Leveraging Linked Data to Discover Semantic Relations within Data Sources

Mohsen Taheriyan

Craig A. Knoblock

Pedro Szekely

Jose Luis Ambite

USC Viterbi School of Engineering

Information Sciences Institute

Map Structured Data to Ontologies

Map the source to the classes & properties in an ontology

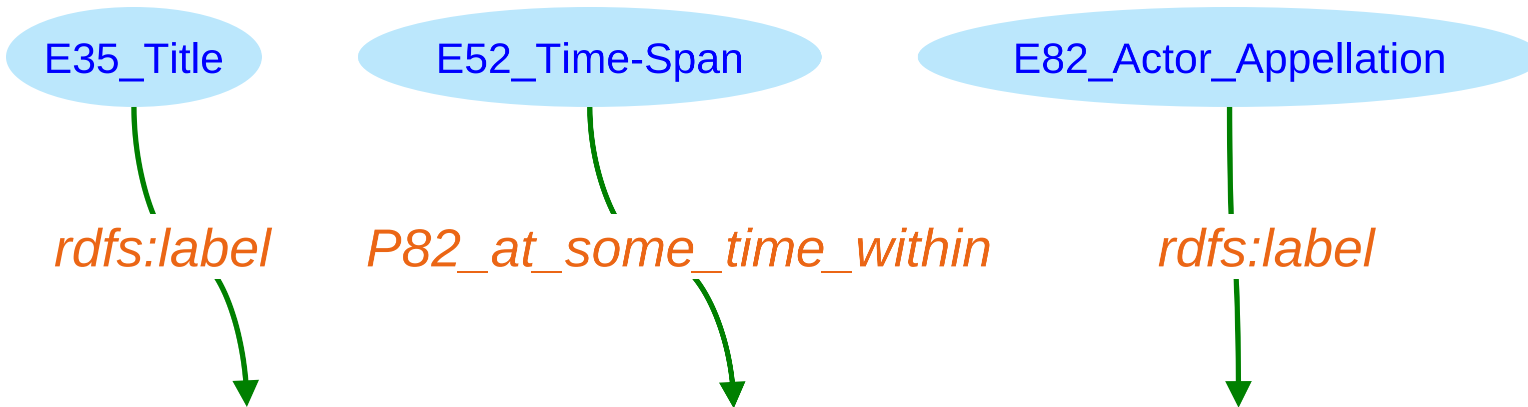
Source

	title	date	name
1	The Island	2009	Walton Ford
2	Excavation at Night	1908	George Wesley Bellows
3	Rose Garden	1901	Maria Oakey Dewing

Domain Ontology

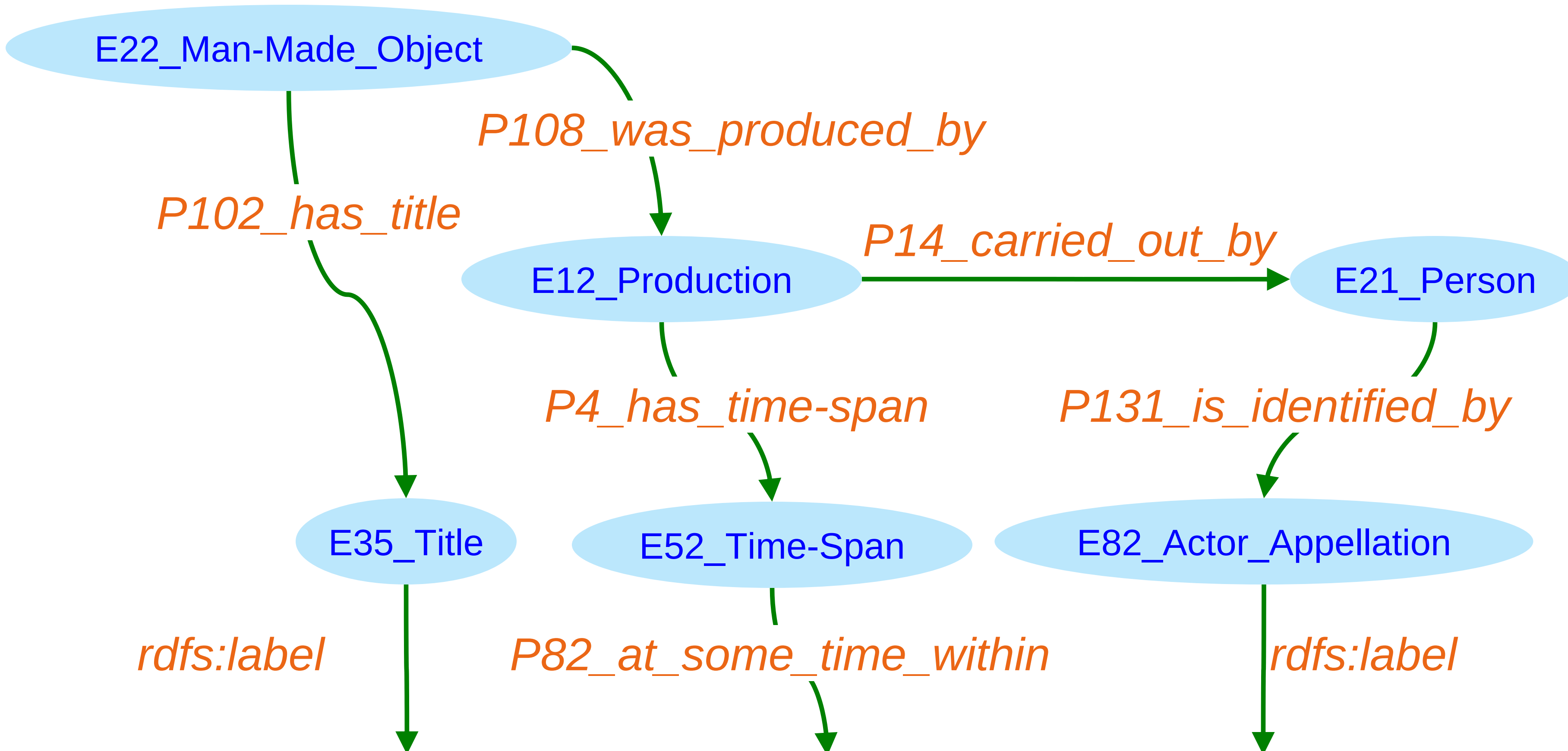
CIDOC-CRM

Semantic Types



	title	date	name
1	The Island	2009	Walton Ford
2	Excavation at Night	1908	George Wesley Bellows
3	Rose Garden	1901	Maria Oakey Dewing

