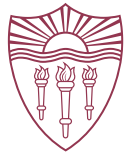


Automatic Source Modeling

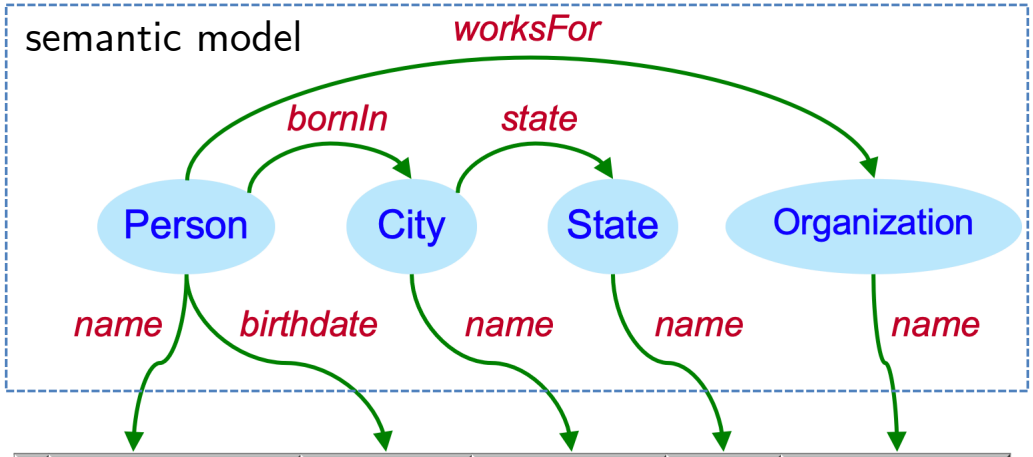
INF 558: Building Knowledge Graph

Binh Vu

*Based on slides by Pedro Szekely, Craig Knoblock

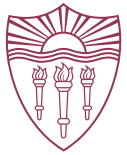
The Problem

- Build a semantic model that describes a data source using classes and relationships in target ontologies.
- The semantic model can be used to create a source mapping (e.g., R2RML) for data discovery and data integration.

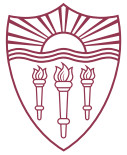


	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Overview



- Supervised approach
  - Build semantic models using known semantic models
- Unsupervised approach
  - Build semantic models using knowledge bases.



Supervised Approach

Learning the semantics of structured data sources.
Taheriyan et al., JWS 2015

Main Idea

Sources in the same domain often have similar data

Étagère
Alexander Roux
(1813–1886)
Date: ca. 1855
Medium: Rosewood, chestnut, poplar, bird's-eye maple
Accession Number: 1971.219
On view in Gallery 736

Pitcher
French Maker
Date: 1859–60
Medium: Porcelain, overglaze enamel decoration and gilding
Accession Number: 1995.26
On view in Gallery 774

Main Idea

Sources in the same domain often have similar data

[Harvest Home](#)
1887
Edwin Austin Abbey, born Philadelphia, PA 1852–
died London, England 1911
pen and ink on paperboard
sheet: 21 x 14 1/4 in. (53.3 x 36.2 cm)
Smithsonian American Art Museum, Gift of Sen. Stuart
Symington and Rep. James W. Symington
1973.28.2
Not on view

ple

[The Queen in "Hamlet"](#)
1895
Edwin Austin Abbey, born Philadelphia, PA 1852–
died London, England 1911
pastel on paperboard
sheet: 27 7/8 x 21 7/8 in. (70.8 x 55.6
medium
Smithsonian American Art Museum, Gift of Sen. Stuart
Symington and Rep. James W. Symington
1973.28.1
Not on view

nd gilding

Main Idea

Sources in the same domain often have similar data

Harvest Home

First Previous 1 2 3 .. 715 716 Next Last

view lightbox view list view single item

View

Copyright Info

Woman's marriage or ceremonial veil

Date: 1900-1930

Dimensions: Overall: 60 x 52 1/2 in. (1 m 52.4 cm x 133.35 cm)

Medium: Wool and natural dyes, including henna

Credit Line: Dallas Museum of Art, Textile Purchase Fund

Description: Resist dyed (tie-dye) technique

Geographic location:

Not on view

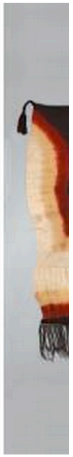
Main Idea

Sources in the same domain often have similar data

Harvest Home

First Previous 1 2 3 .. 715 716 Next Last

view lightbox



Geographic location:

Not on view

NEXT WORK

COURT OF BENIN, EDO CULTURE
Nigeria
Commemorative Head of a King
16th–17th century
Copper alloy
11 1/2 x 9 x 9 inches

The Museum of Fine Arts, Houston
Museum purchase with funds provided by the Alice Pratt Brown Museum Fund and gift of Oliver E. and Pamela F. Cobb

Department of the Arts of Africa, Oceania, & the Americas

Arts of Africa

ABOUT

The most important Benin artworks were life-size heads of the *obas*, the spiritual and corporeal kings of Benin. Ordered in pairs by every new king to honor his predecessor, these heads were arranged symmetrically on altars as representations of the institution of divine kingship. This king's head dates

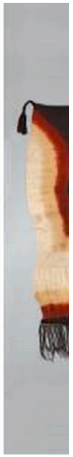
Main Idea

Sources in the same domain often have similar data

Harvest Home

First Previous 1 2 3 .. 715 716 Next Last

view lightbox



Geographic location:

Not on view

NEXT WORK

COURT OF BENIN, EDO CULTURE
Nigeria
Commemorative Head of a King
16th–17th century
Copper alloy
11 1/2 x 9 x 9 inches

The Museum of Fine Arts, Houston
Museum purchase with funds provided by the Alice Pratt Brown Museum Fund and gift of Oliver E. and Pamela F. Cobb

Department of the Arts of Africa, Oceania, & the Americas

Arts of Africa

ABOUT

The most important Benin artworks were life-size heads of the *obas*, the spiritual and corporeal kings of Benin. Ordered in pairs by every new king to honor his predecessor, these heads were arranged symmetrically on altars as representations of the institution of divine kingship. This king's head dates

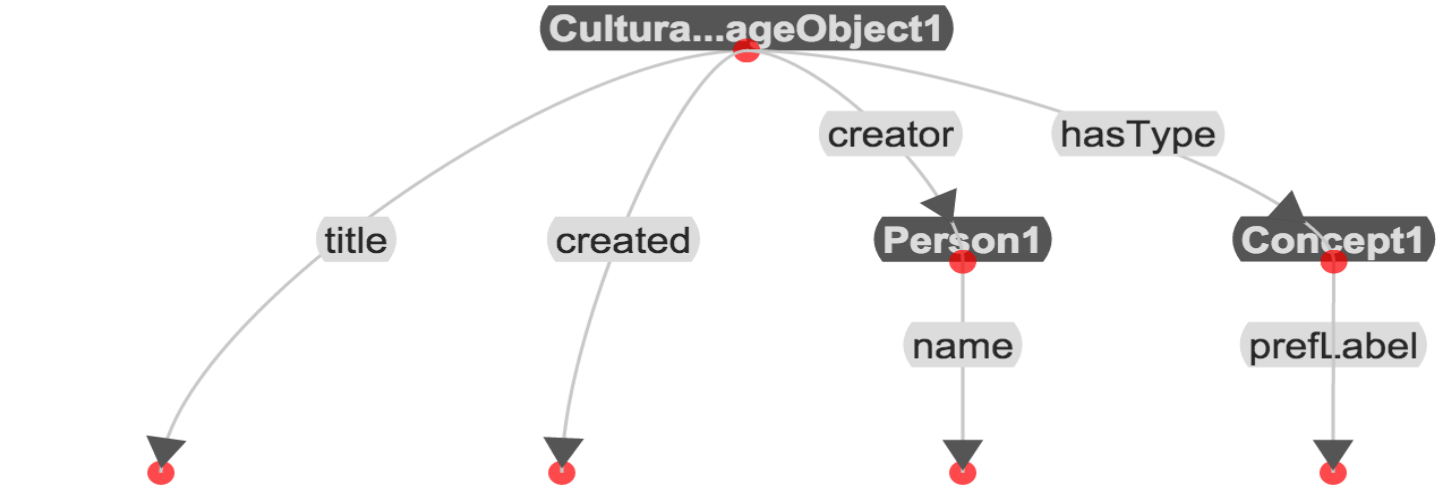
Exploit knowledge of known semantic models to hypothesize a semantic model for a new sources

Example

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: Dallas Museum of Art ➔ **dma(title,creationDate,name,type)**



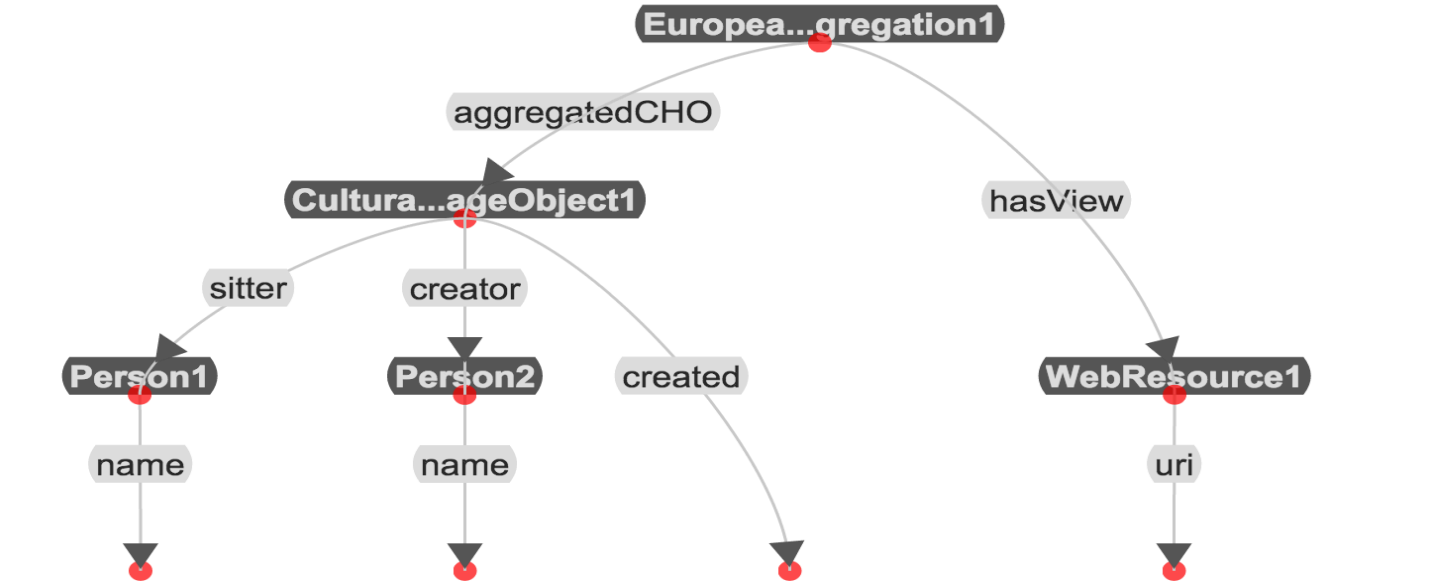
title	creationDate	name	type
Stream in the Mountains	1825	George Inness	Paintings
Lady Godiva	1856	Anne Whitney	Sculpture

Example

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: National Portrait Gallery ➔ **npg(name,artist,year,image)**



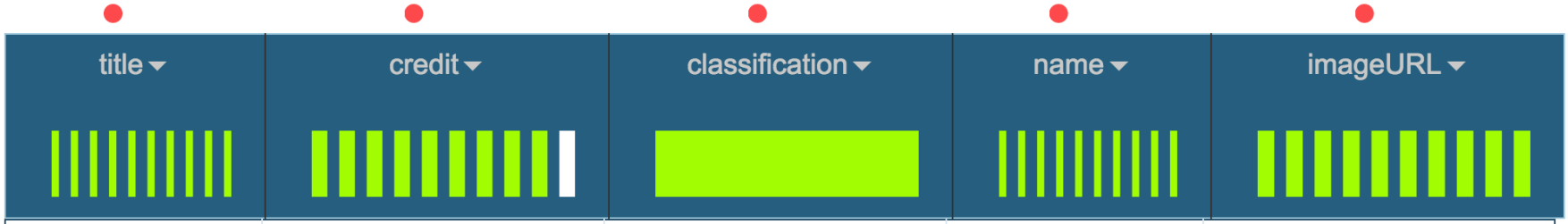
name	artist	year	image
Neil Armstrong	Louis Glanzman	1969	http://npgportraits.si.e…
Henry Larcom Abbot	Nahum Ball Onthank	1857	http://npgportraits.si.e…

Example

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: Detroit Institute of Art ➔ **dia(title,credit,classification,name,imageURL)**



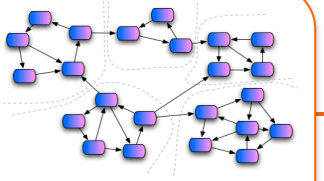
title	credit	classification	name	imageURL
Indian Telegraph	Formerly in the collection of:The artist's student, Mrs. M. Noble Brainard, 1868-	Paintings	John Mix Stanley	http://www.dia.org/us… S1.jpg
Winter Landscape	*Empty*	Paintings	Mortimer L. Smith	http://www.dia.org/us… S1.jpg
The River Seine	Theodore Scott DaboDIA in 1906	Paintings	Theodore Scott Dabo	http://www.dia.org/us… D1.jpg

