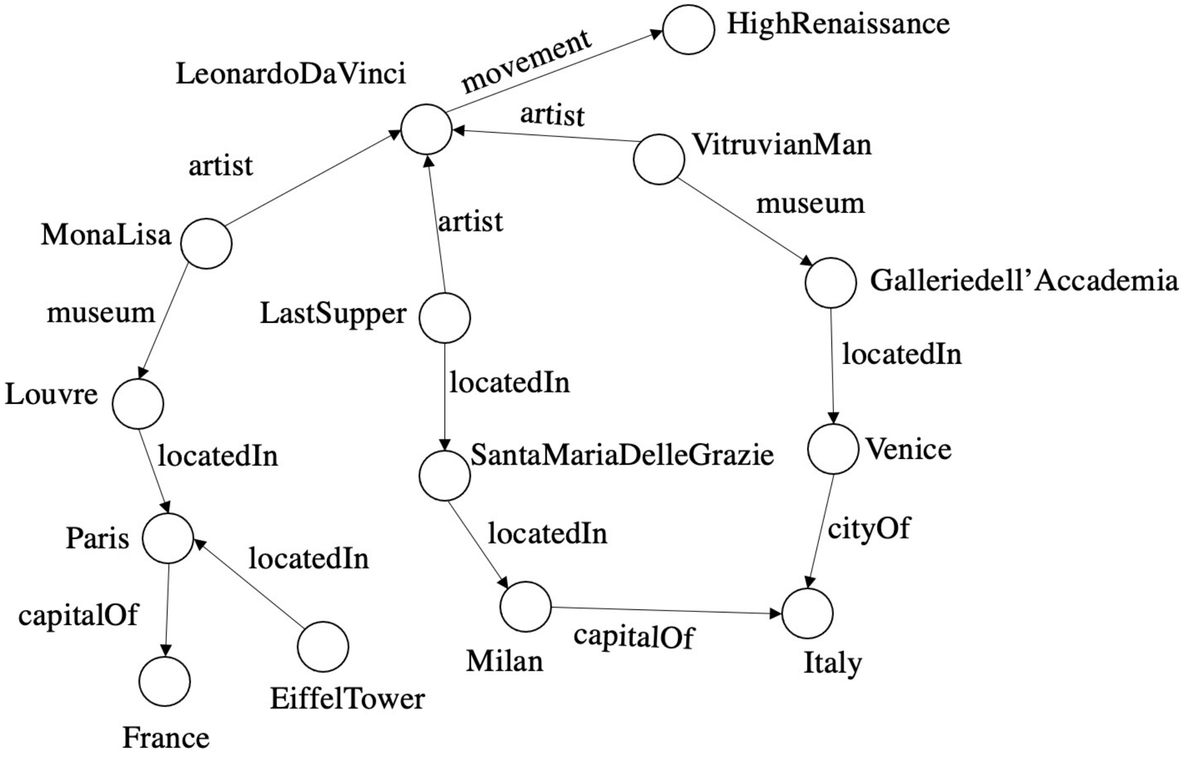
# Knowledge Graphs


Amith Singhal, [Introducing the Knowledge Graph: things, not strings](), Google Blog, May 16, 2012

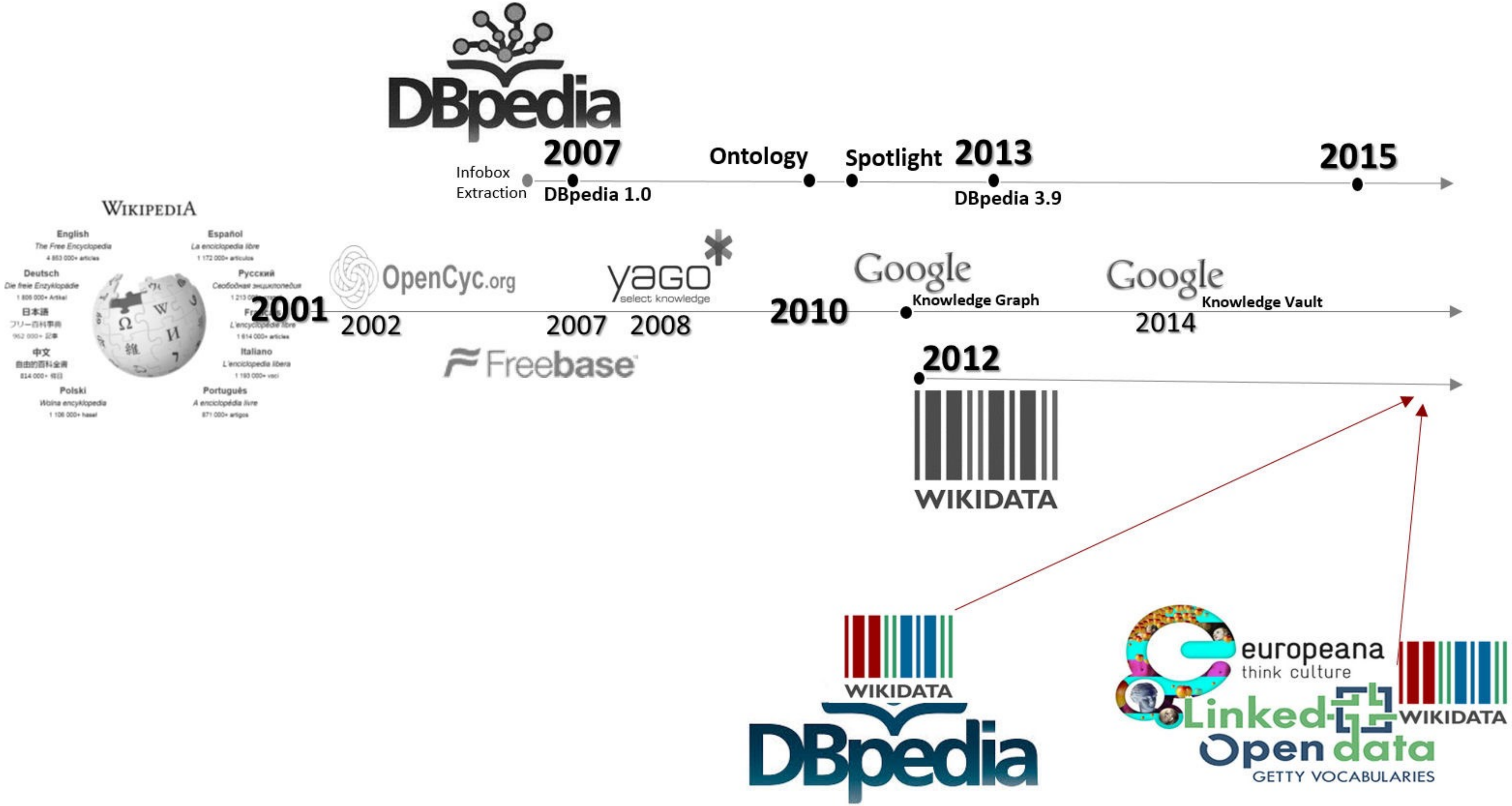
# Knowledge Graphs

**Definition (Knowledge Graph).** Given a KG $G = (E, R)$, where

- $E$ is the set of entities,
- $R$ is the set of relations.
- $< e_h, r, e_t > \in T$, represents a triple belonging to the set of triples T in the KG,
  - $e_h, e_t \in E$ are the head and tail entities,
  - $r \in R$ represents relation between them.