Relationships



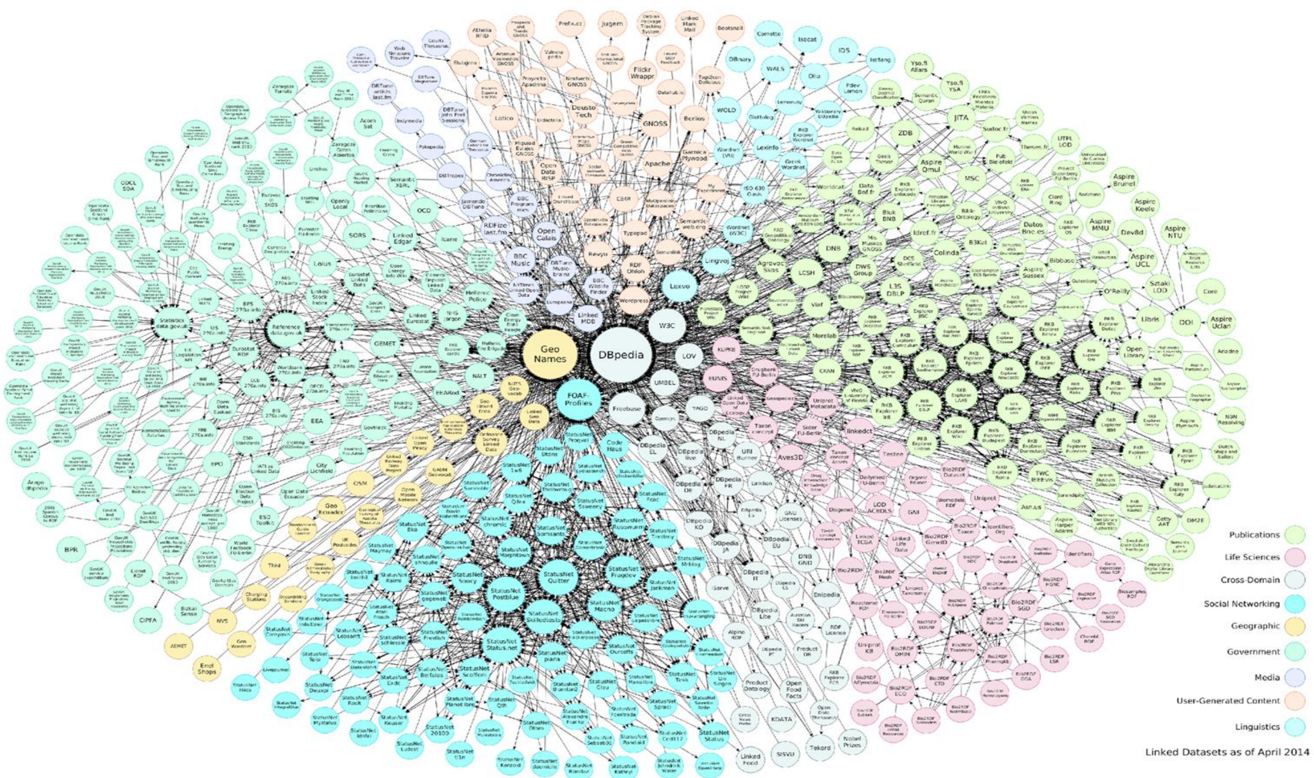
	title	date	name
1	The Island	2009	Walton Ford
2	Excavation at Night	1908	George Wesley Bellows
3	Rose Garden	1901	Maria Oakey Dewing

Problem:

How to automatically infer semantic relations?

Idea

Exploit the relationships within already published linked data



Approach

Input

- Target source (S)
- Domain Ontologies (O)
- Semantic labels of S
- Linked Data (in the same domain)

Output

A ranked set of semantic models for S

1. Extract schema-level graph patterns from LD
2. Construct a graph from LD patterns and the ontology
3. Generate and rank semantic models

Approach

Input

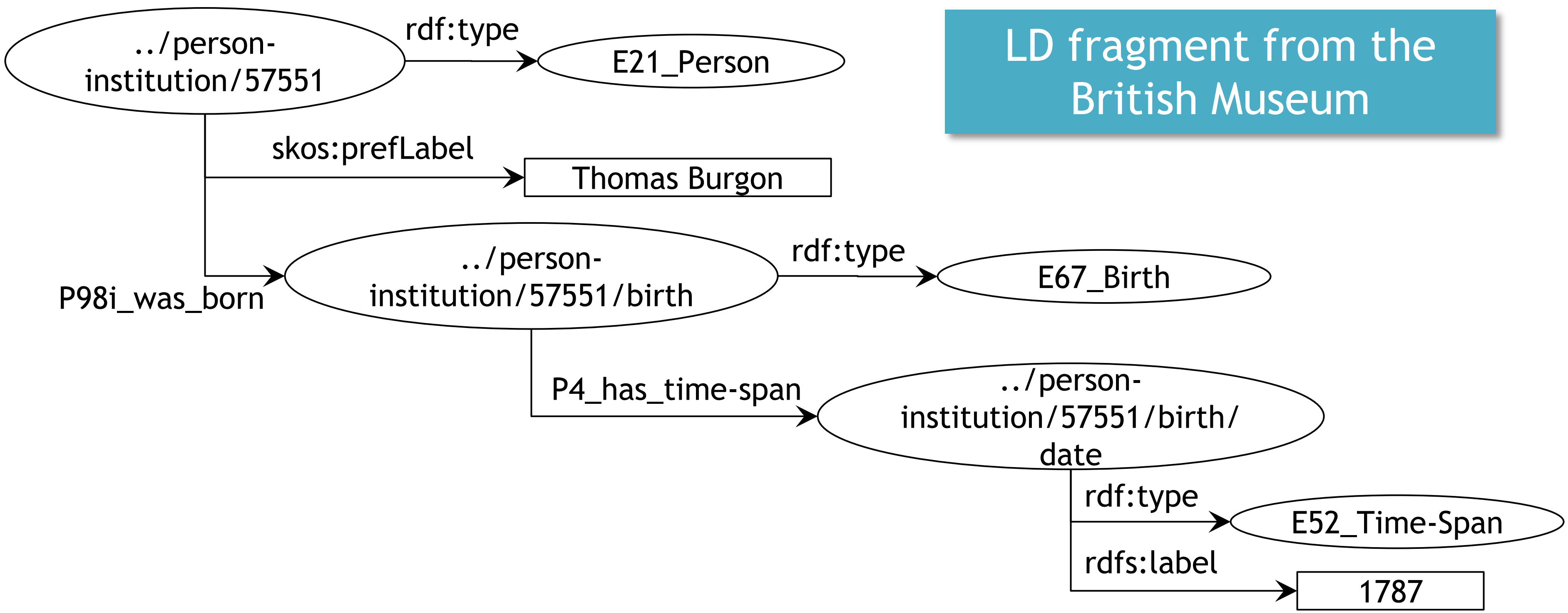
- Target source (S)
- Domain Ontologies (O)
- Semantic labels of S
- Linked Data (in the same domain)

Output

A ranked set of semantic models for S

1. Extract schema-level graph patterns from LD
2. Construct a graph from LD patterns and the ontology
3. Generate and rank semantic models