Goal: Automatically suggest a semantic model for ***dia***

Approach

Sample Data

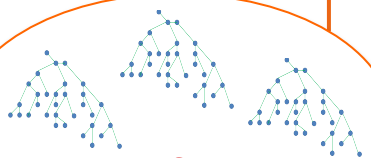
Learn Semantic Types



Construct a Graph

Generate Candidate Models

Domain Ontology



Known Semantic Models

Rank Results

Implemented in **Karma**

Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

1. Learn semantic types for attributes(s)
2. Construct Graph G=(V,E)
3. Generate mappings between attributes(S) and V
4. Generate and rank semantic models

Output

- A ranked set of semantic models for S

Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

1. Learn semantic types for attributes(s)

   Construct Graph G=(V,E)

   Generate mappings between attributes(S) and V

   Generate and rank semantic models

Output

- A ranked set of semantic models for S

Learn Semantic Types

- Learn *Semantic Types* for each attribute from its data
- Semantic Type: **<class_uri, property_uri>**
- Method
  - Textual attributes: cosine similarity between TF/IDF vectors
  - Numerical attributes: statistical hypothesis testing
- Pick top K semantic types according to their confidence values

dia(title,credit, classification,name,imageURL)		
title	<aac:CulturalHeritageObject, dcterms:title>	0.49
	<aac:CulturalHeritageObject, rdfs:label>	0.28
credit	<aac:CulturalHeritageObject, dcterms:provenance>	0.83
	<aac:Person, ElementsGr2:note>	0.06
classification	<skos:Concept, skos:prefLabel>	0.58
	<skos:Concept, rdfs:label>	0.41
name	<aac:Person, foaf:name>	0.65
	<fofa:Person, fofaf:name>	0.32
imageURL	<foaf:Document, uri>	0.47
	<edm:WebResource, uri>	0.40

Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

✓ Learn semantic types for attributes(s)

2 Construct Graph G=(V,E)

Generate mappings between attributes(S) and V

Generate and rank semantic models

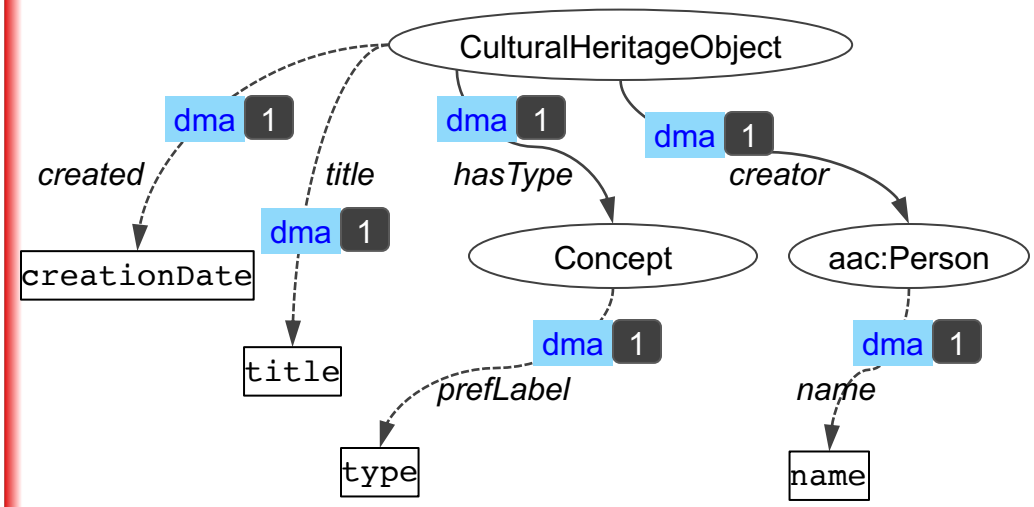
Output

- A ranked set of semantic models for S

Build Graph G: Add Known Models

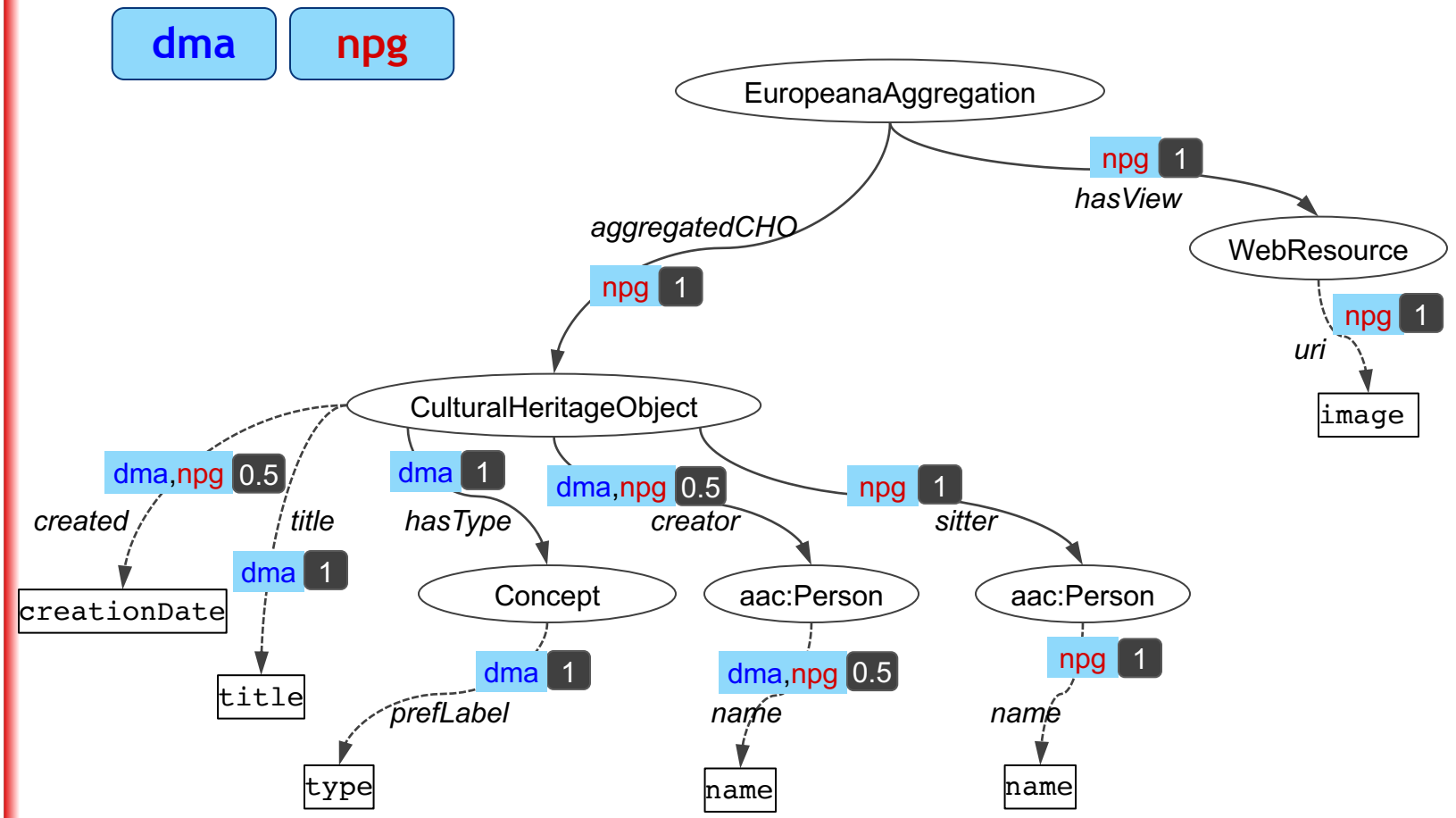
- Annotate nodes and links with list of supporting models
- Adjust weight based on the number of supporting models

dma



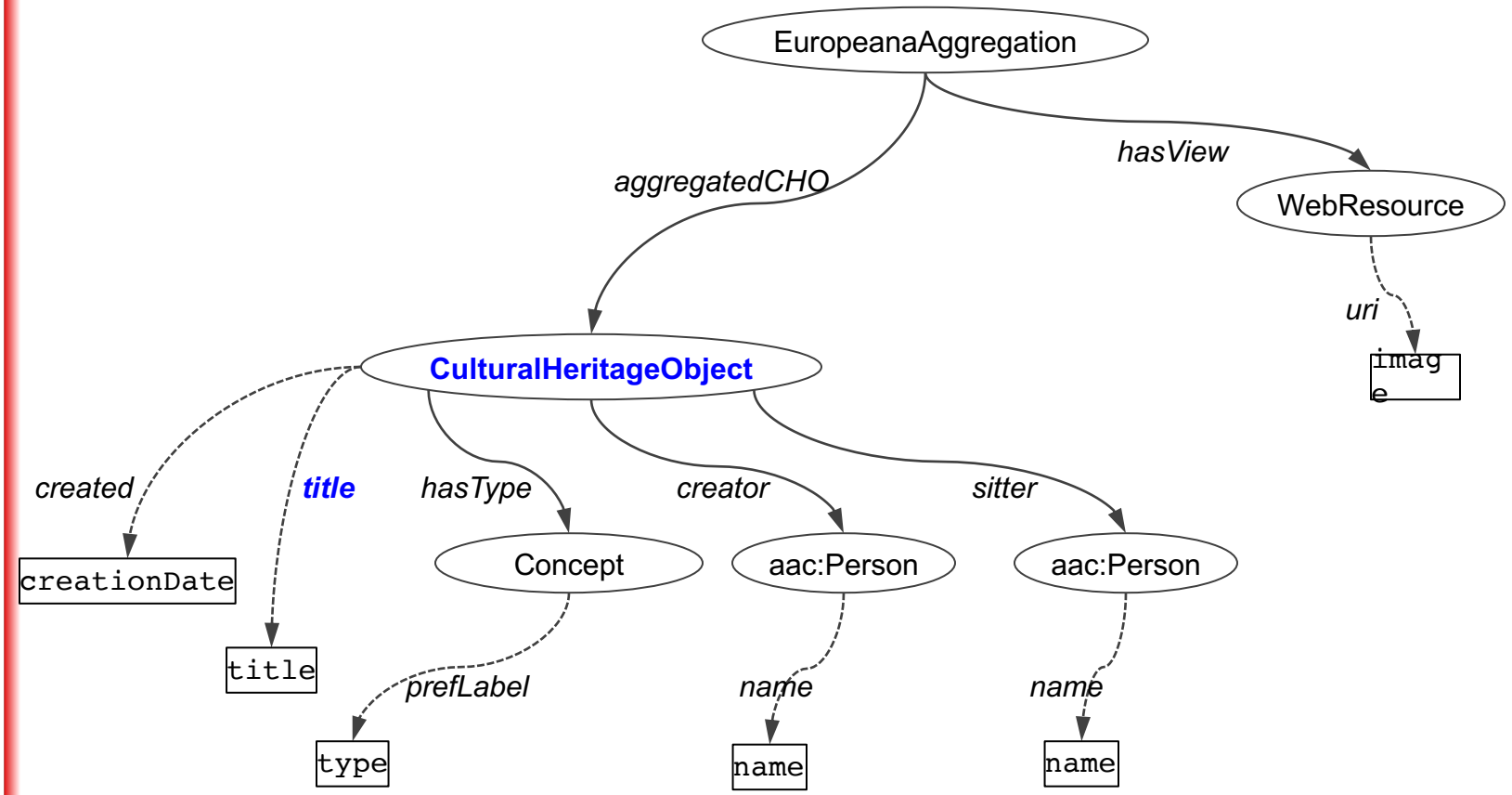
Build Graph G: Add Known Models

- Annotate nodes and links with list of supporting models
- Adjust weight based on the number of supporting models



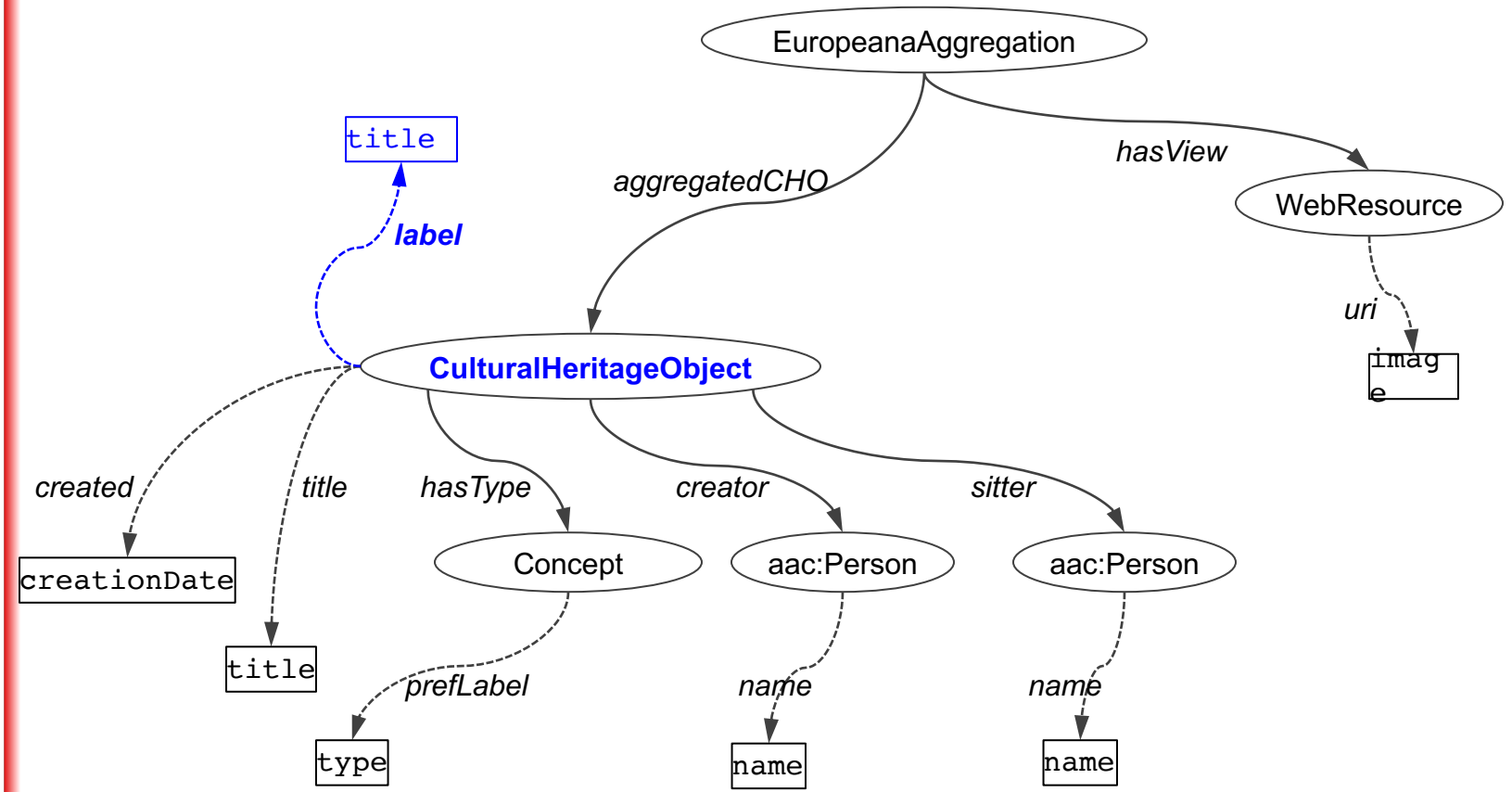
Build Graph G: Add Semantic Types

title	**<CulturalHeritageObject,title>** <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