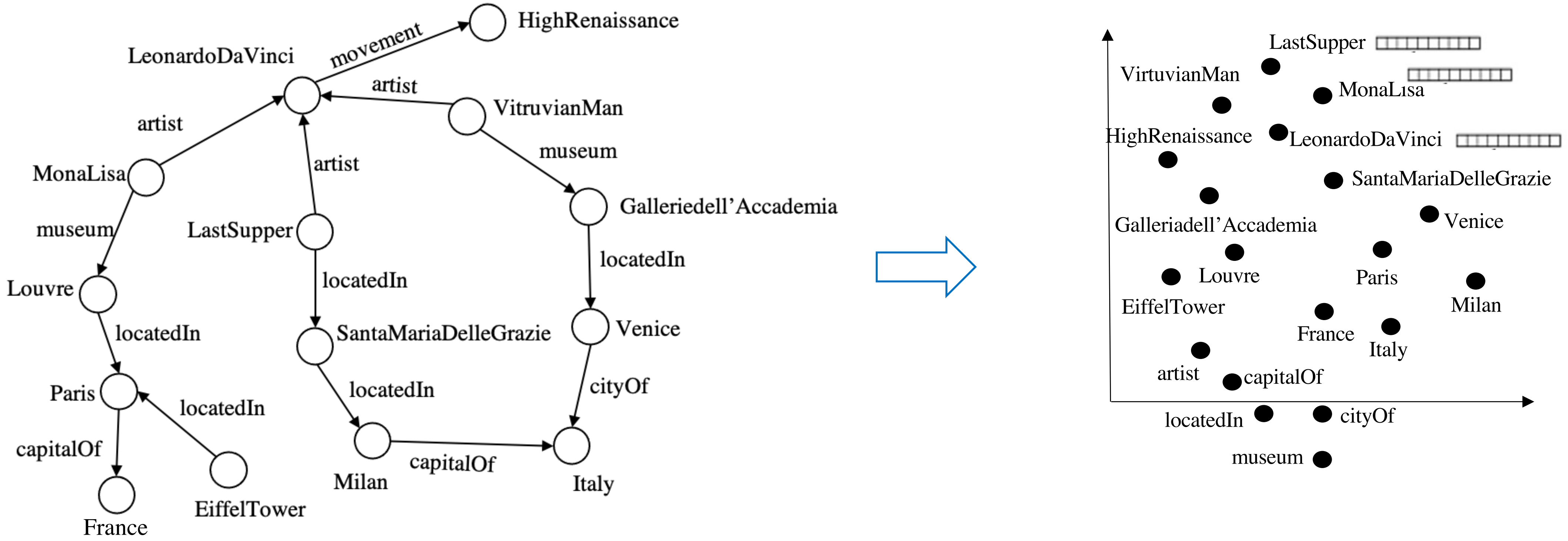


# Timeline of Knowledge Graphs



# Knowledge Graph Embeddings

- Automatic, supervised learning of embeddings, i.e. **projections of entities and relations into a continuous low-dimensional space**




*Q. Wang, Z. Mao, B. Wang, L. Guo. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2017.*

# Knowledge Graph Completion Tasks

**Triple Classification**
LeonardoDaVinci $\xrightarrow{\text{occupation}}$ Artist ?
yes, 95%

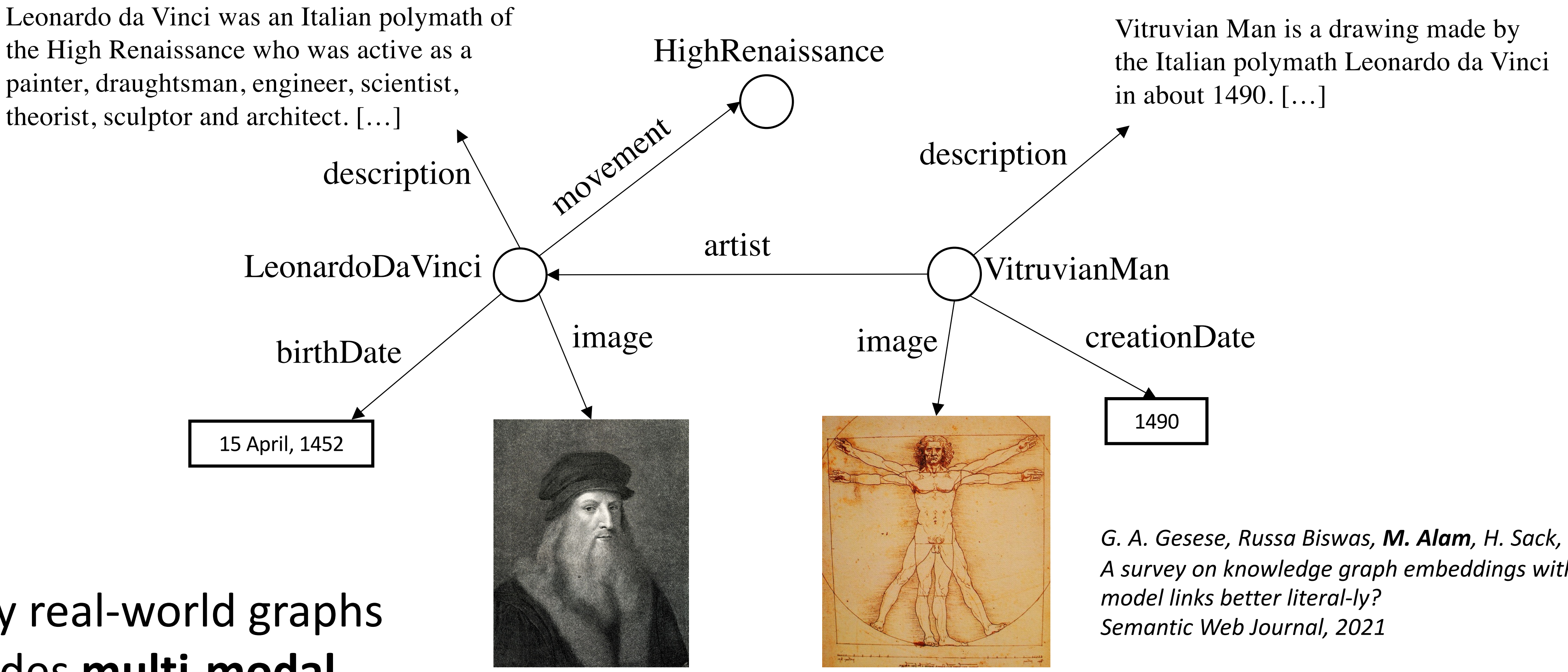
**Head Prediction**
? $\xrightarrow{\text{occupation}}$ Artist
(1, LeonardoDaVinci, 0.91)
(2, Michaelangelo, 0.90)

**Tail Prediction**
LeonardoDaVinci $\xrightarrow{\text{occupation}}$ ?
(1, artist, 0.95),
(2, sculptor, 0.93) ...

**Relation Prediction**
LeonardoDaVinci $\xrightarrow{\text{?}}$ Artist
(1, occupation, 0.95)

**Entity Classification (Type Prediction)**
LeonardoDaVinci $\xrightarrow{\text{type}}$ ?
(1, Person, 0.99)
(2, Human, 0.99),...