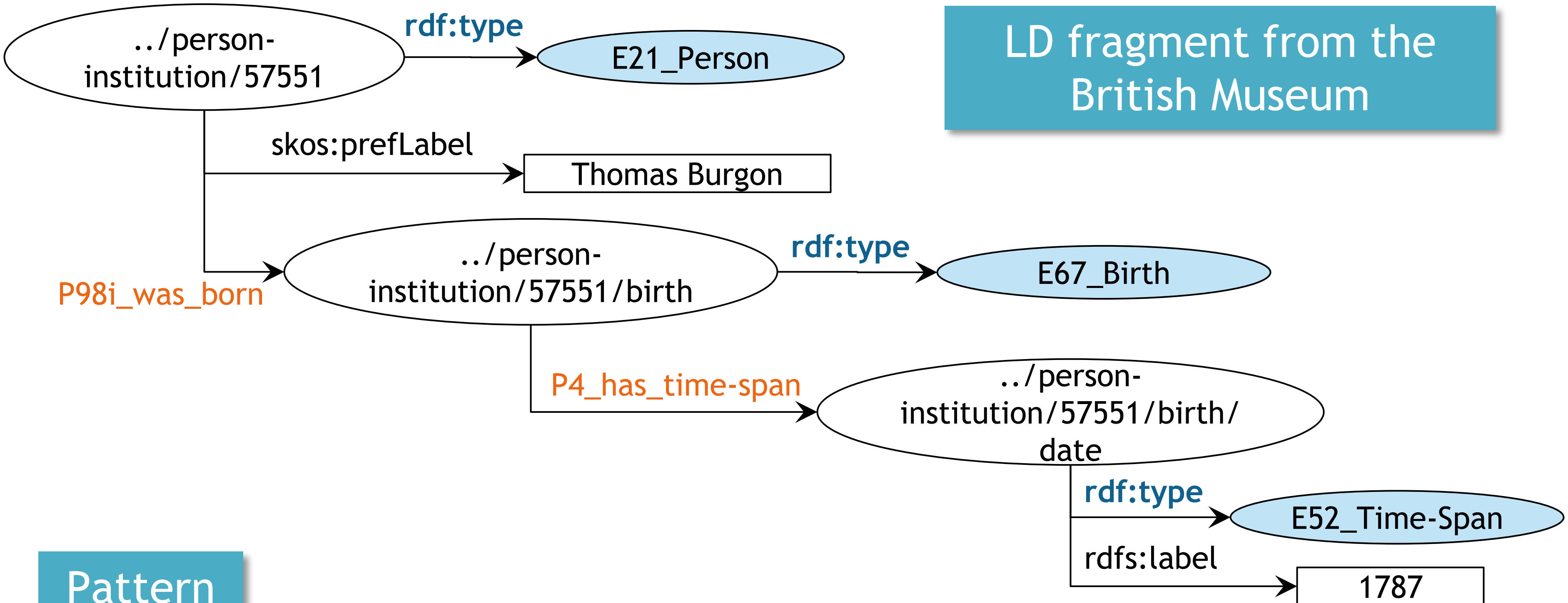
Schema-Level LD Patterns



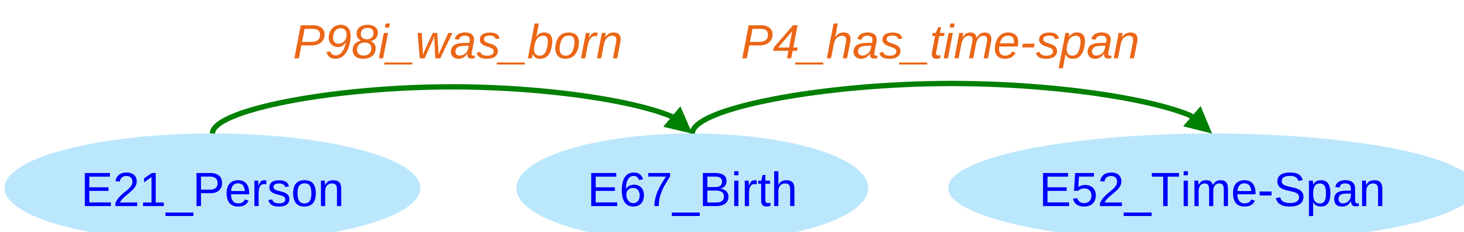
LD fragment from the British Museum

Schema-Level LD Patterns

LD fragment from the British Museum

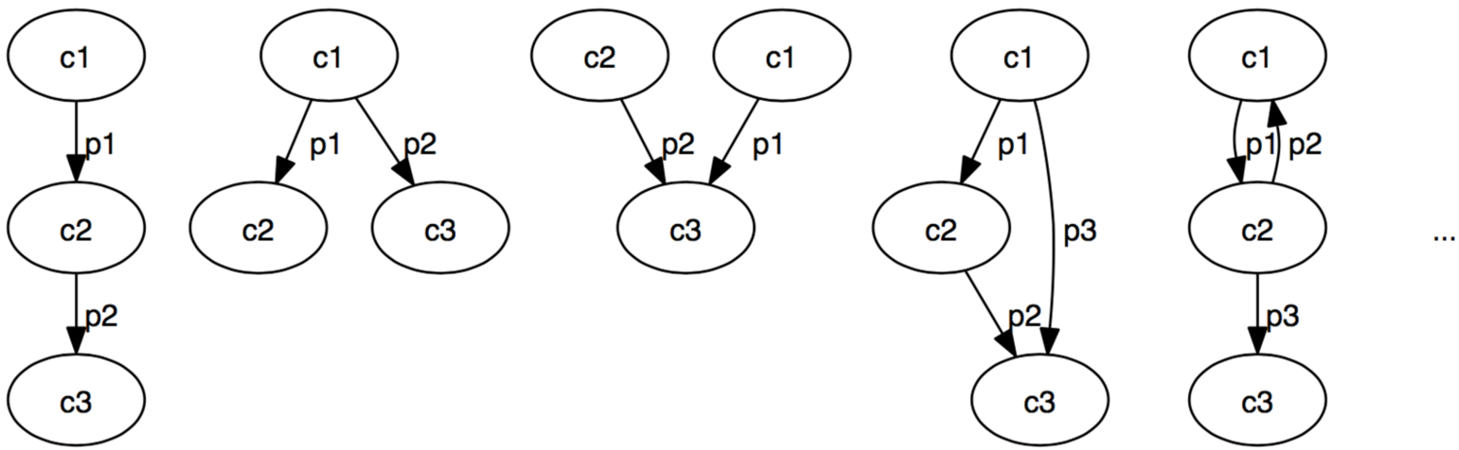


Pattern



Pattern Templates

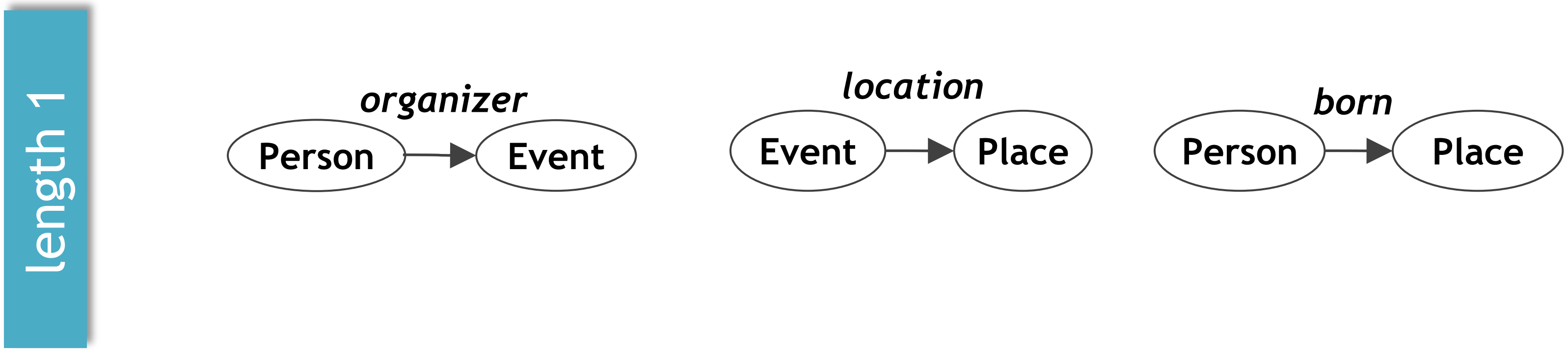
- Many possible templates for patterns
  - Example: patterns for classes C1, C2, C3