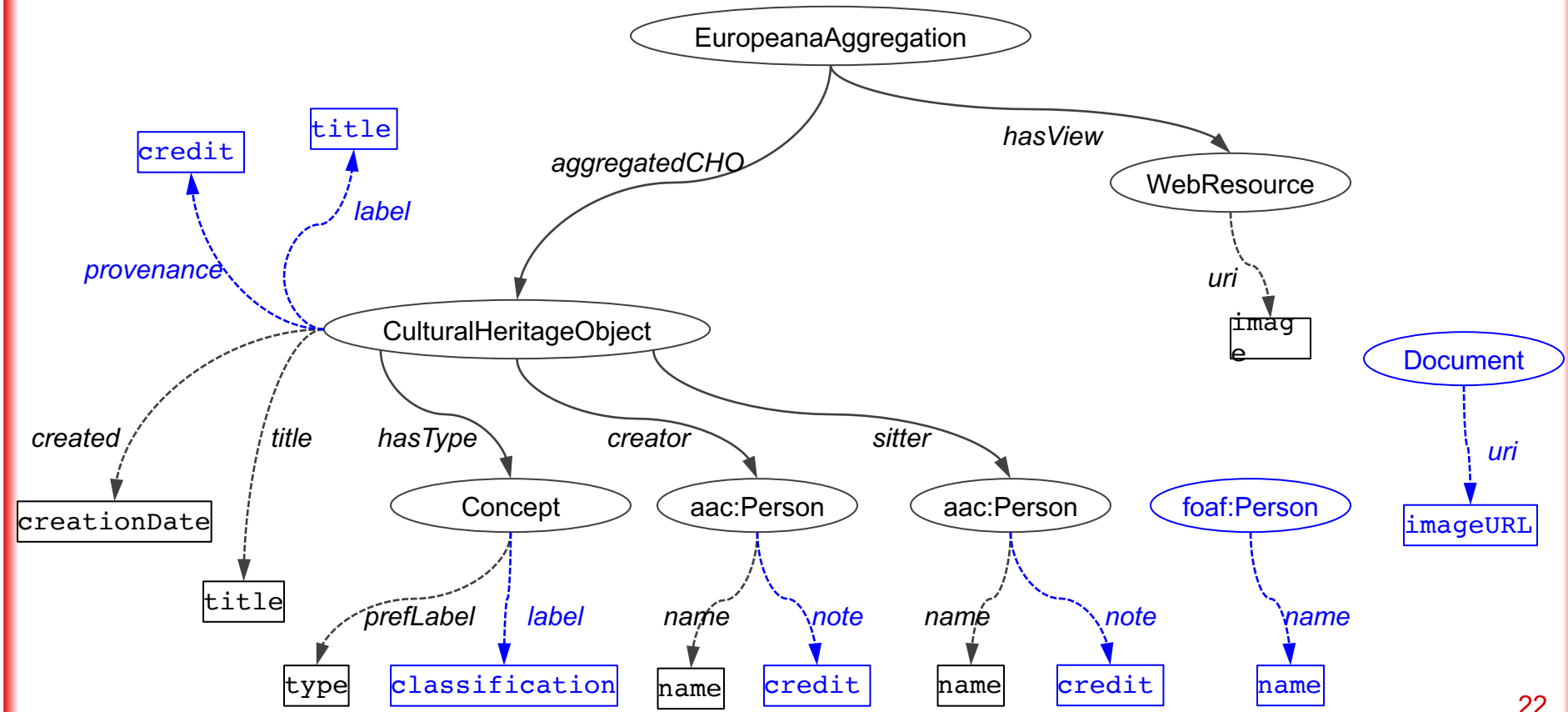
Build Graph G: Add Semantic Types

title	<CulturalHeritageObject,title> **<CulturalHeritageObject,label>**
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



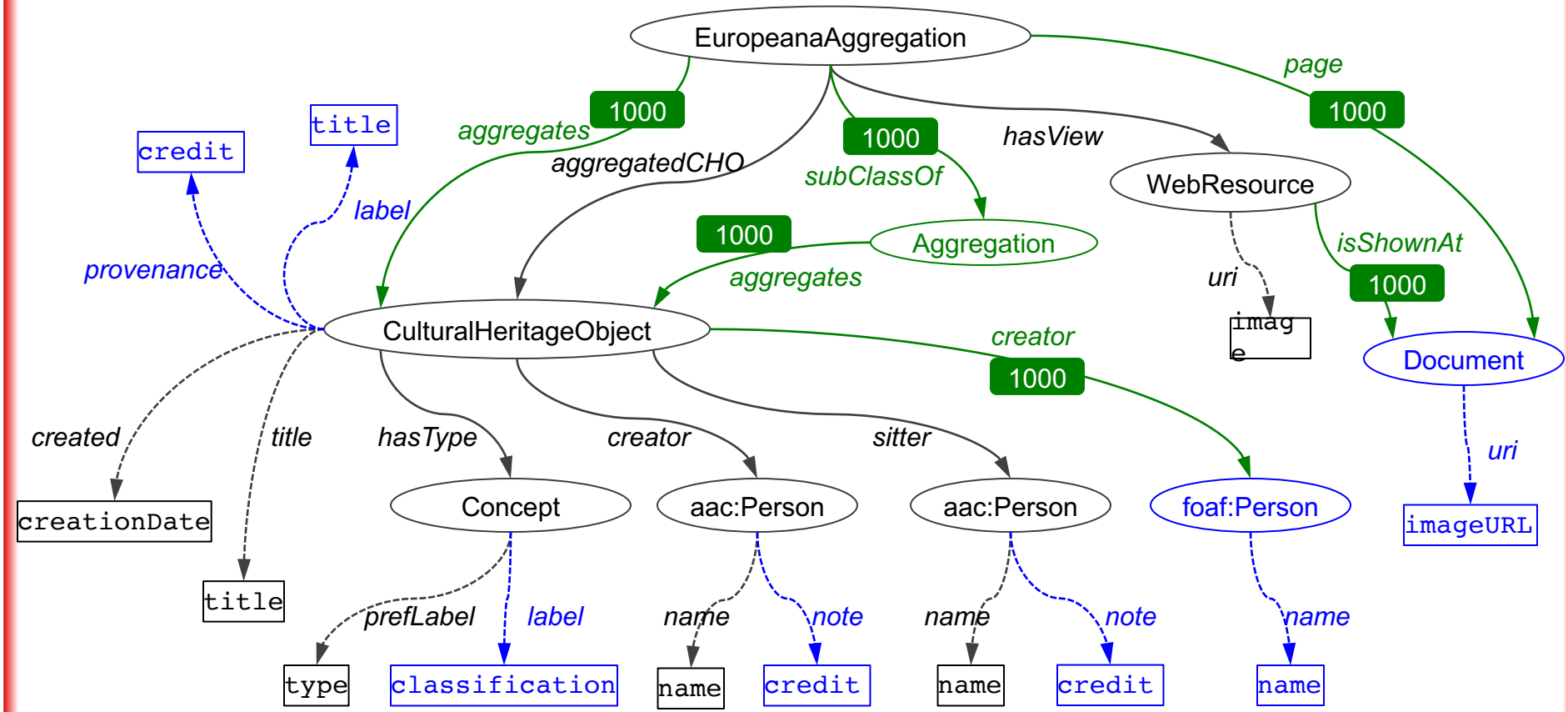
Build Graph G: Add Semantic Types

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



Build Graph G: Expand with Paths from Ontology

- Assign a high weight to the links coming from the ontology



Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

✓ Learn semantic types for attributes(s)

✓ Construct Graph G=(V,E)