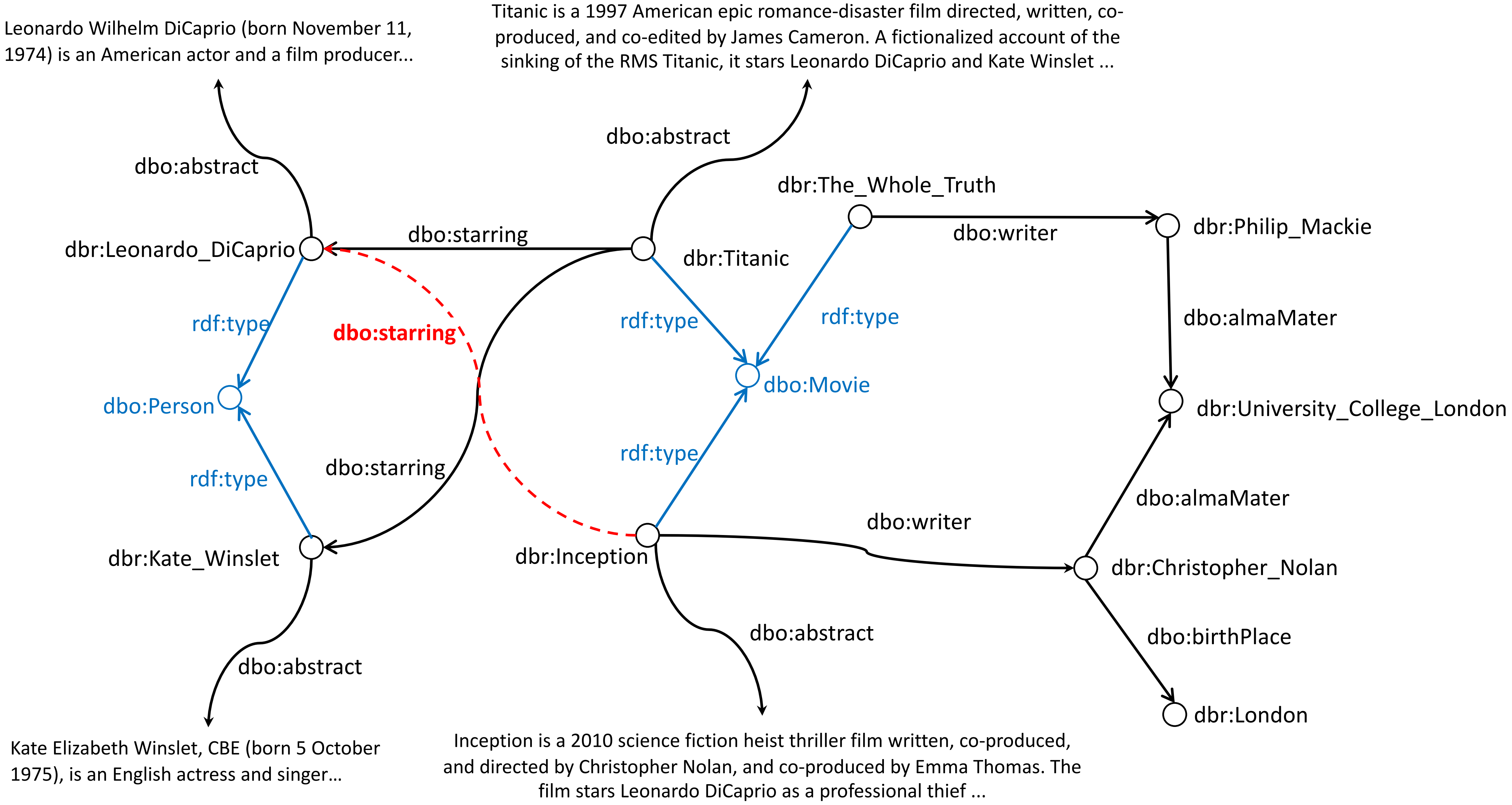
# Multimodal Knowledge Graphs



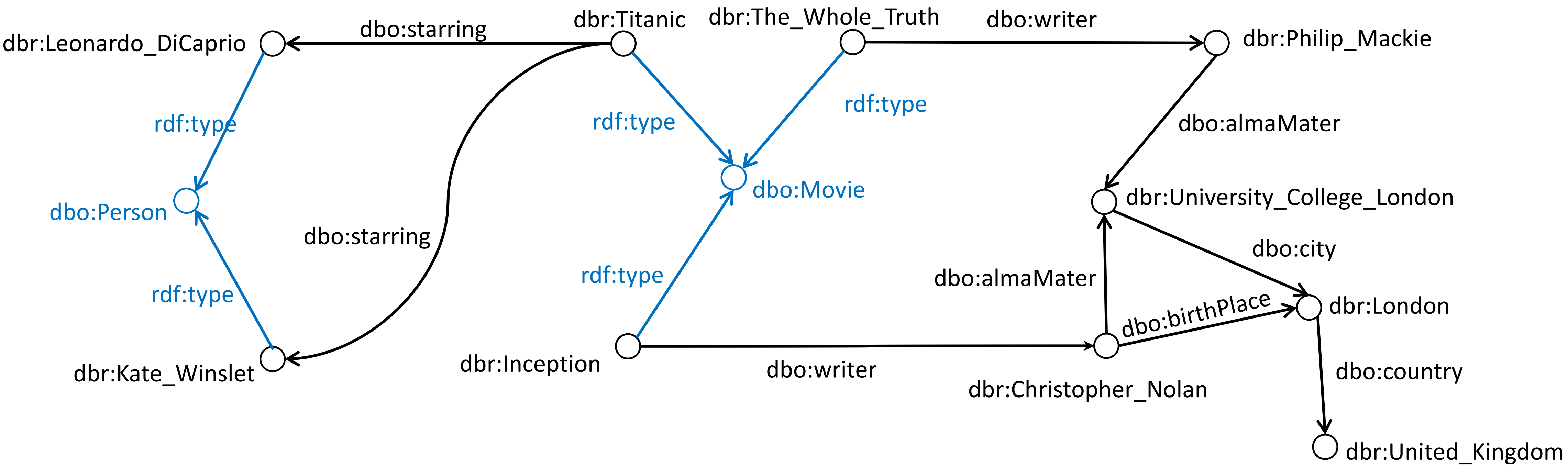

*G. A. Gesese, Russa Biswas,* ***M. Alam****, H. Sack,*
*A survey on knowledge graph embeddings with literals: Which model links better literal-ly?*
*Semantic Web Journal, 2021*