- Consider only tree patterns
- Limit the length of the patterns

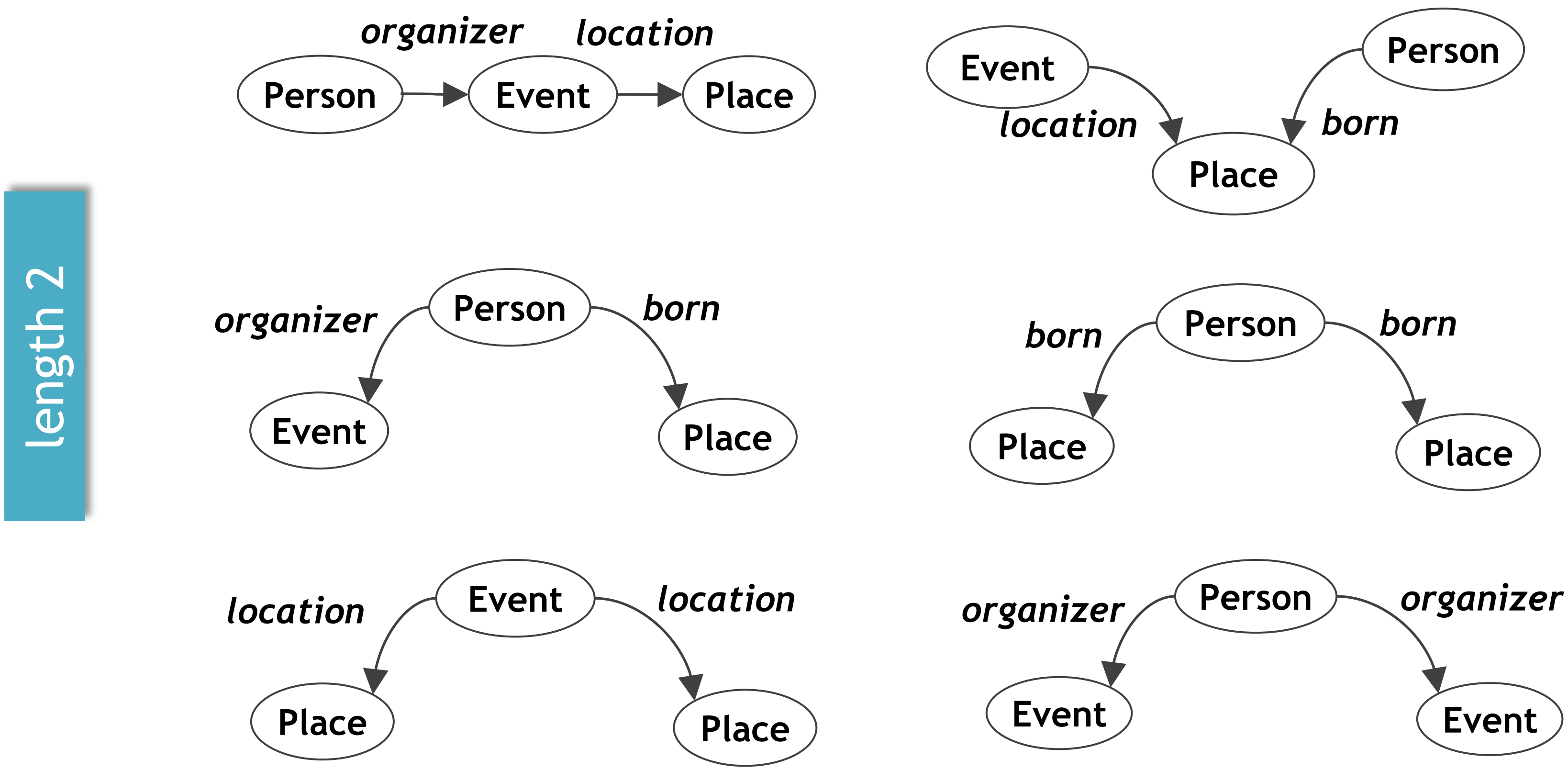
Extracting LD Patterns

- Use SPARQL to extract patterns of length one



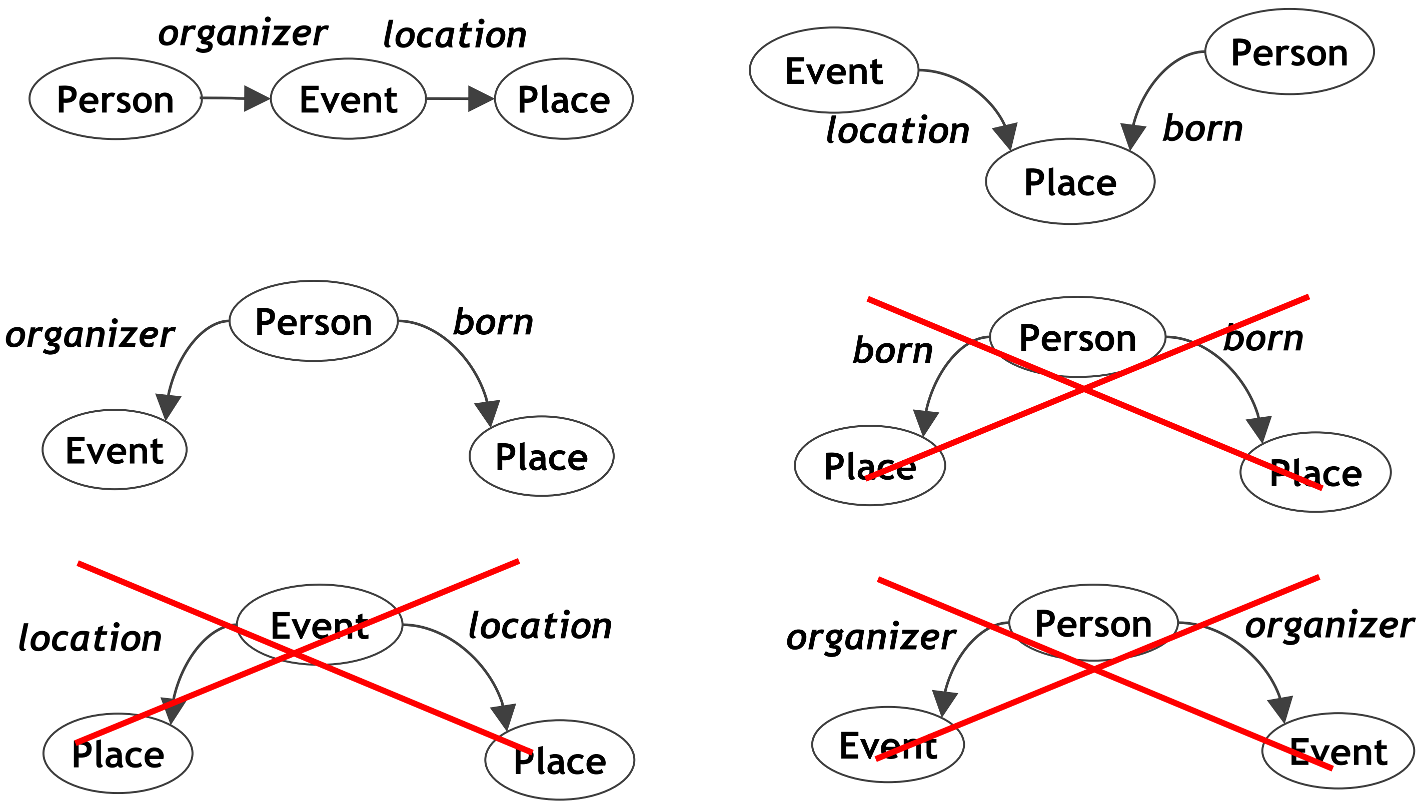
Extracting LD Patterns

- Iteratively construct larger patterns by joining with patterns of length 1



Extracting LD Patterns

- Filter out the patterns not appearing in the data



Approach

Input

- Target source (S)
- Domain Ontologies (O)
- Semantic labels of S
- Linked Data (in the same domain)