3 Generate mappings between attributes(S) and V

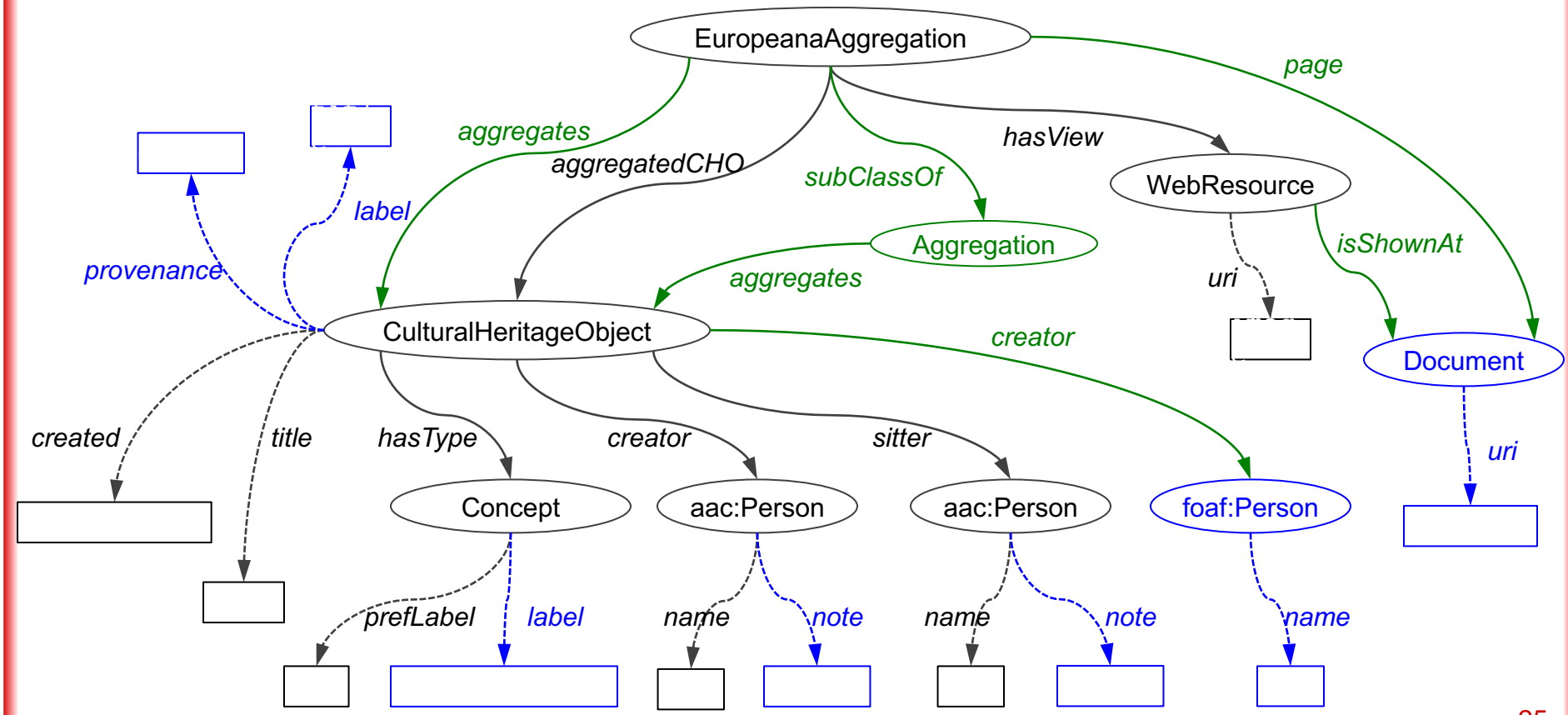
Generate and rank semantic models

Output

- A ranked set of semantic models for S

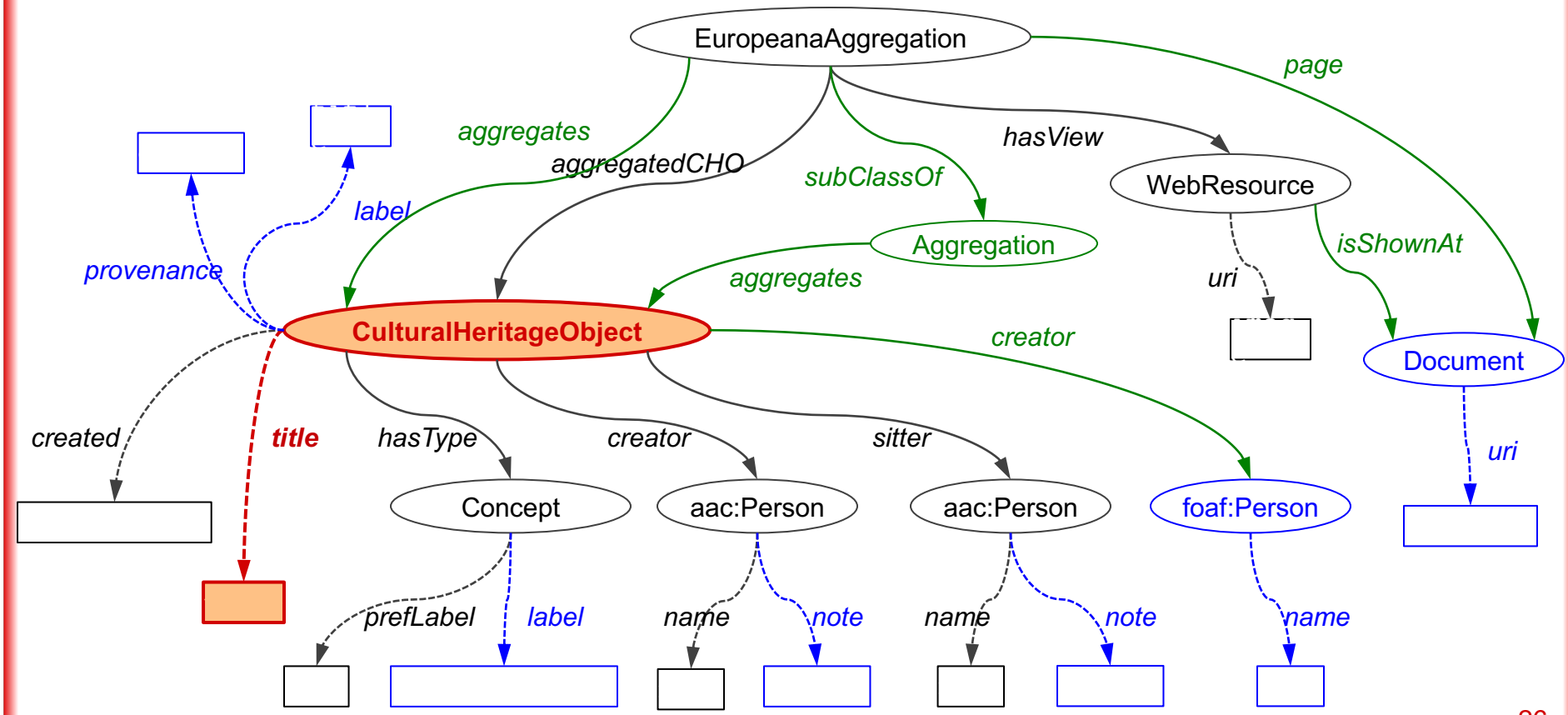
Map Source Attributes to the Graph

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



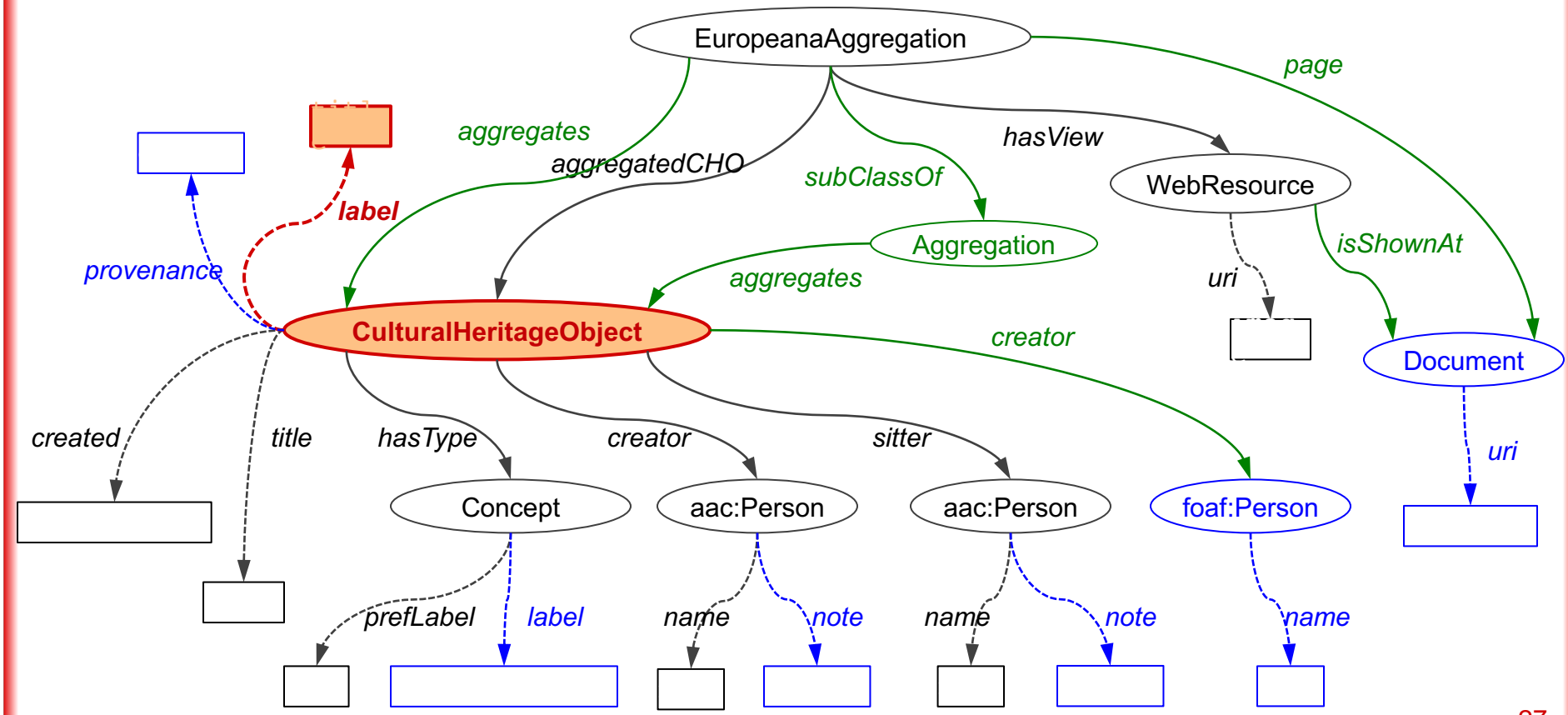
Map Source Attributes to the Graph

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



Map Source Attributes to the Graph

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



Scalability Issue

- Multiple mappings from attributes(S) to nodes of G
  - Each attribute has more than one semantic type
  - Multiple matches for each semantic type
- Not feasible to generate all possible mappings
  - The size of graph may be large
  - The source may have many attributes
- Exponential in terms of number of attributes
  - N attributes, M matches for each → M^N mappings

Prune the Mappings

- Score the partial mappings after mapping each attribute
  - Coherence: number of nodes in a mapping that belong to same component
  - Confidence: average confidence of semantic types in m
  - Score = arithmetic mean of coherence and confidence
- Beam Search
  - Keep only top K mappings after mapping each attribute
- Number of mappings will not exceed a fixed size after mapping each attribute

Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

✓ Learn semantic types for attributes(s)

✓ Construct Graph G=(V,E)

✓ Generate mappings between attributes(S) and V

4 Generate and rank semantic models

Output

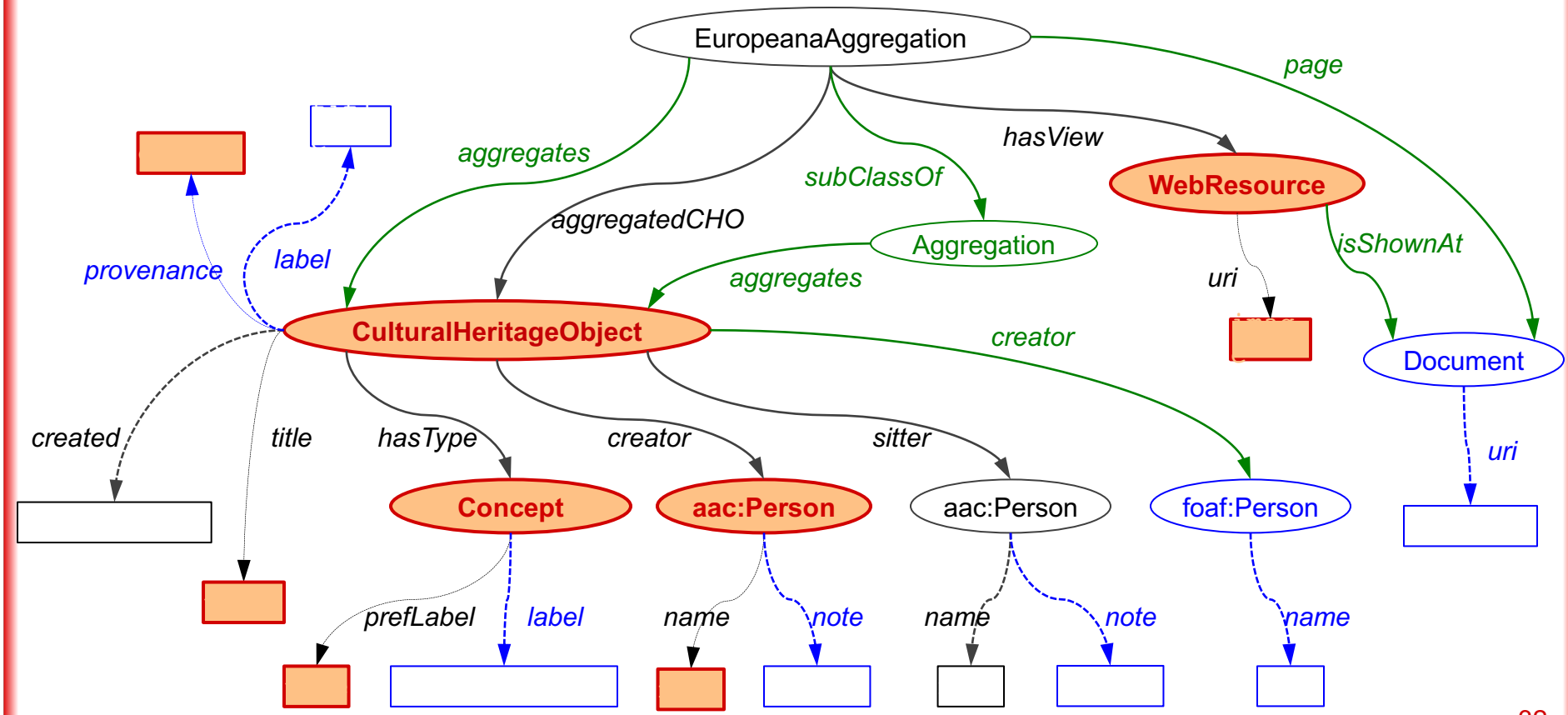
- A ranked set of semantic models for S

Generate Semantic Models

- Compute Top-K Steiner tree for each mapping
  - A minimal tree connecting nodes of mapping
  - A customization of BANKS algorithm [Bhalotia et al., 2002]
- Each tree is a candidate model
- Rank candidate models (Steiner trees)
  - Coherence of the links
  - Cost

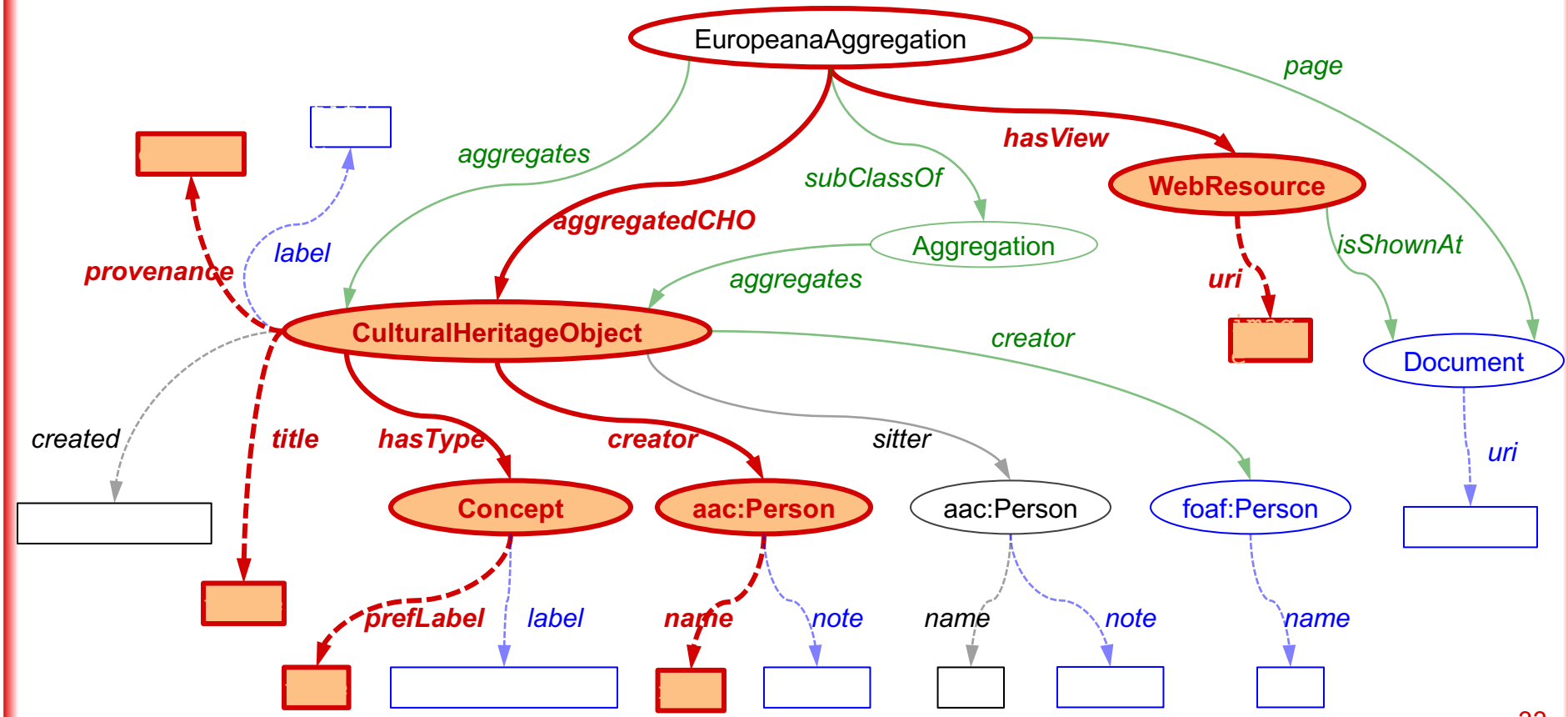
Example Mapping

title	**<CulturalHeritageObject,title>** <CulturalHeritageObject,label>
credit	**<CulturalHeritageObject,provenance>** <Person,note>
classification	**<Concept,prefLabel>** <Concept,label>
name	**<aac:Person,name>** <foaf:Person,name>
imageURL	<Document,uri> **<WebResource,uri>**