Many real-world graphs includes **multi-modal attributes.**

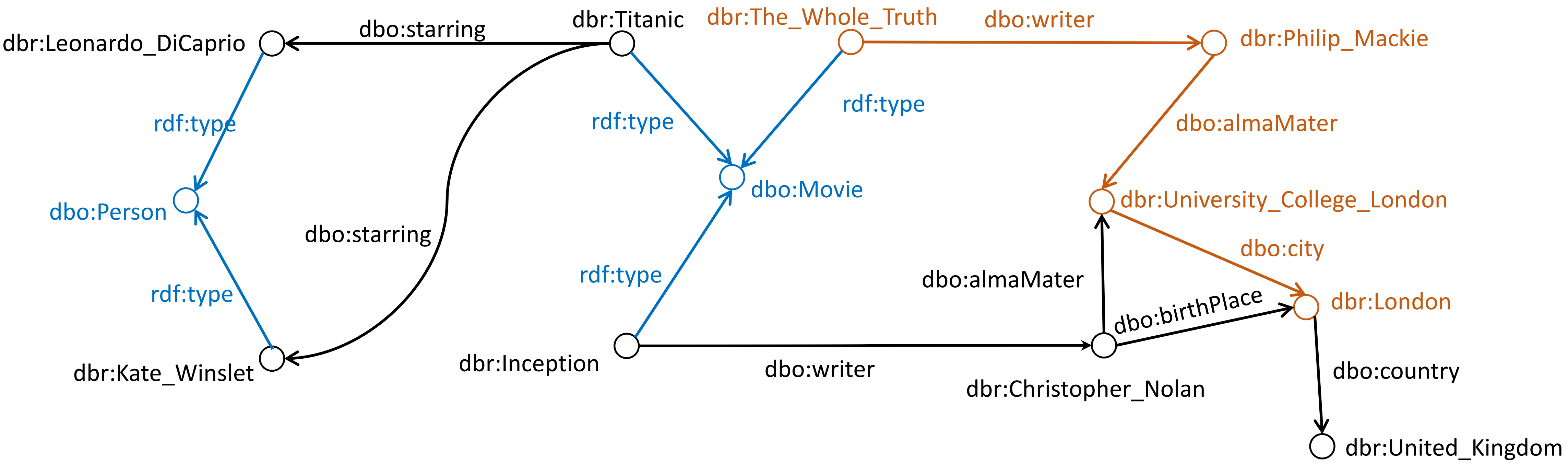
# MADLINK: Attentive Multihop and Entity Descriptions for Link Prediction in Knowledge Graphs


Leonardo Wilhelm DiCaprio (born November 11, 1974) is an American actor and a film producer...
Titanic is a 1997 American epic romance-disaster film directed, written, co-produced, and co-edited by James Cameron. A fictionalized account of the sinking of the RMS Titanic, it stars Leonardo DiCaprio and Kate Winslet ...
dbo:abstract
dbo:abstract
dbr:The_Whole_Truth
dbo:starring
dbr:Leonardo_DiCaprio
dbo:writer
dbr:Philip_Mackie
dbr:Titanic
rdf:type
dbo:starring
rdf:type
rdf:type
dbo:almaMater
dbo:Movie
dbo:Person
dbr:University_College_London
rdf:type
dbo:starring
rdf:type
dbo:almaMater
dbo:writer
dbr:Kate_Winslet
dbr:Inception
dbr:Christopher_Nolan
dbo:abstract
dbo:birthPlace
dbo:abstract
dbr:London
Kate Elizabeth Winslet, CBE (born 5 October 1975), is an English actress and singer...
Inception is a 2010 science fiction heist thriller film written, co-produced, and directed by Christopher Nolan, and co-produced by Emma Thomas. The film stars Leonardo DiCaprio as a professional thief ...

# Extracting Entity Context



# Extracting Entity Context




**Random Walks:**

(Depth = 3): dbr:The_Whole_Truth → dbo:writer → dbr:Philip_Mackie → dbo:almaMater → dbr:University_College_London → dbo:city → dbr:London

# Predicate Frequency – Inverse Triple Fequency

- Predicates are selected at each hop using PF-ITF
- Predicate Frequency $(pf)$

$$pf_o^e(r, G) = \frac{|\varepsilon_o(e)|_{\pi(r)}}{|\varepsilon_o(e)|}$$

$|\varepsilon_o(e)|_{\pi(r)}$: # of outgoing edges from *e* w.r.t. relation *r*

$\pi(r)$: set of relations

$|\varepsilon_o(e)|$: total # of outgoing edges from *e*

- Inverse Triple Frequency $(itf)$

$$itf(r, G) = log \frac{|\varepsilon|}{|\varepsilon_o|_{\pi(r)}}$$

$|\varepsilon|$: total # of triples

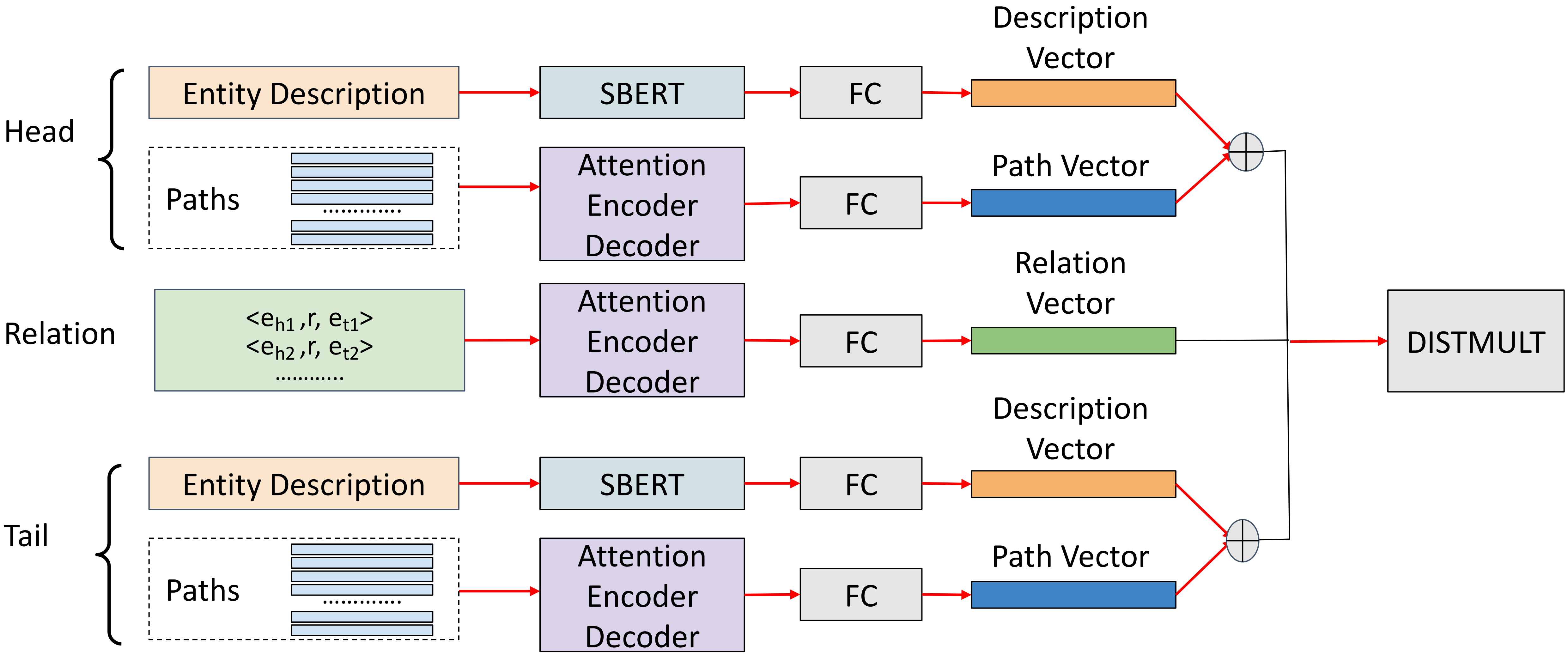
$|\varepsilon_o|_{\pi(r)}$: total # of triples w.r.t. relation *r*

- $pf - itf$

$$pfitf_e(r, G) = pf_e \times itf$$

- Triples are then ranked based on pf-itf, ***top-n*** predicates are selected at each hop.