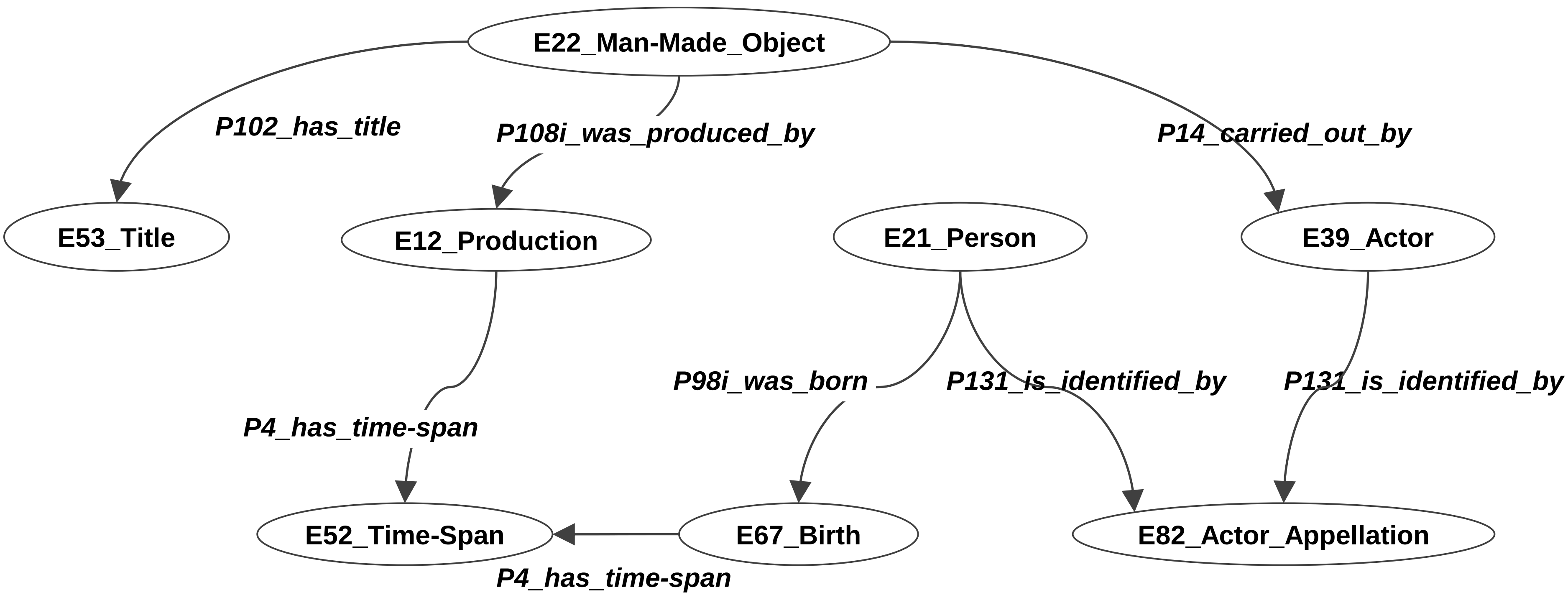
Output

A ranked set of semantic models for S

1. Extract schema-level graph patterns from LD
2. Construct a graph from LD patterns and the ontology
3. Generate and rank semantic models

Merge the Patterns into a Graph

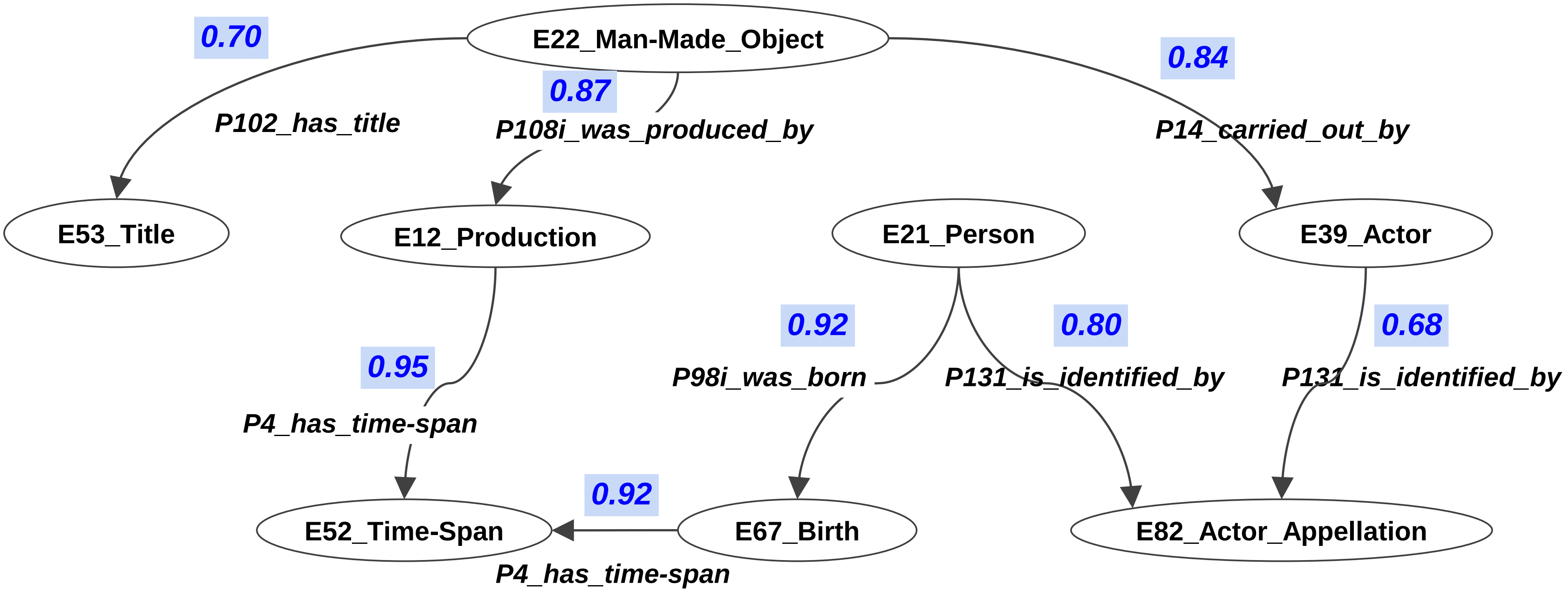
Start from longer patterns, skip the ones already in the graph



Weighting the Links

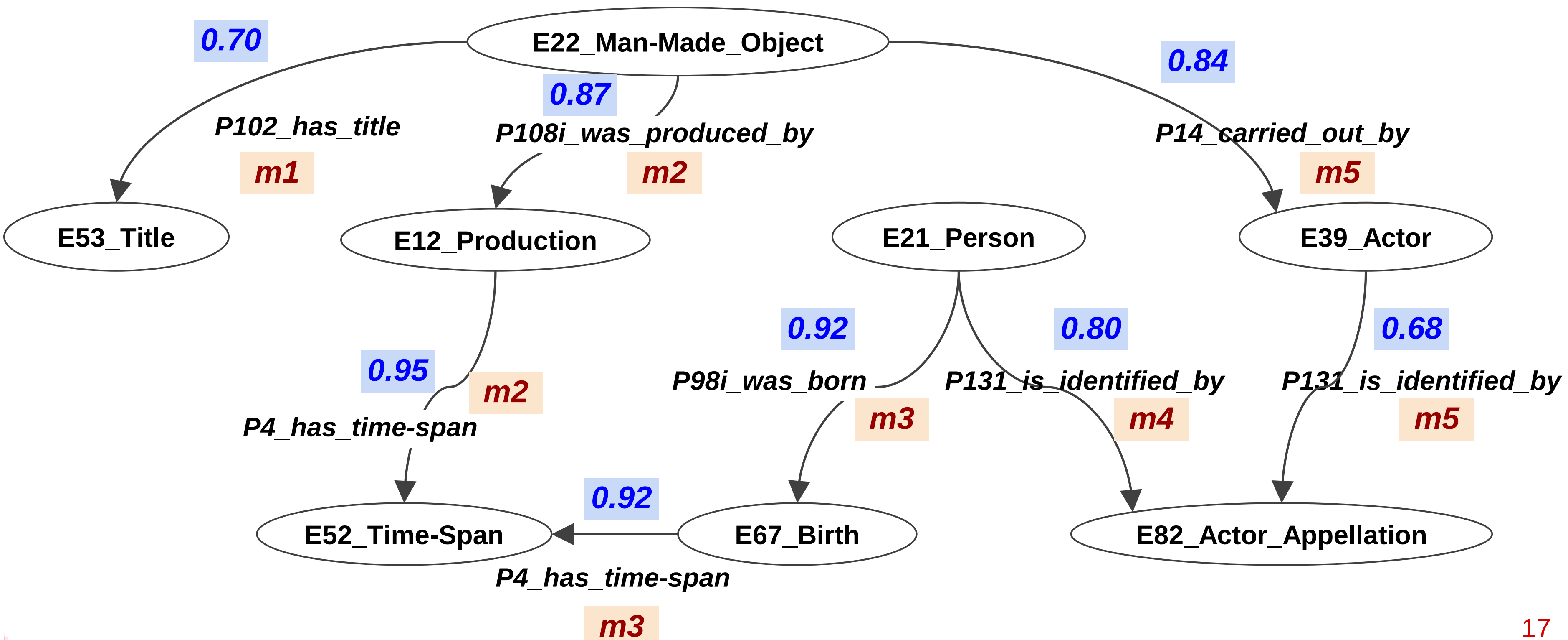
Less weight for more popular links

W = (1 - freq)/(total count of links)