Steiner Tree

title	**<CulturalHeritageObject,title>** <CulturalHeritageObject,label>
credit	**<CulturalHeritageObject,provenance>** <Person,note>
classification	**<Concept,prefLabel>** <Concept,label>
name	**<aac:Person,name>** <foaf:Person,name>
imageURL	<Document,uri> **<WebResource,uri>**

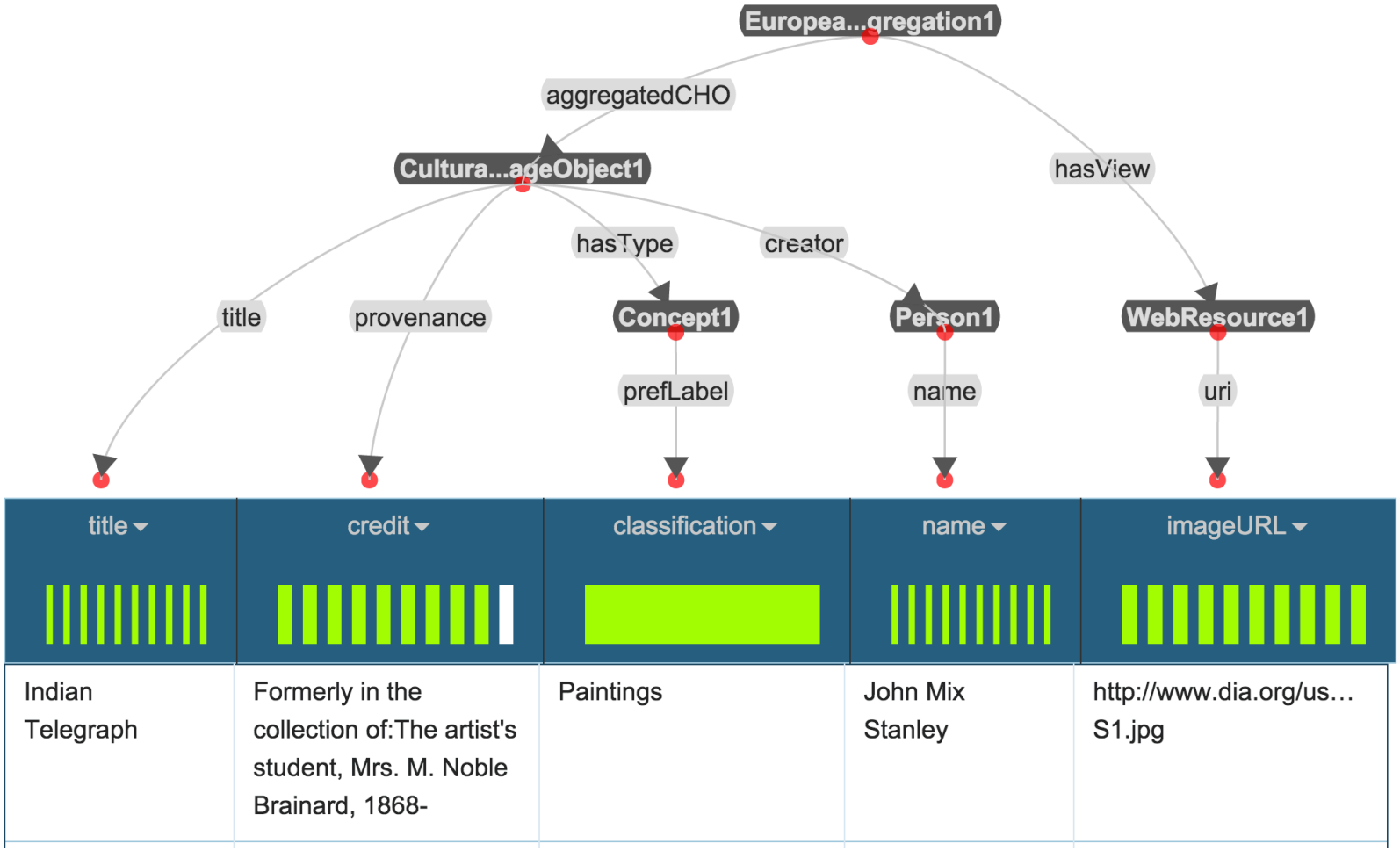


Final Model in Karma

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: Detroit Institute of Art ➔ **dia(title,credit,classification,name,imageURL)**



Evaluation

Evaluation Dataset	EDM	CRM
# sources	29	29
# classes in the ontologies	119	147
# properties in the ontologies	351	409
# nodes in the gold standard models	473	812
# data nodes in the gold standard models	331	418
# class nodes in the gold standard models	142	394
# links in the gold standard models	444	785

Compute precision and recall between learned models and correct models

$$precision = \frac{rel(sm) \cap rel(sm')}{rel(sm')}$$

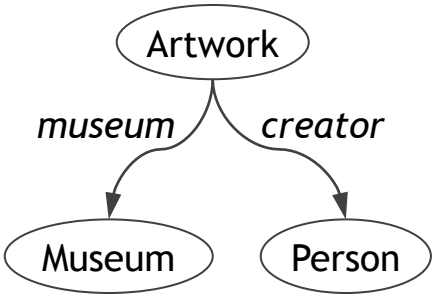
How many of the learned relationships are correct?

$$recall = \frac{rel(sm) \cap rel(sm')}{rel(sm)}$$

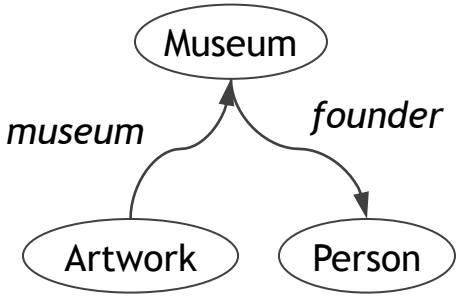
How many of the correct relationships are learned?

rel(sm) is the set of triples <source, link, target> in the semantic model

Example



correct model



learned model

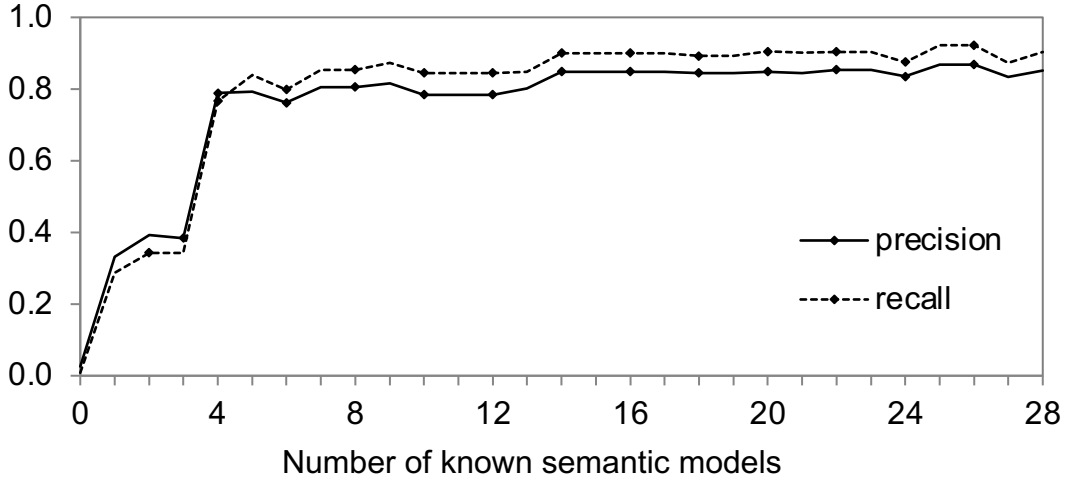
<Artwork,museum,Museum>
<Artwork,creator,Person>

<Artwork,museum,Museum>
<Museum,founder,Person>

Precision: 0.5
Recall: 0.5

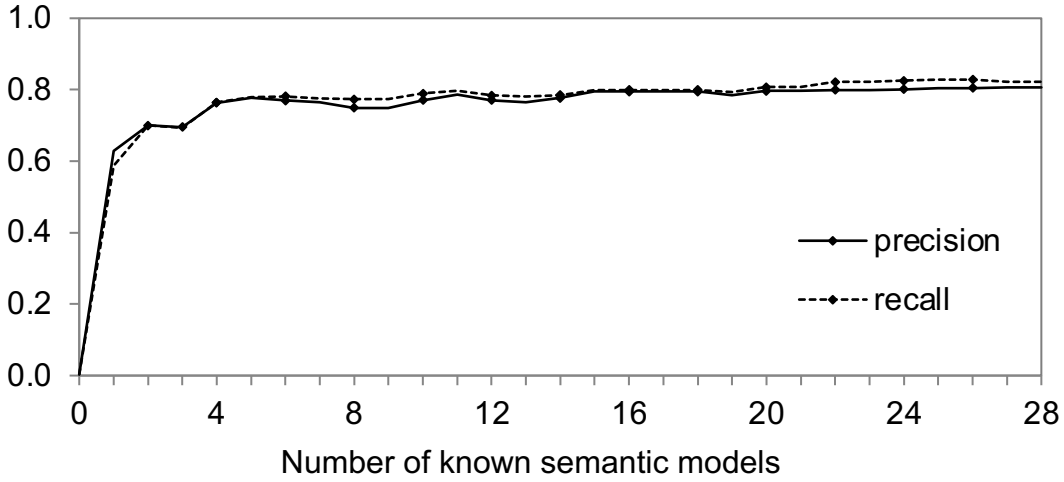
Experiment 1

correct semantic types are given



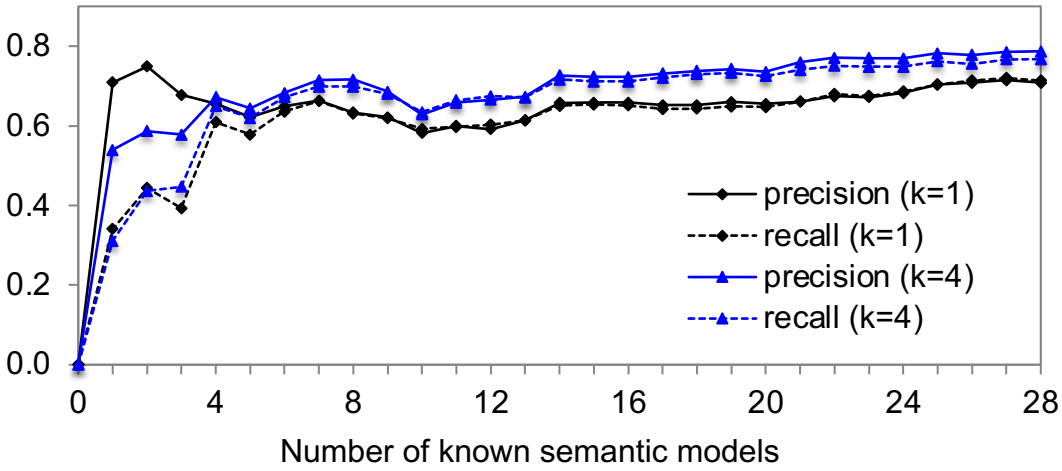
EDM

CRM



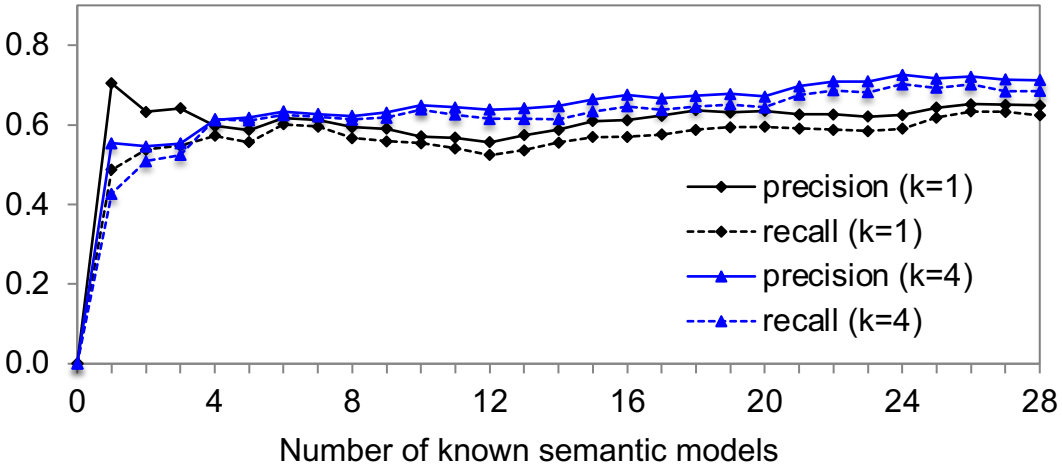
Experiment 2

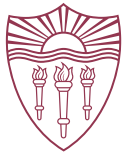
learn semantic types, pick top K candidates