# MADLINK – Overall Architecture




R. Biswas, H. Sack, **M. Alam**, MADLINK: Attentive Multihop and Entity Descriptions for Link Prediction in Knowledge Graphs, *Semantic Web Journal, 2022* (Accepted)

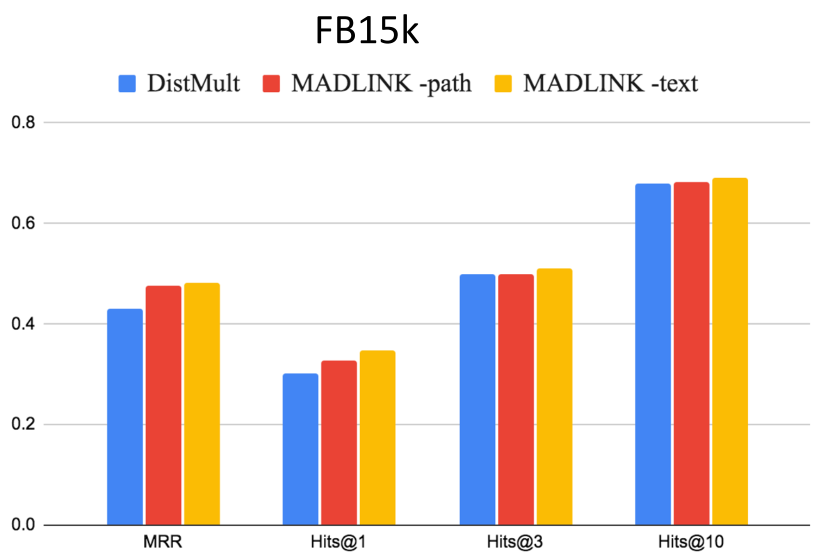
# Comparison with other Text-based Models

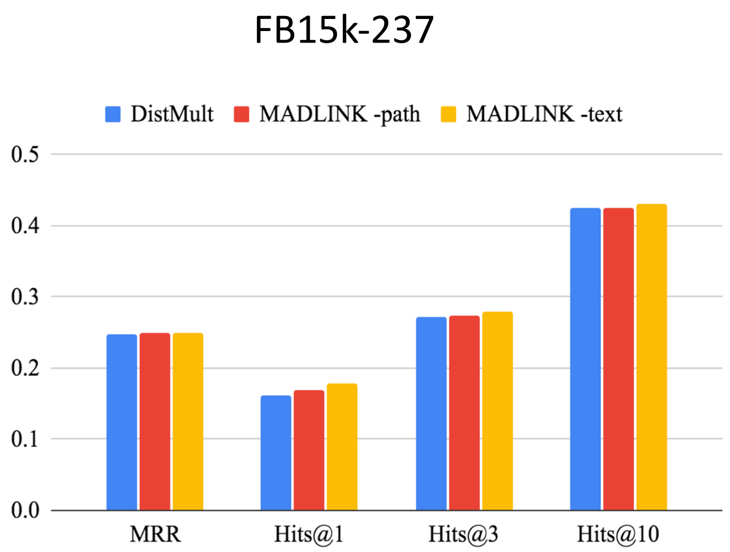
| Datasets | Models | MRR | Hits@1 | Hits@3 | Hits@10 |
|---|---|---|---|---|---|
| FB15k-237 | DKRL | 0.19 | 0.11 | 0.167 | 0.215 |
| | Jointly (ALSTM) | 0.21 | 0.19 | 0.21 | 0.258 |
| | MADLINK | 0.347 | 0.252 | 0.38 | 0.529 |
| WN18RR | DKRL | 0.112 | 0.05 | 0.146 | 0.288 |
| | Jointly (ALSTM) | 0.21 | 0.112 | 0.156 | 0.31 |
| | MADLINK | 0.477 | 0.438 | 0.479 | 0.549 |
| FB15k | DKRL | 0.311 | 0.192 | 0.359 | 0.548 |
| | Jointly (ALSTM) | 0.345 | 0.21 | 0.412 | 0.65 |
| | MADLINK | 0.712 | 0.722 | 0.788 | 0.81 |
| WN18 | DKRL | 0.51 | 0.31 | 0.542 | 0.61 |
| | Jointly (ALSTM) | 0.588 | 0.388 | 0.596 | 0.77 |
| | MADLINK | 0.95 | 0.898 | 0.911 | 0.96 |
| YAGO3-10 | DKRL | 0.19 | 0.119 | 0.234 | 0.321 |
| | Jointly (ALSTM) | 0.22 | 0.296 | 0.331 | 0.41 |
| | MADLINK | 0.538 | 0.457 | 0.580 | 0.68 |

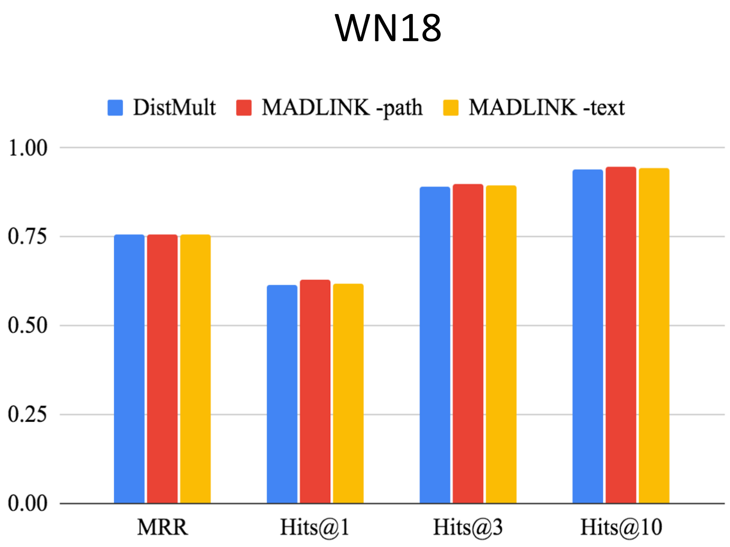
Representation Learning of Knowledge Graphs with Entity Descriptions R. Xie, Z. Liu, J. Jia, H. Luan, M. Sun, AAAI 2016

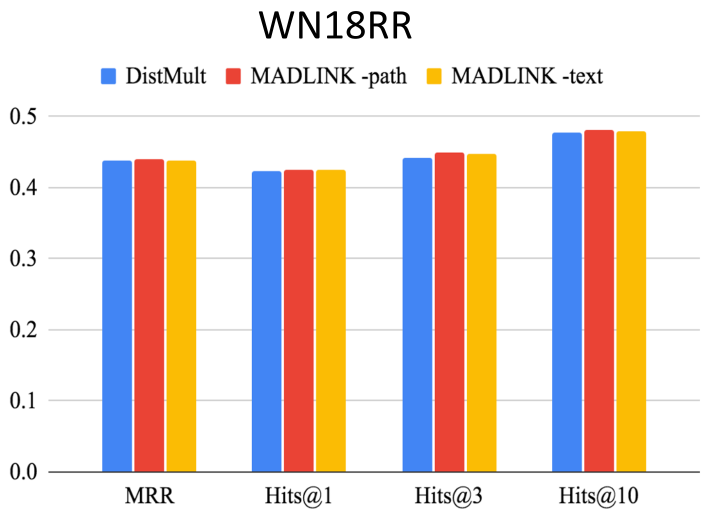
Knowledge Graph Representation with Jointly Structural and Textual Encoding J. Xu, K. Chen, X. Qiu, X. Huang, IJCAI 2017