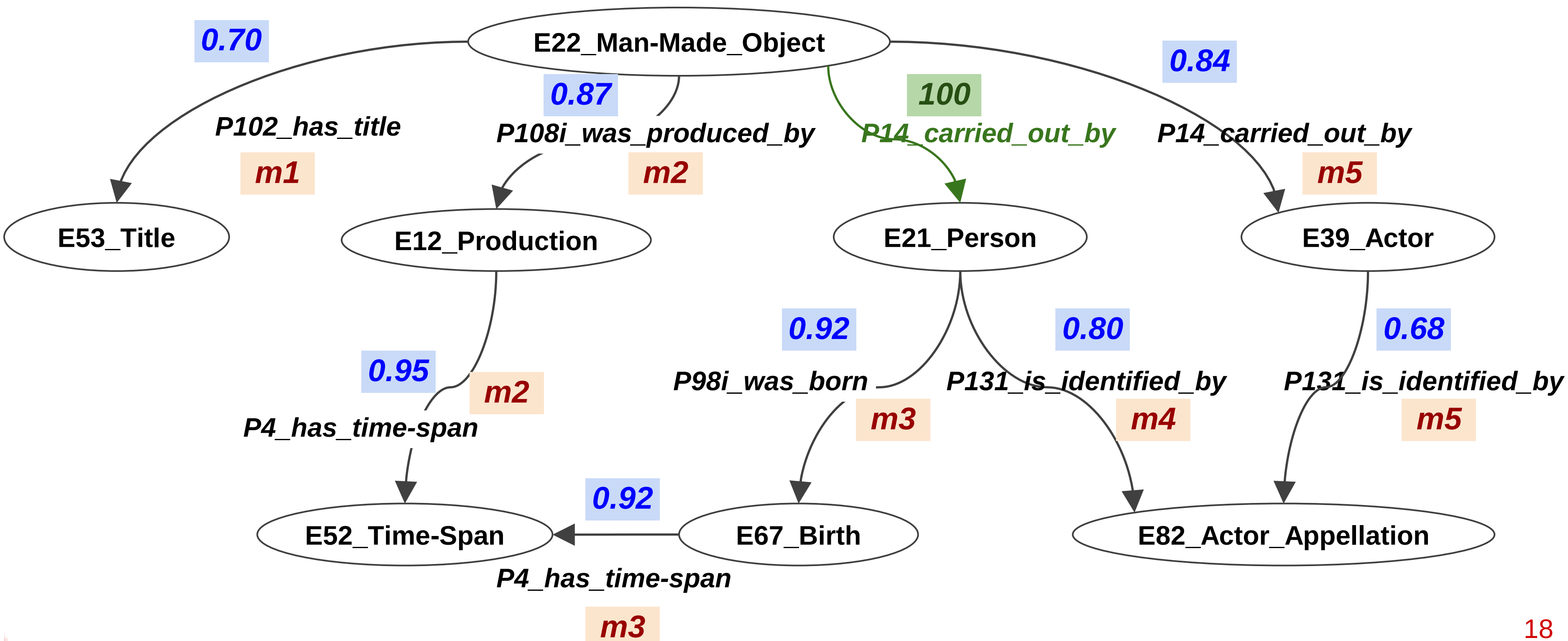
Coherence

Links from the same pattern have the same tag



Add the paths from the Ontology

High weights for links that do not have any instance in the data



Approach

Input

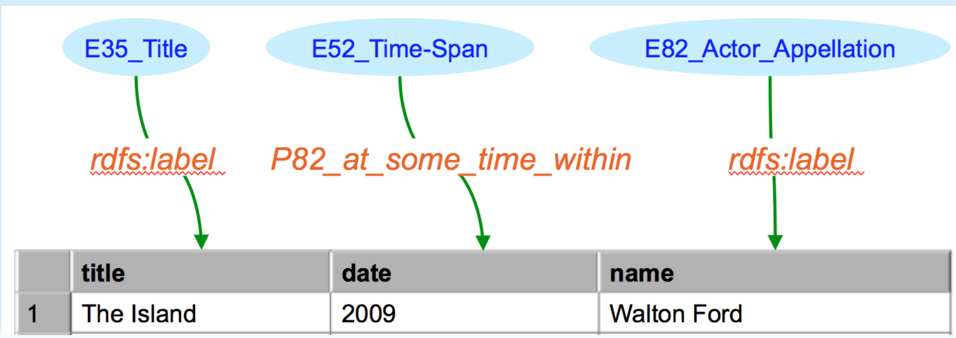
- Target source (S)
- Domain Ontologies (O)
- Semantic labels of S
- Linked Data (in the same domain)