EDM

CRM





Unsupervised Approach

Main Idea

- Information of entities in knowledge graphs can help source modeling

Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Table of Third Presidents

Eva in Wikidata

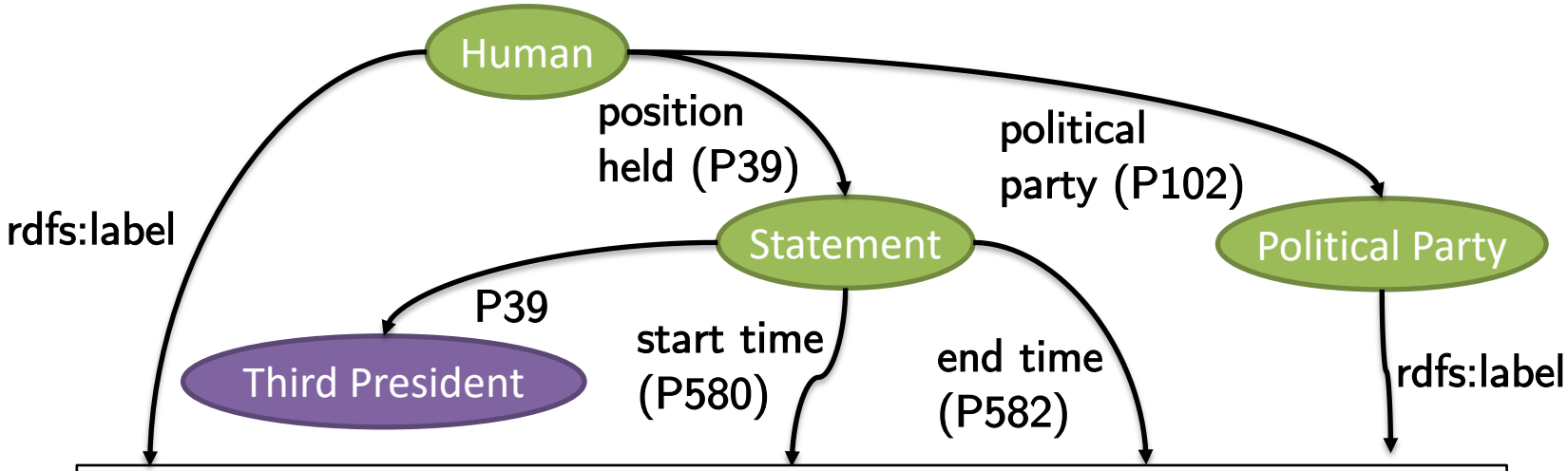
Eva Glawischnig-Piesczek (Q93870)

Austrian politician

edit

member of political party	Die Grünen
position held	Third President of the National Council of Austria
start time	30 October 2006
end time	28 October 2008

Challenges in Modeling Web Tables



Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Table of Third Presidents

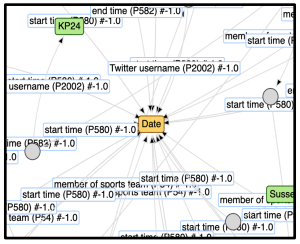
- The position (*third president*) is critical in understand the semantics of the table. However, it's in the table context.
- Need n-ary relationships (*position held*) to model the table

Approach

Drug	Main indication	Trade name
adalimumab	rheumatoid arthritis	Humira
apixaban	anticoagulant	Eliquis
lenalidomide	multiple myeloma	Revlimid
nivolumab	oncology	Opdivo

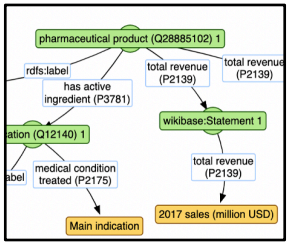
Linked table

Build graph of possible relationships in the table & context



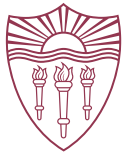
Semantic graph

Inference



Semantic Model

Approach



Inputs

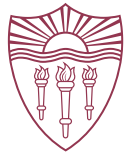
- A linked relational table $T(c_1, c_2, \ldots, c_n)$
- Contextual values of T: $C = \{v_1, v_2, \ldots\}$
- A target knowledge graph (e.g., Wikidata (WD))

1. **Construct semantic graph of relationships between columns and contextual values using KG**
2. Infer semantic model

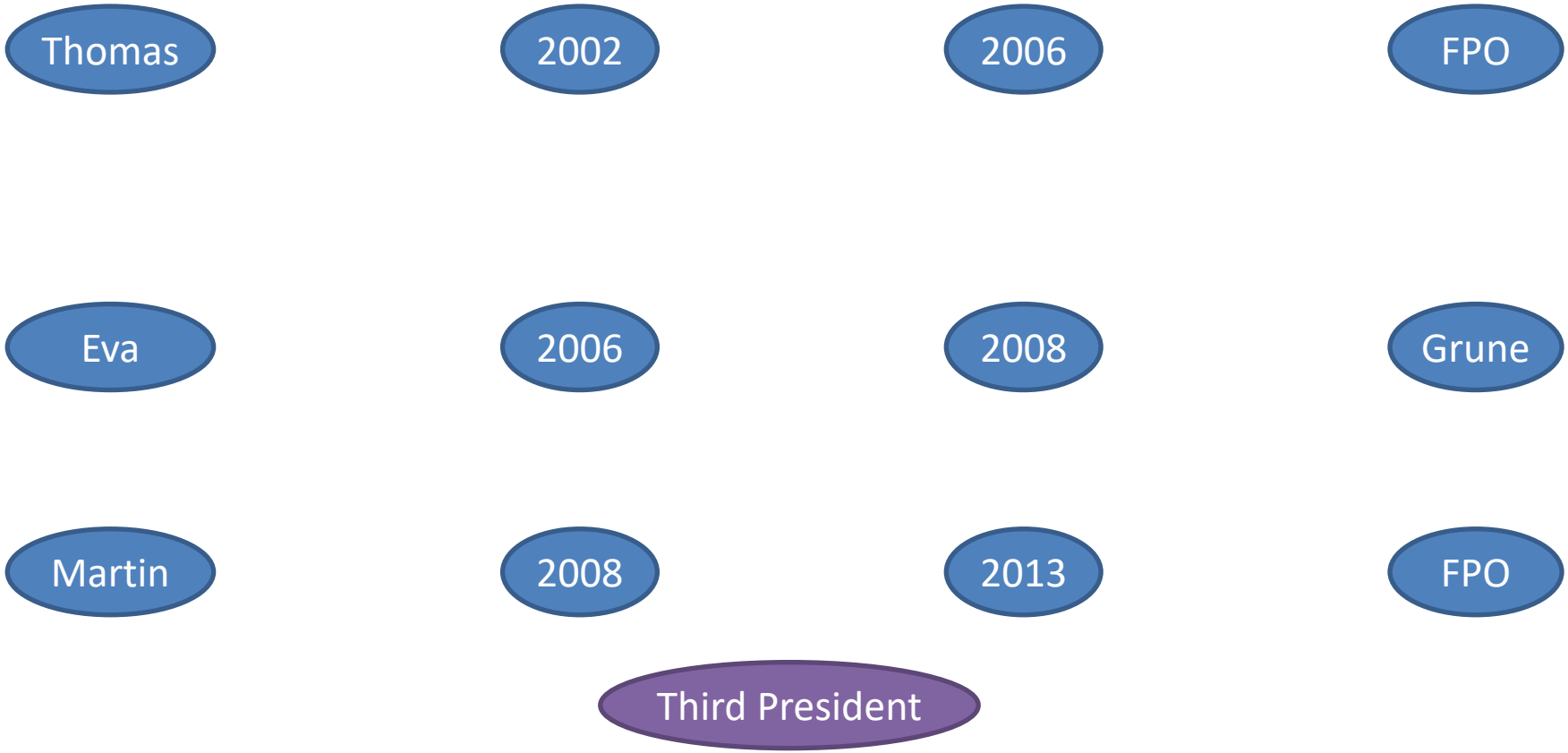
Outputs

- A semantic model of (T, C)

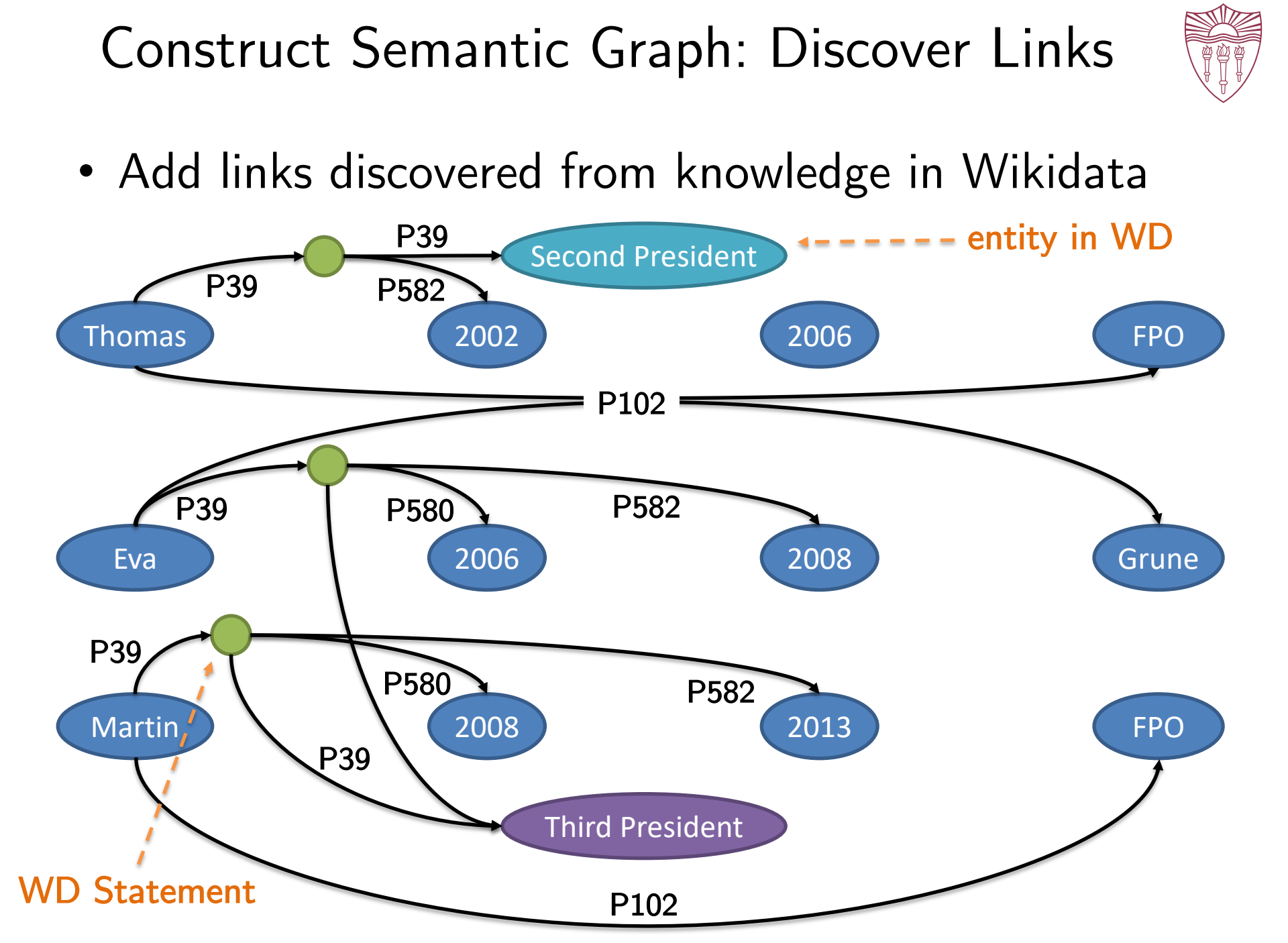
Construct Semantic Graph: Discover Links