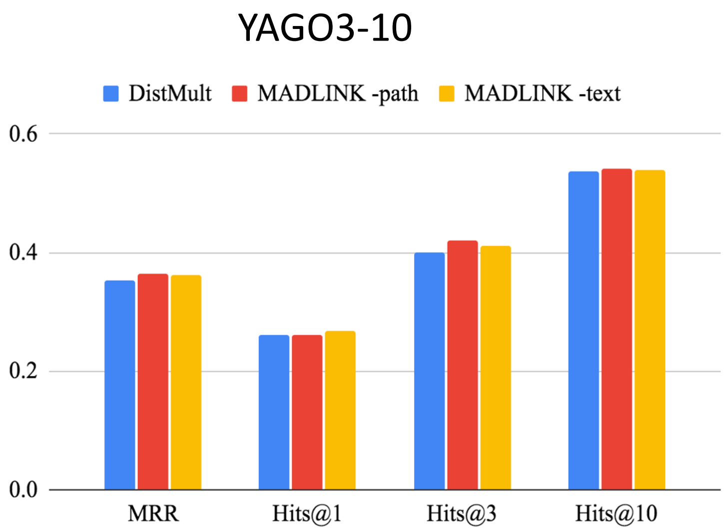
# Impact of Textual Descriptions and Paths



# Existing Datasets for KG Completion

- Freebase Extracts:
  - FB15K and FB15K-237 are widely used.
  - FB15K-237 was designed by removing inverse-duplicates.
  - FB15K-237 contains many triples with **skewed relations towards either some head or tail entities**.
- WordNet Extracts:
  - WN18 and WNRR are most widely used.
  - They do not contain numeric literals.
- Wikidata & Wikipedia Extracts:
  - Not include any numeric attribute
  - If extracted from Wikidata, only limited number of entities contain numeric attributes

# LiterallyWikidata - A Benchmark for KG Completion using Literals

## Overall Workflow for Generating Literally Wikidata

WIKIDATA

Full dump released on September 07, 2020 (13,958,526,390 triples)

**Extracting triples with datatype properties**

- Identifying **properties taking numeric values**
- Extracting triples with these properties having **non-class entities as subjects**

entities

**Extracting triples with object properties**

- Retrieving triples
- Removing **inverse-duplicate** and **symmetric- duplicate** properties

triples

**Filtering Triples**

- Seeding the entities
- Extracting triples based on seed entities
- Creating k-core sub-graphs
- Further filtering

dataset

**Splitting the datasets**

Splitting the datasets into train, valid, and test sets

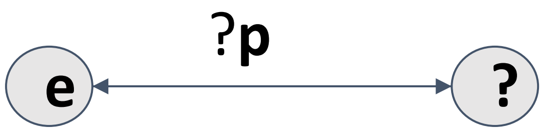
# LiterallyWikidata Pipeline

**3** **Filtering the triples**

- Seeding the entities
- Extracting triples based on seed entities
- Creating k-core sub-graphs
- Further filtering

## Creating dataset 1:

- Seeding entities:
  - Top **N** entities with the highest number of datatype properties. ( N = 200,000 )
- Extracting triples
  - Getting triples where the seed entities occur either as a head or as a tail, i.e., extending the entities with their **1-hop** neighbours.




Where **e** is a seed entity and **p** is an object property.

# LiterallyWikidata Pipeline

3 **Filtering the triples**

- Seeding the entities
- Extracting triples based on seed entities
- Creating k-core sub-graphs
- Further filtering

## Creating dataset 1:

- Seeding entities
- Extracting triples
- Generating k-core
    - A k-core is a maximal subgraphs G’ of a given graph G such that every node in G’ has a degree of at least k.
    - K=6

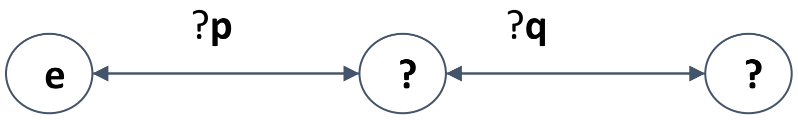
# LiterallyWikidata Pipeline

**3** **Filtering the triples**

- Seeding the entities
- Extracting triples based on seed entities
- Creating k-core sub-graphs
- Further filtering

## Creating dataset 2:

- Seeding entities:
  - N = 50,000
  - K = 15
- Extracting triples
  - Getting triples considering the seed entities both **one-hop** and **2-hop** neighbours.




Where **e** is a seed entity and **p and q are** object properties.

# LiterallyWikidata Pipeline

**3** **Filtering the triples**

- Seeding the entities
- Extracting triples based on seed entities
- Creating k-core sub-graphs
- Further filtering

## Creating dataset 3:

- Seeding entities: N = 200,000
- Extracting triples: 1-hop
- Generating k-core: k=15

# LiterallyWikidata Statistics

| | LitWD1K | LitWD19K | LitWD48K |
|---|---|---|---|
| **#Entities** | 1533 | 18986 | 47998 |
| **#Relations** | 47 | 182 | 257 |
| **#Attributes** | 81 | 151 | 291 |
| **#StruTriples** | 29017 | 288933 | 336745 |
| **#AttrTriples** | 10988 | 63951 | 324418 |
| **#Train** | 26115 | 260039 | 303117 |
| **#Test** | 1451 | 14447 | 16838 |
| **#Valid** | 1451 | 14447 | 16838 |
| **Connectivity** | Yes | Yes | No |
| **Diameter** | 5 | 7 | 8 |
| **Density** | 0.01235 | 0.0008 | 0.00014 |

# Textual Information

- Textual information includes
  - Wikidata labels, aliases, descriptions of entities, relations, and attributes.
- Multi-linguality
  - For each entity summary part extracted from Wikipedia
  - Languages targeted: English, German, Russian, and Chinese.