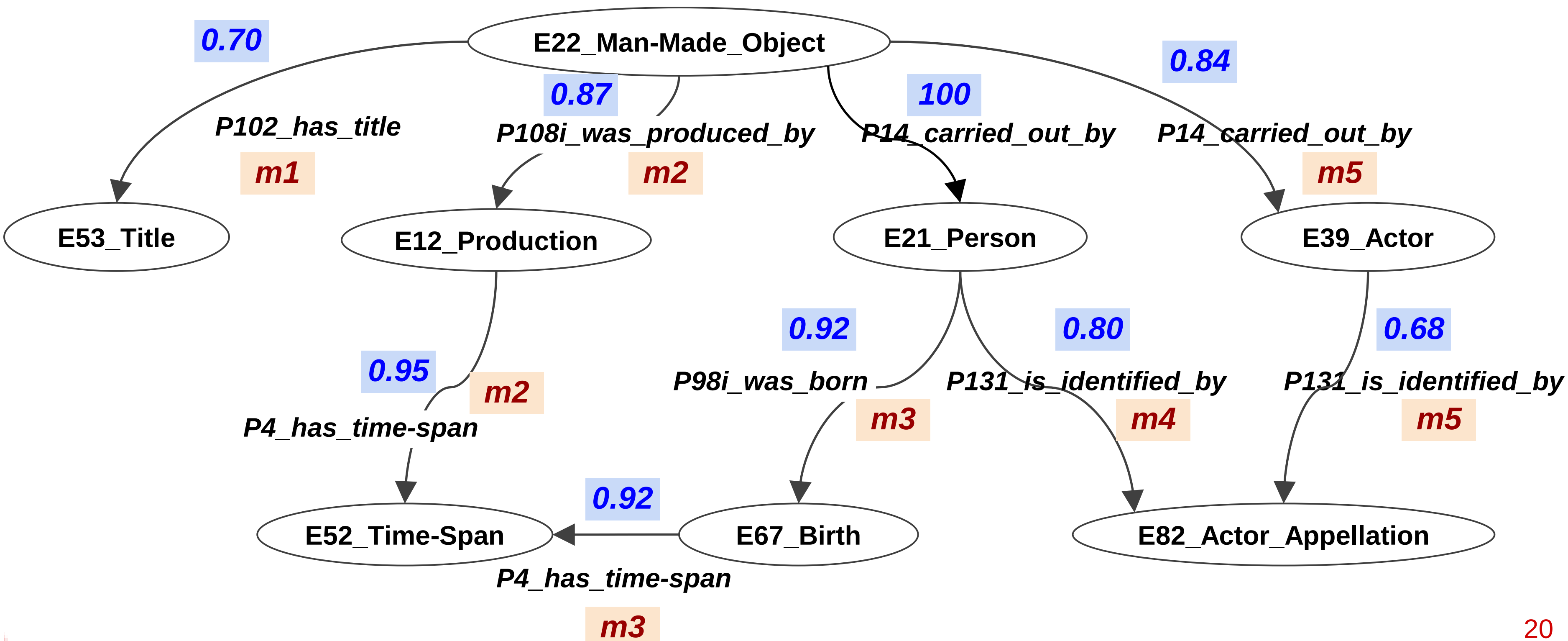
Output

A ranked set of semantic models for S

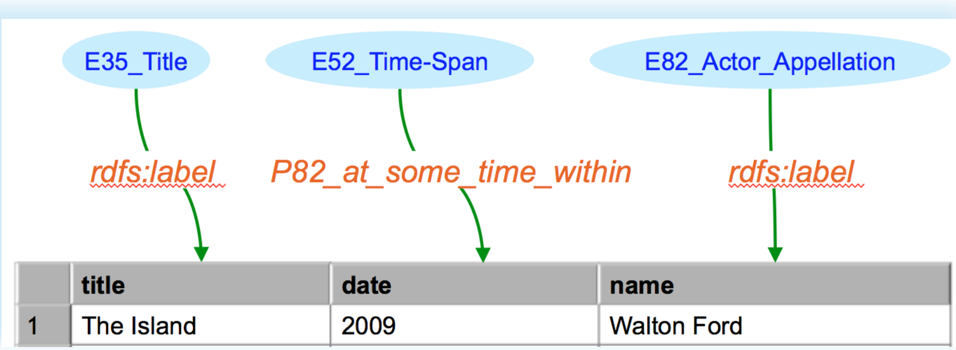
1. Extract schema-level graph patterns from LD
2. Construct a graph from LD patterns and the ontology
3. Generate and rank semantic models