- Create a graph of cells in T and C



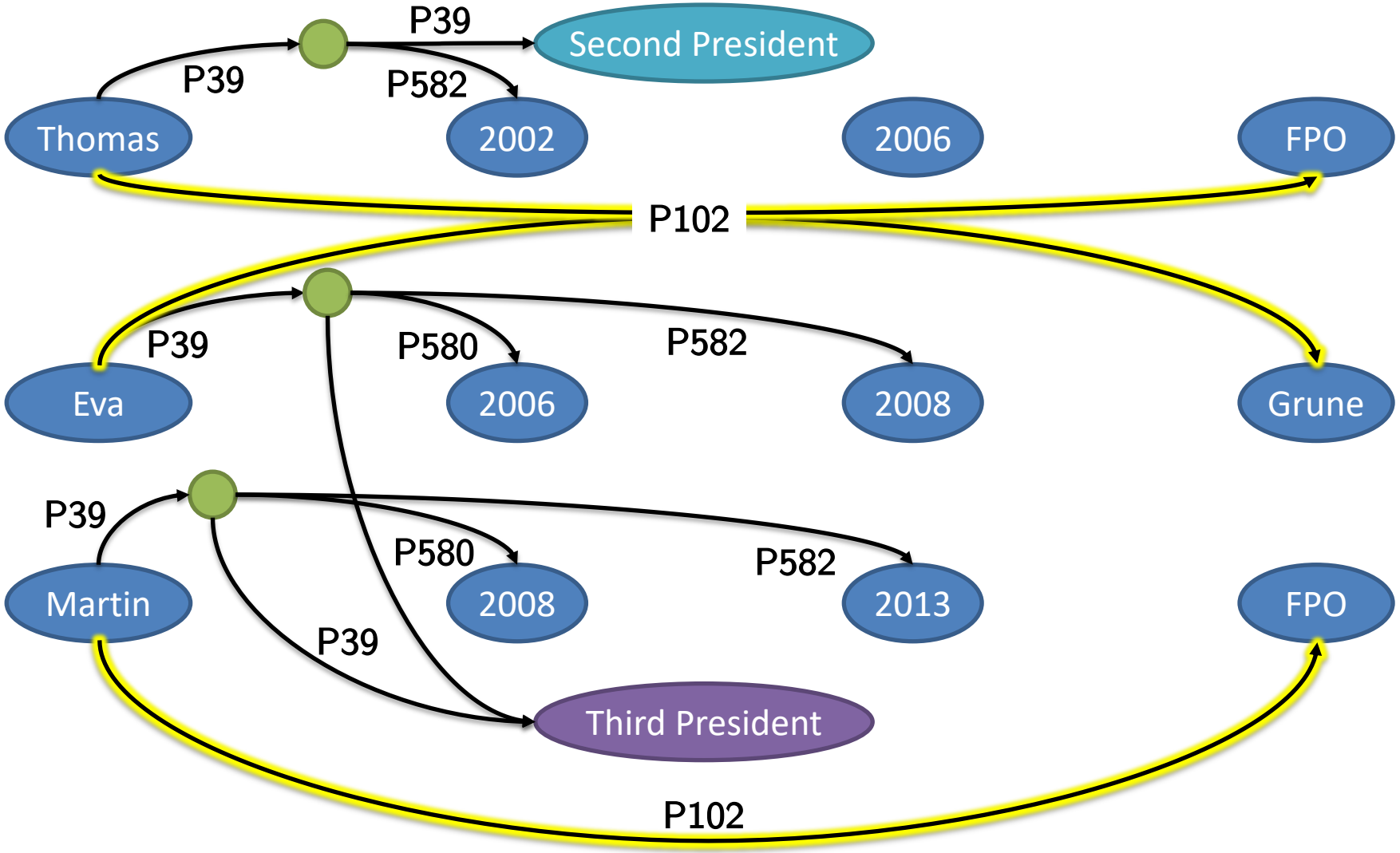
Construct Semantic Graph: Discover Links



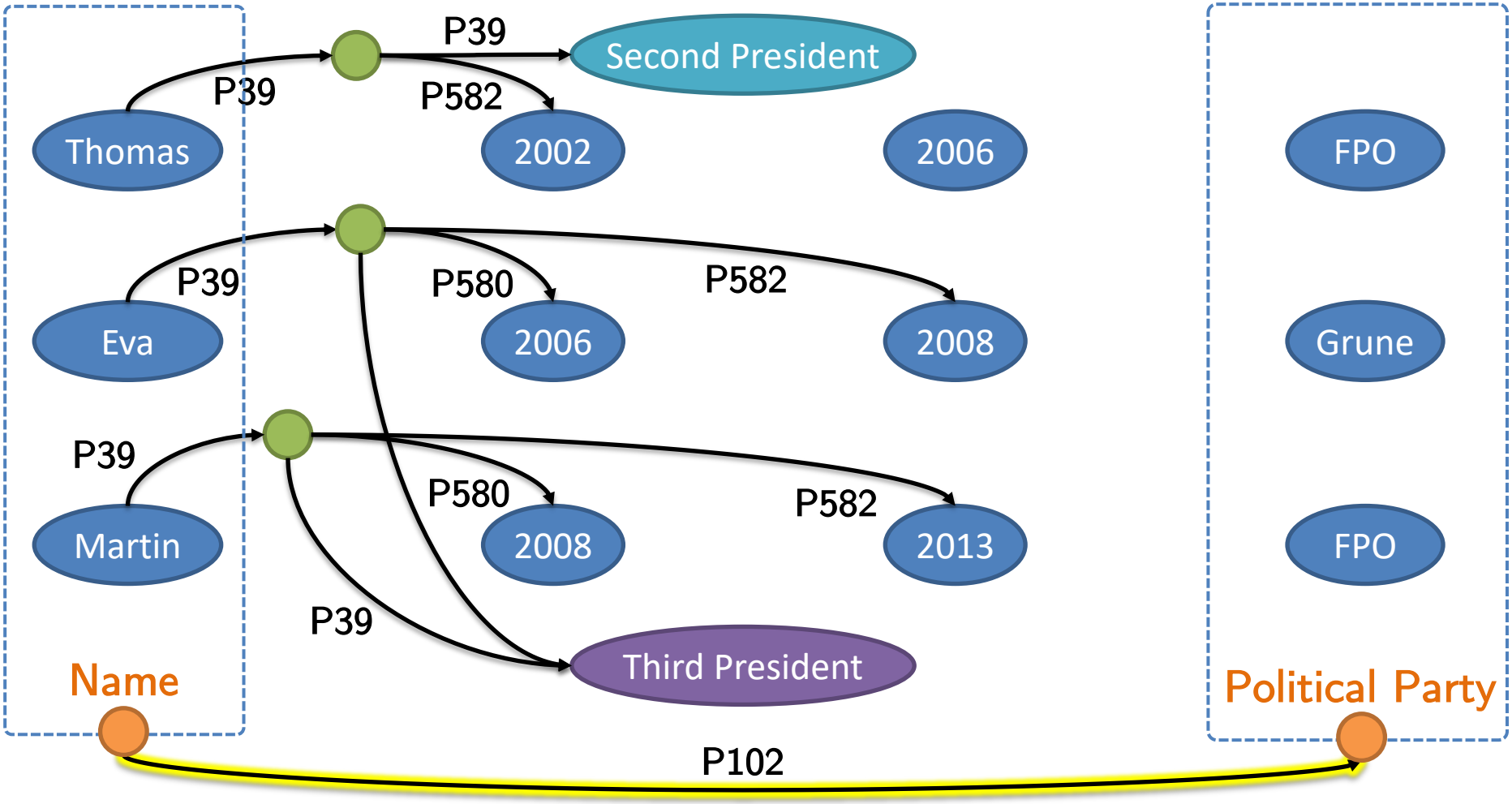
- Add links discovered from knowledge in Wikidata

Construct Semantic Graph: Summarization

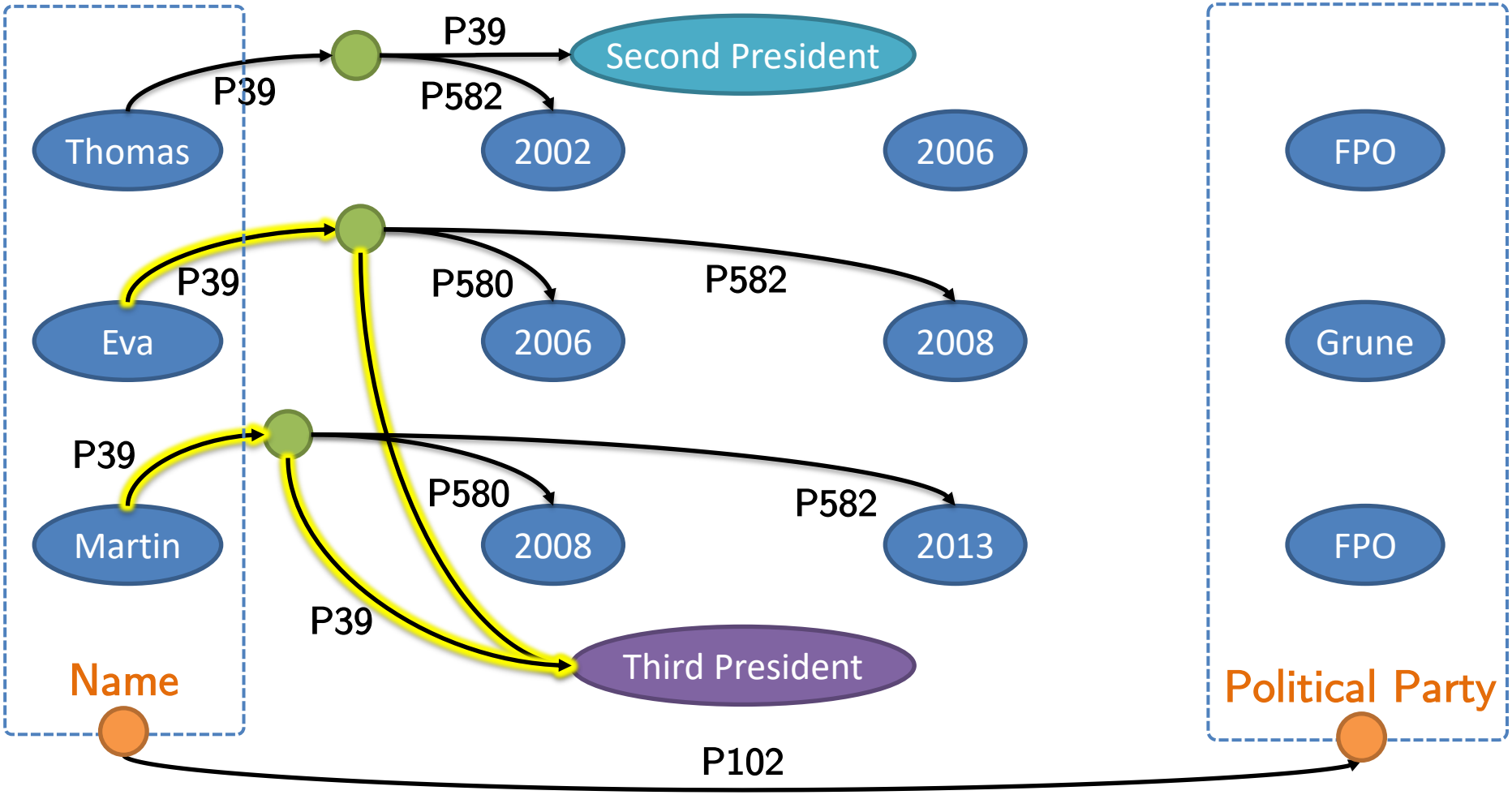
- Group links of same source & target columns/context