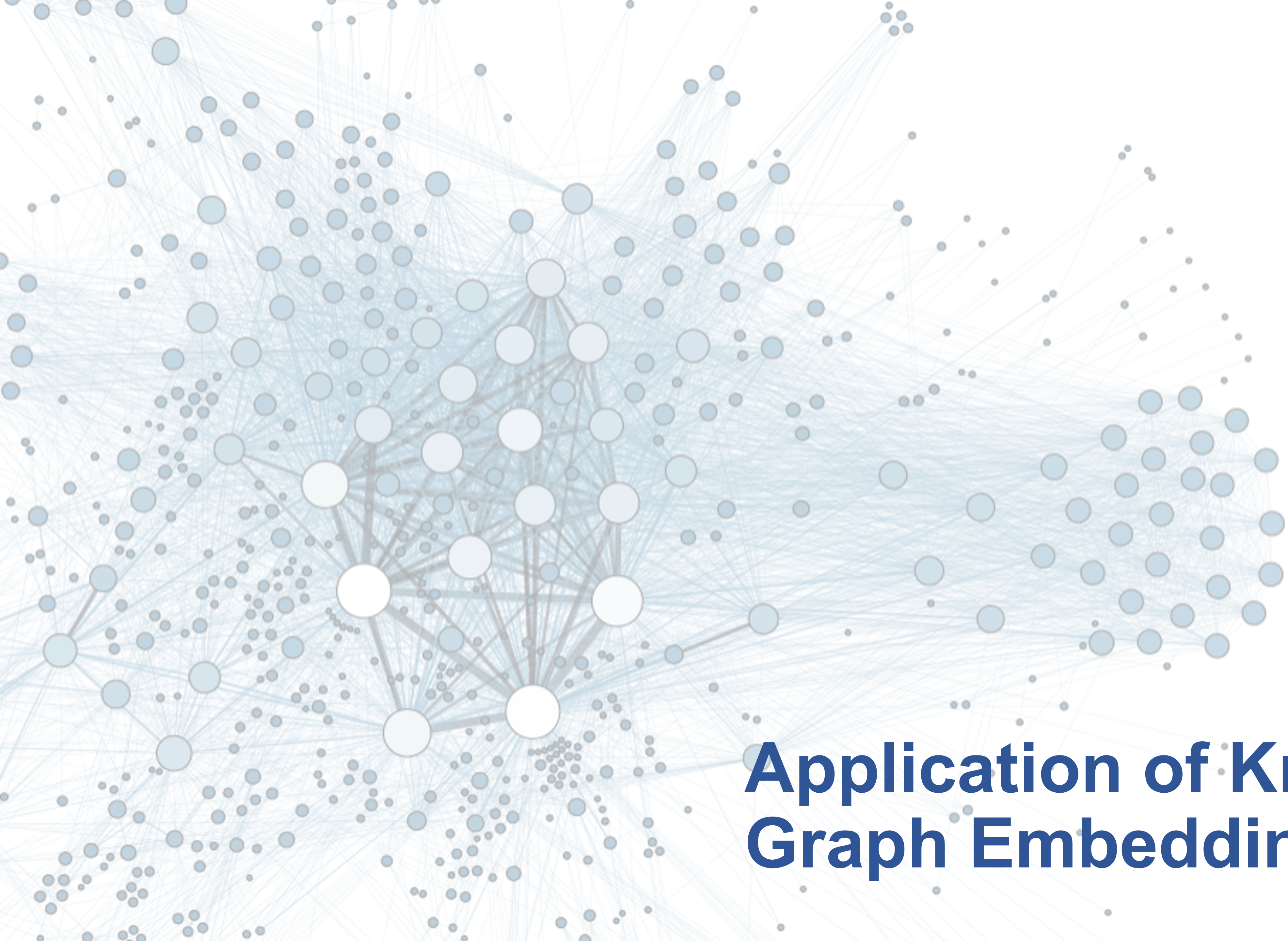
| | **Wikipedia Summary** | | | |
|---|---|---|---|---|
| | **en** | **de** | **ru** | **zh** |
| **LitWD1K** | 100 | 78 | 72 | 66 |
| **LitWD19K** | 100 | 80 | 65 | 39 |
| **LitWD48K** | 100 | 88 | 75 | 29 |

# Experimentation on Link Prediction

| Dataset | Model | MRR | HITS@1 | HIT@10 |
|---|---|---|---|---|
| LitWD1K | DistMult | 0.419 | 0.283 | 0.679 |
| | ComplEX | 0.413 | 0.28 | 0.673 |
| | DistMultLiteral | **0.431** | **0.297** | **0.703** |
| LitWD19K | DistMult | 0.195 | 0.138 | 0.308 |
| | ComplEX | 0.181 | 0.122 | 0.296 |
| | DistMultLiteral | **0.245** | **0.168** | **0.399** |
| LitWD48K | DistMult | 0.261 | 0.195 | 0.4 |
| | ComplEX | 0.277 | **0.207** | 0.428 |
| | DistMultLiteral | **0.279** | 0.204 | **0.434** |

| | | | | |
|---|---|---|---|---|
| FB15K-237 | DistMult | 0.343 | 0.250 | 0.531 |
| | ComplEX | 0.348 | 0.253 | 0.536 |
| CoDEx-M | ComplEX | 0.337 | 0.262 | 0.476 |

*G. A. Gesese, **M. Alam**, H. Sack, LiterallyWikidata - A Benchmark for Knowledge Graph Completion using Literals, International Semantic Web Conference, 2021.*



# Application of Knowledge Graph Embeddings

# Author Name Disambiguation

Cutaneous squamous-cell carcinoma

M Alam, D Ratner - New England Journal of Medicine, 2001 - Mass Medical Soc

Nonmelanoma skin cancer is the most common cancer in the United States, with over 1.3 million cases expected to occur in the year 2001. Approximately 80 percent of nonmelanoma …

☆ Save 99 Cite Cited by 1485 Related articles All 9 versions

**Are these the same people?**

Framester: A wide coverage linguistic linked data hub

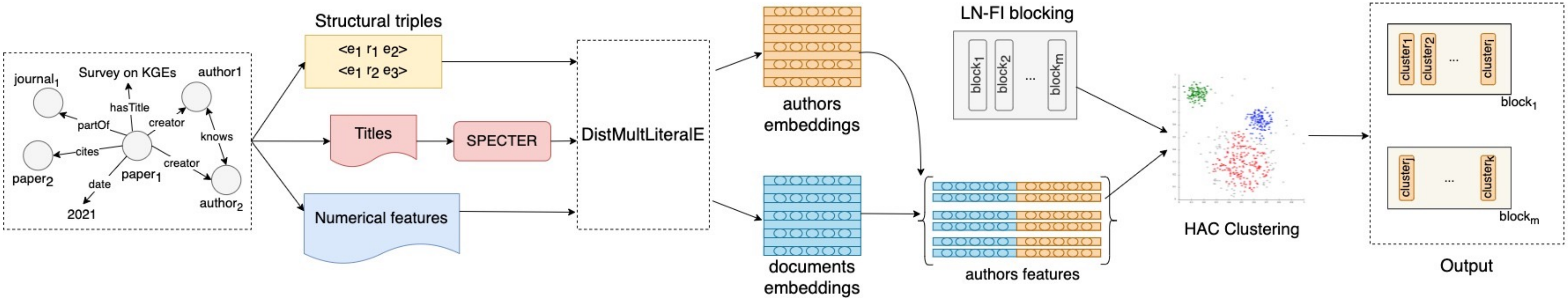
A Gangemi, M Alam, L Asprino, V Presutti… - European knowledge …, 2016 - Springer

Semantic web applications leveraging NLP can benefit from easy access to expressive lexical resources such as FrameNet. However, the usefulness of FrameNet is affected by its …

☆ Save 99 Cite Cited by 76 Related articles All 8 versions

# Overall Architecture of LAND

## Literally Author Name Disambiguation (LAND)




*C. Santini, G. A. Gesese, S. Peroni, A. Gangemi, H. Sack, **M. Alam**,*
A Knowledge Graph Embeddings based Approach for Author Name Disambiguation using Literals *(Accepted), Scientometrics Journal, 2022.*

# Entity Typing in Knowledge Graphs using Wikipedia Categories

# Wikipedia Categories

## Category:Machine learning

❓ Help

From Wikipedia, the free encyclopedia

**Machine learning** is a branch of statistics and computer science, which studies algorithms and architectures that learn from observed facts.