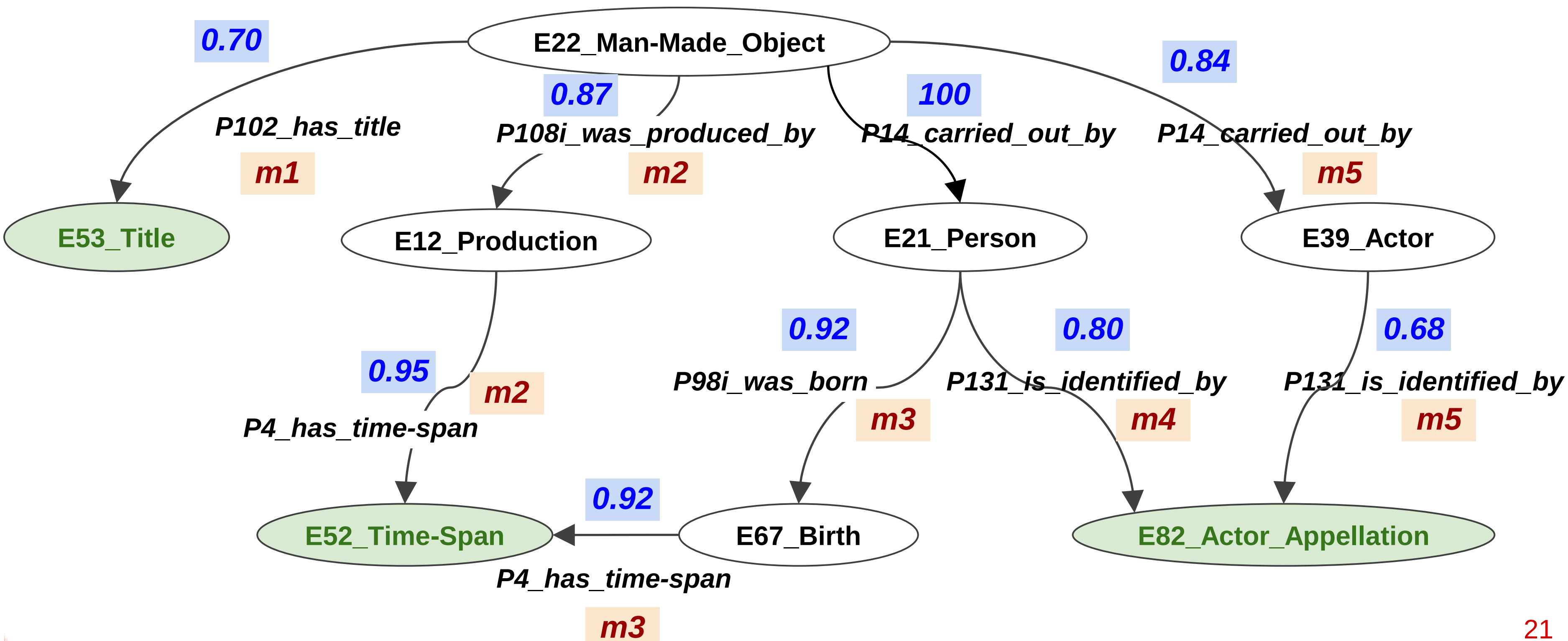
Map Semantic Labels to the Graph





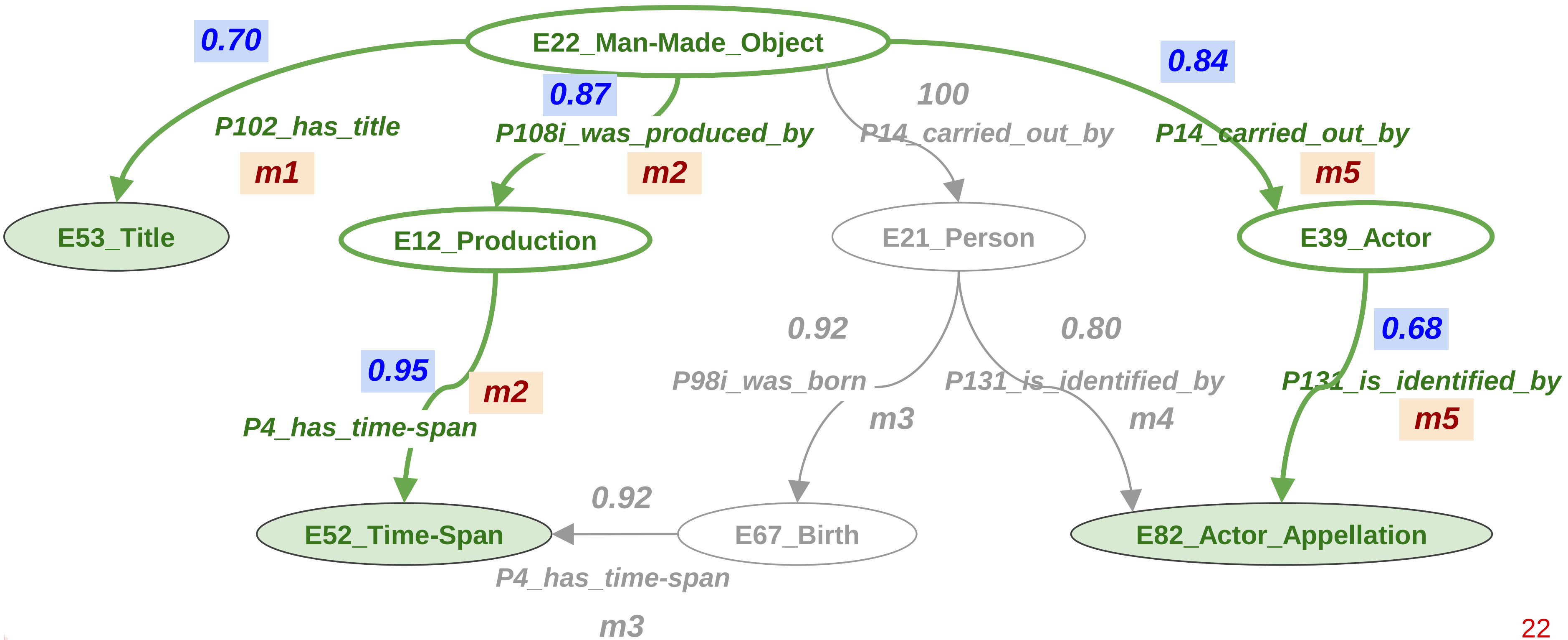
Map Semantic Labels to the Graph





Generate Semantic Models

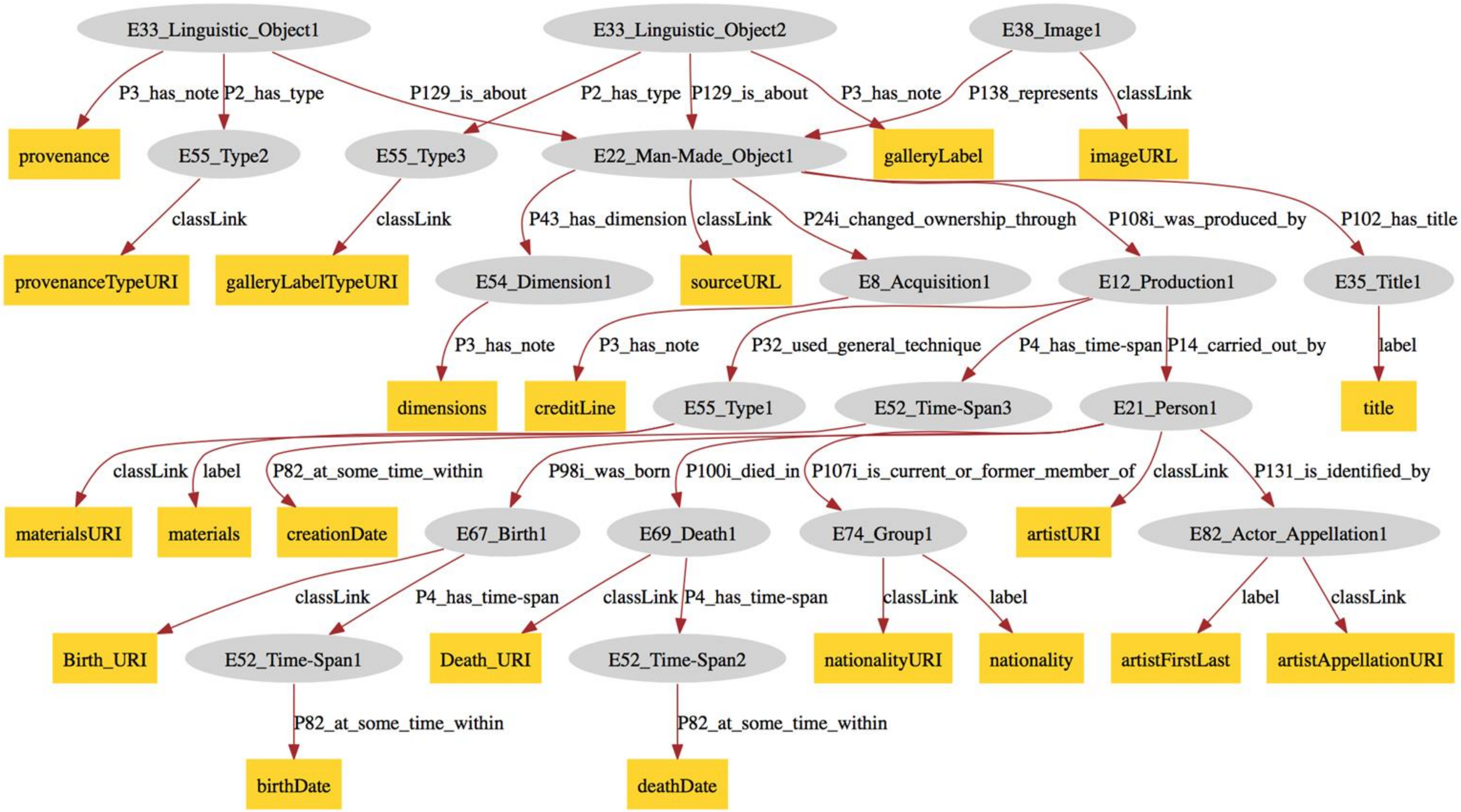
- Compute top k minimal trees
- Consider both coherence and popularity



Evaluation

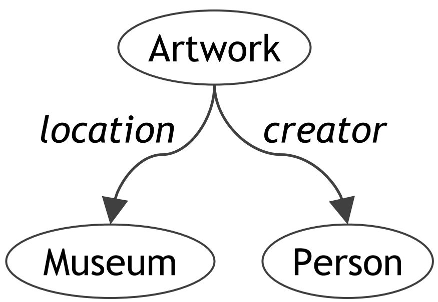
Dataset	Ontology	Gold Standard Models	Linked Data
29 museum data sources 458 attributes (columns)	**CRM** 147 classes 409 properties	852 nodes 825 links	RDF generated from the same dataset (leave-one-out)
29 museum data sources 458 attributes	**CRM** 147 classes 409 properties	852 nodes 825 links	RDF published by Smithsonian American Art Museum (more than 3 million triples)
29 museum data sources 329 attributes	**EDM** 147 classes 409 properties	470 nodes 441 links	RDF generated from the same dataset (leave-one-out)
15 sources containing data about **weapon ads** 175 attributes	**schema.org** (ext) 736 classes 1081 properties	261 nodes 246 links	RDF generated from the same dataset (leave-one-out)

Example Gold Standard Models

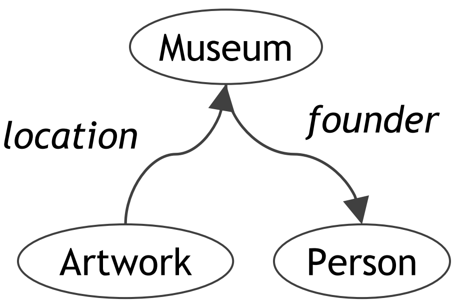


Evaluation

- Compute precision and recall (between learned links and correct links)
- Correct semantic labels are given



correct model



learned model

<Artwork,location,Museum>
<Artwork,creator,Person>

<Museum,founder,Person>
<Artwork,location,Museum>

Precision: 0.5
Recall: 0.5

Results

max len of patterns	Museum CRM (leave-one-out)		Museum CRM (Smithsonian LD)		Museum EDM		Weapon schema.org	
	precision	recall	precision	recall	precision	recall	precision	recall
0	0.07	0.05	0.07	0.05	0.01	0.01	0.03	0.02
1	0.60	0.60	0.28	0.29	0.85	0.78	0.84	0.79
2	0.64	0.67	0.53	0.58	0.81	0.81	0.83	0.79
…	…	…	…	…	…	…	…	…
5	0.75	0.77	0.61	0.67	0.83	0.82	0.86	0.82

- Very low accuracy if only using the ontology paths
- Considering coherence improves the quality of the models (longer patterns increase the accuracy)
- Higher precision & recall for less complex ontologies

Related Work

- Understand semantics of Web tables [Wang et al., 2012] [Limaye et al., 2010] [Venetis et al., 2011]

- Link table values to the LOD entities [Muoz et al., 2013] [Mulwad et al., 2013]

- Learn semantic models from previously modeled sources (Karma) [Taheriyan et al, 2015]

- Extract schema-level patterns (SLPs, length one) from LOD [Schaible et al., 2016]
  - E.g., ({Person,Player},{knows},{Person,Coach})

Discussion

- Manually constructing semantic models is hard & expensive
  - Needs domain knowledge and expertise in SW technologies
  - Often requires many user interactions in modeling tools
- Infer semantic relations from linked data
  - The suggested model can be refined in tools such as Karma
- Help to publish consistent RDF data