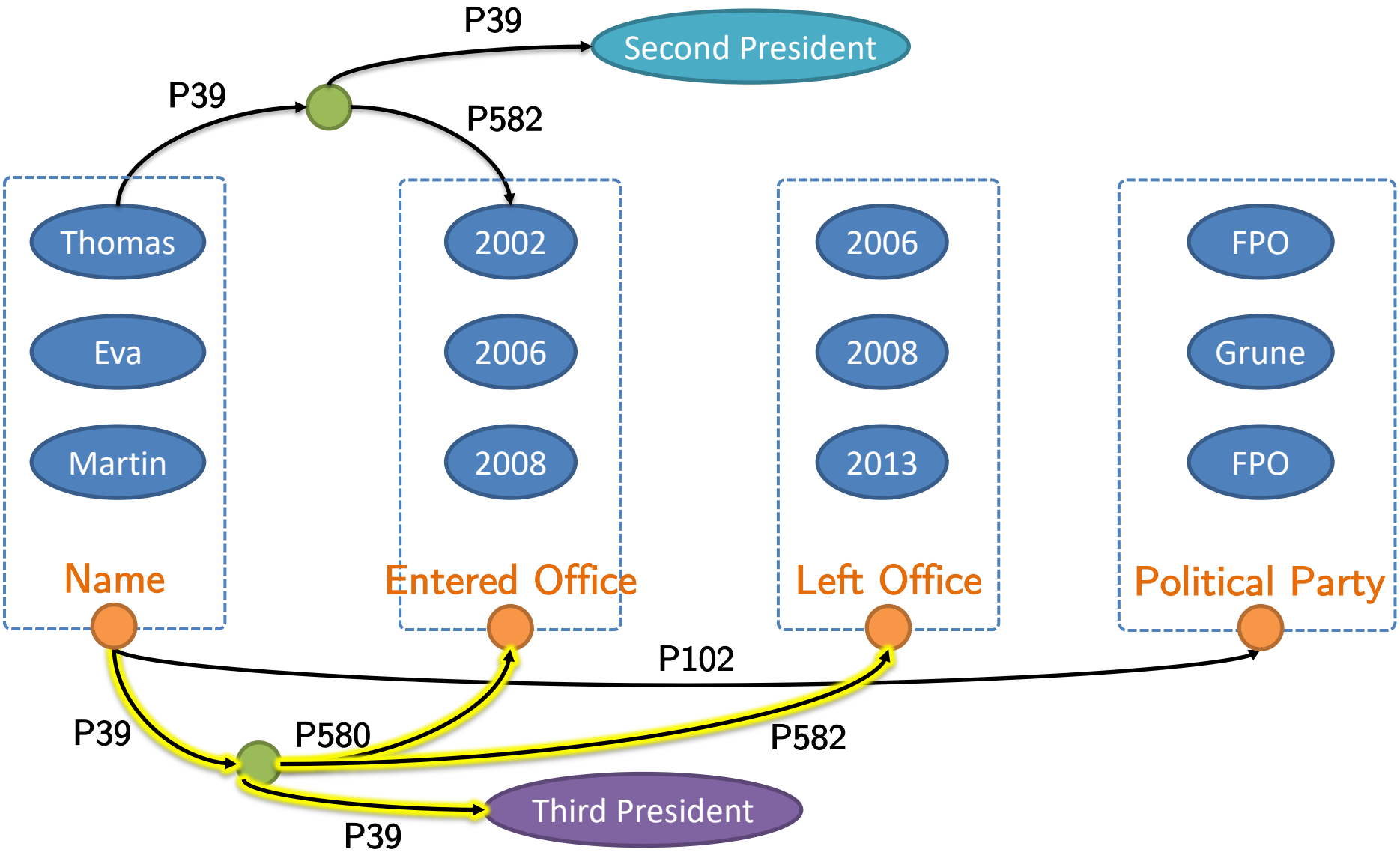
Construct Semantic Graph: Summarization



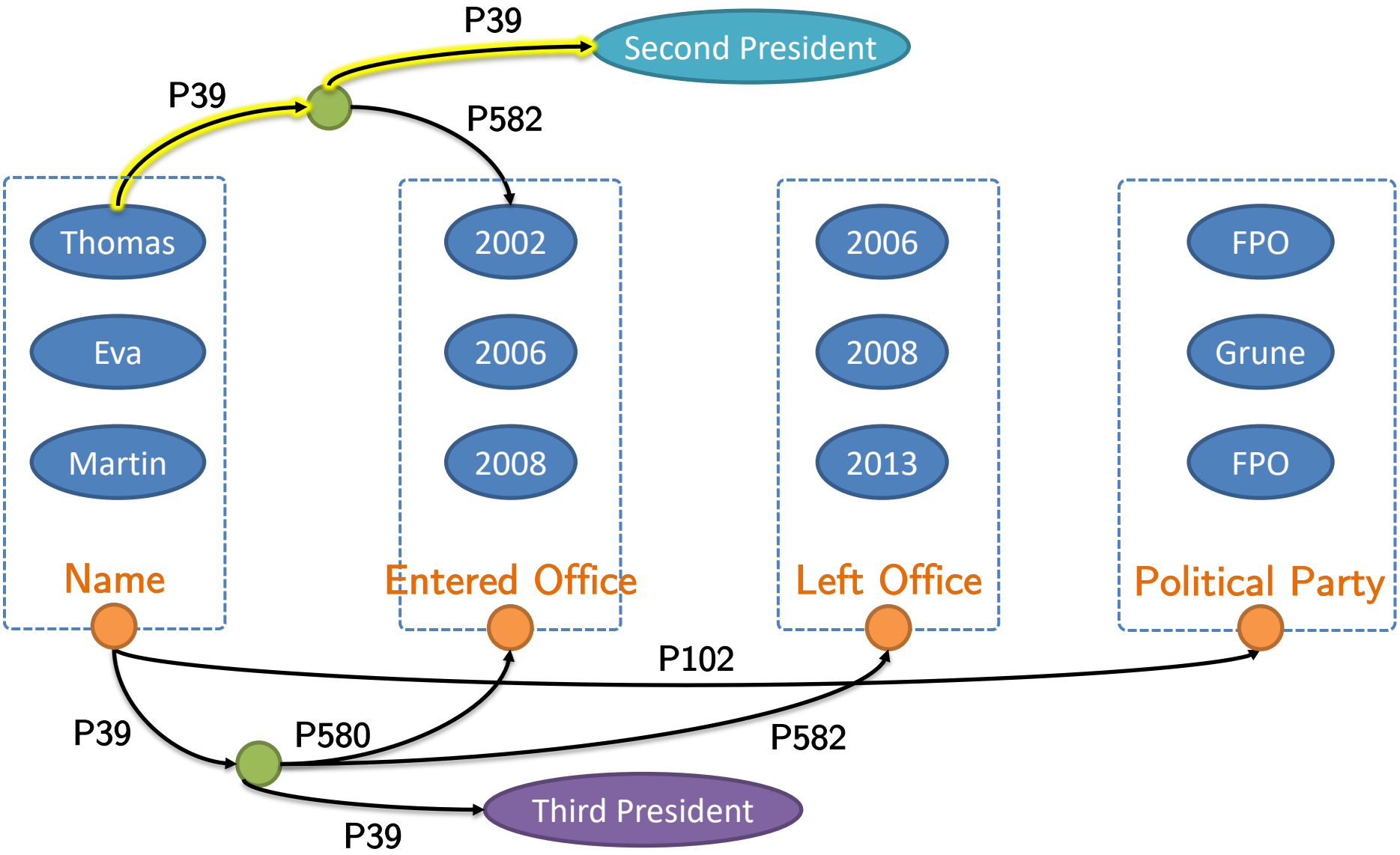
Construct Semantic Graph: Summarization



Construct Semantic Graph: Summarization

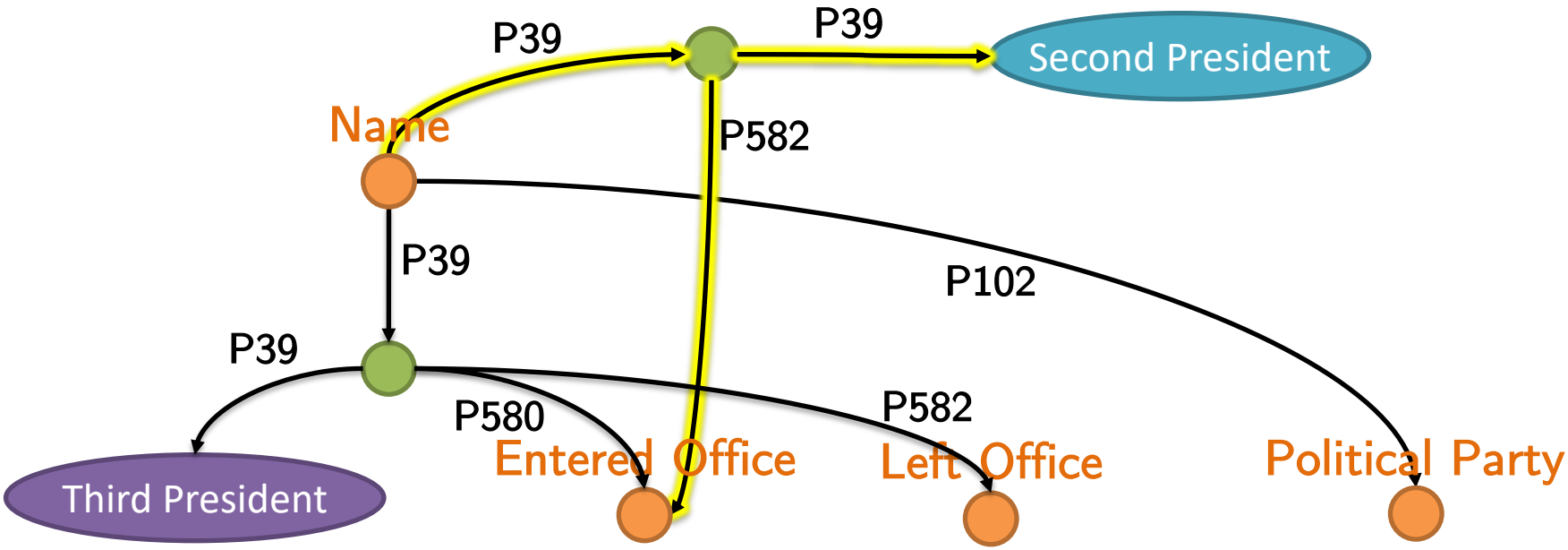


Construct Semantic Graph: Summarization

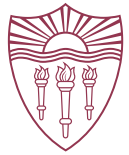


Construct Semantic Graph: Summarization

- Final semantic graph



Approach



Inputs

- A linked relational table $T(c_1, c_2, \ldots, c_n)$
- Contextual values of T: $C = \{v_1, v_2, \ldots\}$
- A target knowledge graph (e.g., Wikidata)

1. Construct semantic graph of relationships between columns and contextual values.
2. **Infer semantic model**

Outputs

- A semantic model of (T, C)

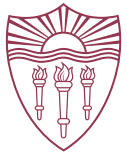
Infer Semantic Model

- Two tasks:
  1. Predict labels (true/false) of links in the semantic graph
  2. Predict columns' type

Classification methods consider each link/column separately and ignore structure of the graph

$\Longrightarrow$ Solve two tasks **collectively** using Probabilistic Soft Logic (PSL).

Infer Semantic Model: PSL Model



- Variable notation:
  - N is a node in the graph
  - P, Q are property or qualifier
  - T is an ontology class (column type)
- Rules apply to a single link/column

By default, all link/type are wrong

$\neg$ Rel(N_1, N_2, P)

$\neg$ Type(N, T)

If there are some features support/oppose the link/type, they will be correct/incorrect

CanRel(N_1, N_2, P) & PosFeat(N_1, N_2, P) $\rightarrow$ Rel(N_1, N_2, P)

CanRel(N_1, N_2, P) & NegFeat(N_1, N_2, P) $\rightarrow$ $\neg$ Rel(N_1, N_2, P)

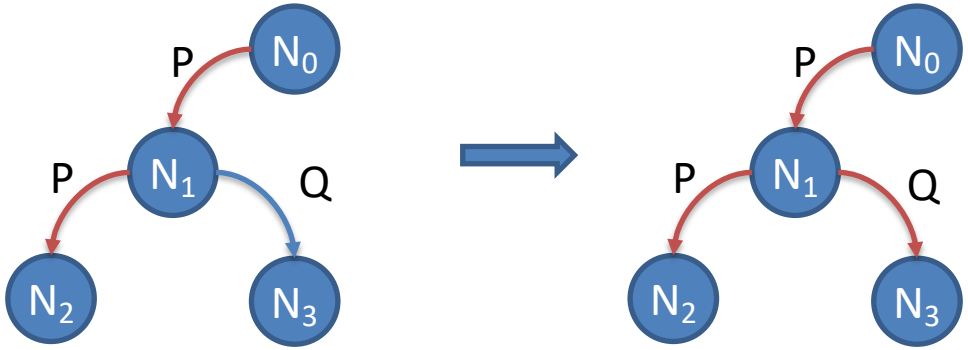
CanType(N_1, T) & PosTypeFeat(N_1, T) $\rightarrow$ Type(N_1, T)

Infer Semantic Model: PSL Model

- Rules apply to a group of links/columns

Incorrect property ⟹ incorrect qualifiers

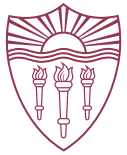
$$\text{CanRel}(N_1, N_2, P) \;\&\; \text{CanRel}(N_1, N_3, Q) \;\&\; \text{IsProp}(N_1, N_2, P) \;\&\; \text{IsQual}(N_1, N_3, Q) \;\&\; \neg\mathbf{Rel}(N_1, N_2, P) \rightarrow \neg\mathbf{Rel}(N_1, N_3, Q)$$



Prefer specific property than a general property

$$\text{CanRel}(N_1, N_2, P_1) \;\&\; \text{CanRel}(N_1, N_2, P_2) \;\&\; \mathbf{SubPropertyOf}(P_1, P_2) \;\&\; \mathbf{Rel}(N_1, N_2, P_1) \rightarrow \neg\mathbf{Rel}(N_1, N_2, P_2)$$

Infer Semantic Model: PSL Model



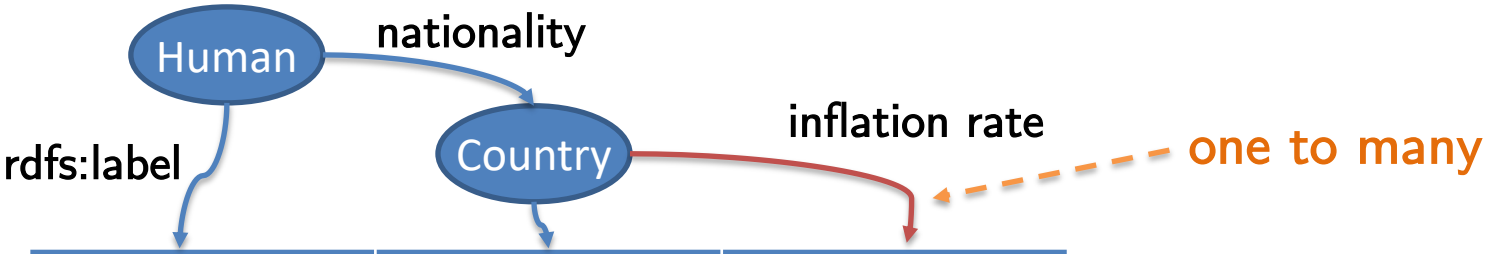
- Rules apply to a group of links/columns

Column type needs to compatible with range of the property

$$\text{CanRel}(N_1, N_2, P) \;\&\; \text{CanType}(N_2, T) \;\&\; \text{Rel}(N_1, N_2, P) \;\&\; \neg\,\textbf{Range}(P, T) \rightarrow \neg\textbf{Type}(N_2, T)$$

Properties' values of non-subject entity should be one to one correspondence with the entity

$$\text{CanRel}(N_1, N_2, P_1) \;\&\; \text{CanRel}(N_2, N_3, P_2) \;\&\; \text{Rel}(N_1, N_2, P) \;\&\; \textbf{OneToMany}(N_2, N_3) \rightarrow \neg\,\textbf{Rel}(N_2, N_3, P2)$$



Player	Country	To par
Gil Morgan	United States	+1
Terry Diehl	United States	+2

Evaluation

- Datasets
  - Wikipedia Tables: 250 tables
  - SemTab2020 Round 4: 2048 synthetic tables
- Tasks:
  - Column Relationship Annotation (CPA)
  - Column Type Annotation (CTA)
- Measurements: Precision, Recall, F1
- Evaluating systems: GRASD (our approach), MantisTable

Evaluation

Dataset	Method	CPA			CTA		
		Precision	Recall	F1	Precision	Recall	F1
Wikipedia Tables	MantisTable	0.534	0.440	0.483	0.885	0.306	0.454
	MantisTable Correct Subject	0.561	0.570	0.565	0.892	0.355	0.508
	GRASD	**0.744**	**0.657**	**0.699**	**0.826**	**0.820**	**0.823**
SemTab2020 Round4	MantisTable	0.987	0.982	0.984			
	GRASD	**0.988**	**0.990**	**0.989**			

Summary

	Supervised Approach	Unsupervised Approach
Pros	- Can choose target ontologies	- Need no/little training data - Easy to extend by adding rules
Cons	- Require training data	- Target ontology is fixed (KG ontology)