*The main article for this category is **Machine learning**.*

*See also the categories data mining, artificial intelligence, decision theory, and Statistical classification*

Wikimedia Commons has media related to ***Machine learning***.

**Contents**



## Subcategories

This category has the following 35 subcategories, out of 35 total.

**A**

▶ Applied machine learning (1 C, 65 P)
▶ Artificial neural networks (2 C, 170 P)

**B**

▶ Bayesian networks (13 P)
▶ Blockmodeling (14 P)

**C**

▶ Classification algorithms (3 C, 86 P)
▶ Cluster analysis (2 C, 20 P)
▶ Computational learning theory (21 P)
▶ Artificial intelligence conferences (22 P)
▶ Signal processing conferences (5 P)

**D**

**E**

▶ Ensemble learning (13 P)
▶ Evolutionary algorithms (4 C, 44 P, 2 F)

**G**

▶ Genetic programming (13 P)

**I**

▶ Inductive logic programming (5 P)

**K**

▶ Kernel methods for machine learning (1 C, 17 P)

**L**

▶ Latent variable models (2 C, 26 P)
▶ Learning in computer vision (5 P)
▶ Log-linear models (2 P)

▶ Machine learning algorithms (1 C, 69 P)
▶ Machine learning task (9 P)
▶ Markov models (2 C, 54 P)

**O**

▶ Ontology learning (computer science) (2 P)

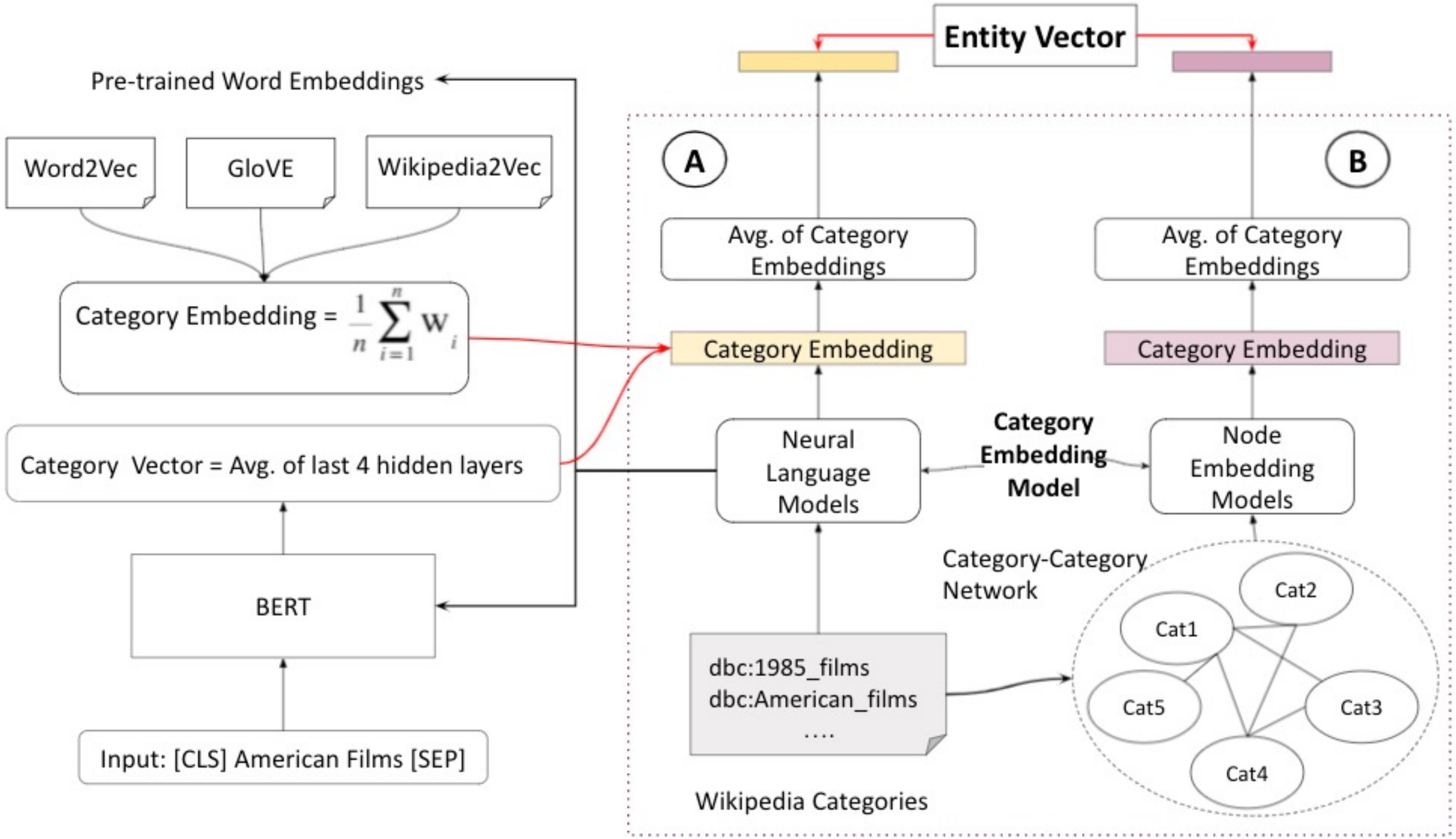
**R**

▶ Reinforcement learning (7 P)
▶ Machine learning researchers (132 P)
▶ Natural language processing researchers (114 P)

**S**

▶ Semisupervised learning (2 P)
▶ Statistical natural language processing (1 C, 35 P)
▶ Structured prediction (1 C, 4 P)
▶ Supervised learning (4 P)

# Overall Architecture



# Entity Classification

- Multi-class Classification
  - Fully Connected Neural Networks (FCNN) with **two dense layers**
  - **ReLU** as activation function
  - A **softmax classifier** with **cross-entropy loss function** is used in the last layer to calculate the probability of the entities belonging to different classes.

$$f(s_i) = \frac{e^{s_i}}{\sum_j^{C_T} e^{s_j}}$$

Scores inferred for each class $C_T$

$$CE_{loss} = -\sum_i^{C_T} t_i \log f(s_i)$$

Score for $i$-th class

Ground truth for $i$-th class

- Multi-label classification
  - a **sigmoid function** with **binary cross-entropy loss** is used in the last layer which sets up a binary classification problem for each class in $C_T$ .

$$CE_{loss} = -\sum_i^{C_T} t_i \log\big(f(s_i)\big) - (1 - t_i)\log(1 - f(s_i))$$

# Summary

- Knowledge Graph (KG) Embedding Algorithm using textual description
  - Entity Context using random walks
  - Language Models
- Benchmark dataset for numeric information and textual description
- Application of KG Embedding algorithms with textual and numeric information on **Author Name Disambiguation**
- Entity Type prediction based on Wikipedia Category Embeddings
- Other Related Studies:
  - Entity Type Prediction using different walk strategies

# Thank you very much for your attention!

✉ mehwish.alam@kit.edu