Linked Data and the Digital Humanities – A Match Made in Heaven or Hell?

Krzysztof Janowicz and Yingjie Hu

STKO Lab, University of California, Santa Barbara, USA

Department of Geography, University of Tennessee, Knoxville, USA

November 2017

Smart Data versus Smart Applications

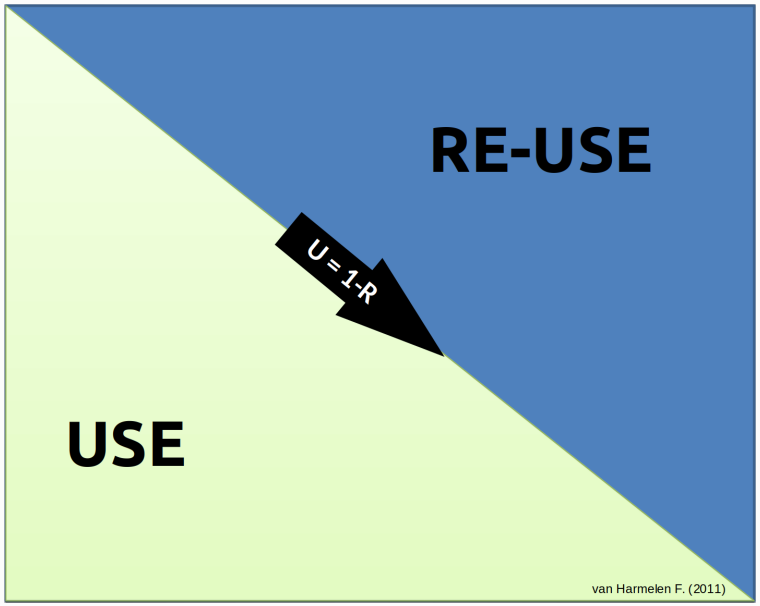
Key insight: smart data will make future applications more (re)usable, flexible, and robust, while smarter applications fail to improve data along the same dimensions. Put differently, applications are like old fish while data is like old **wine**.

Key challenge: most wines do not improve with age. In fact, bottling and storage is often more important than vintage.

Opportunistic/Unexpected Reuse

There are (at least) two major reasons to be concerned about **data aging**.

- Reuse of **own data**, e.g., for future work.
  - Main challenges: **data formats**, data carriers, carrier storage,...
- Reuse **by others**, e.g., for reproducibility.
  - Main challenges: **semantic aging**, provenance, usage restrictions/ rights,...[on top of the challenges above]



The current discussion around **data science** is still largely driven by data size, new tools, and new computational methods, while a significant part of this paradigm shift stems from our changing **reuse** of data. Today, data are often conceptualized, collected, cleaned, and published **by others**.

This opportunistic/**unexpected reuse** from **heterogeneous** sources is a major opportunity and challenge at the same time.

So how to make Data More Reusable?

- Use openly and freely **standardized** formats
- Make data **human & machine-readable**
- Break up the **data-metadata** distinction, i.e., make data describe themselves
- Provide the **application logic** together with the data
- Make the **meaning** of the used terms and procedures **explicit**

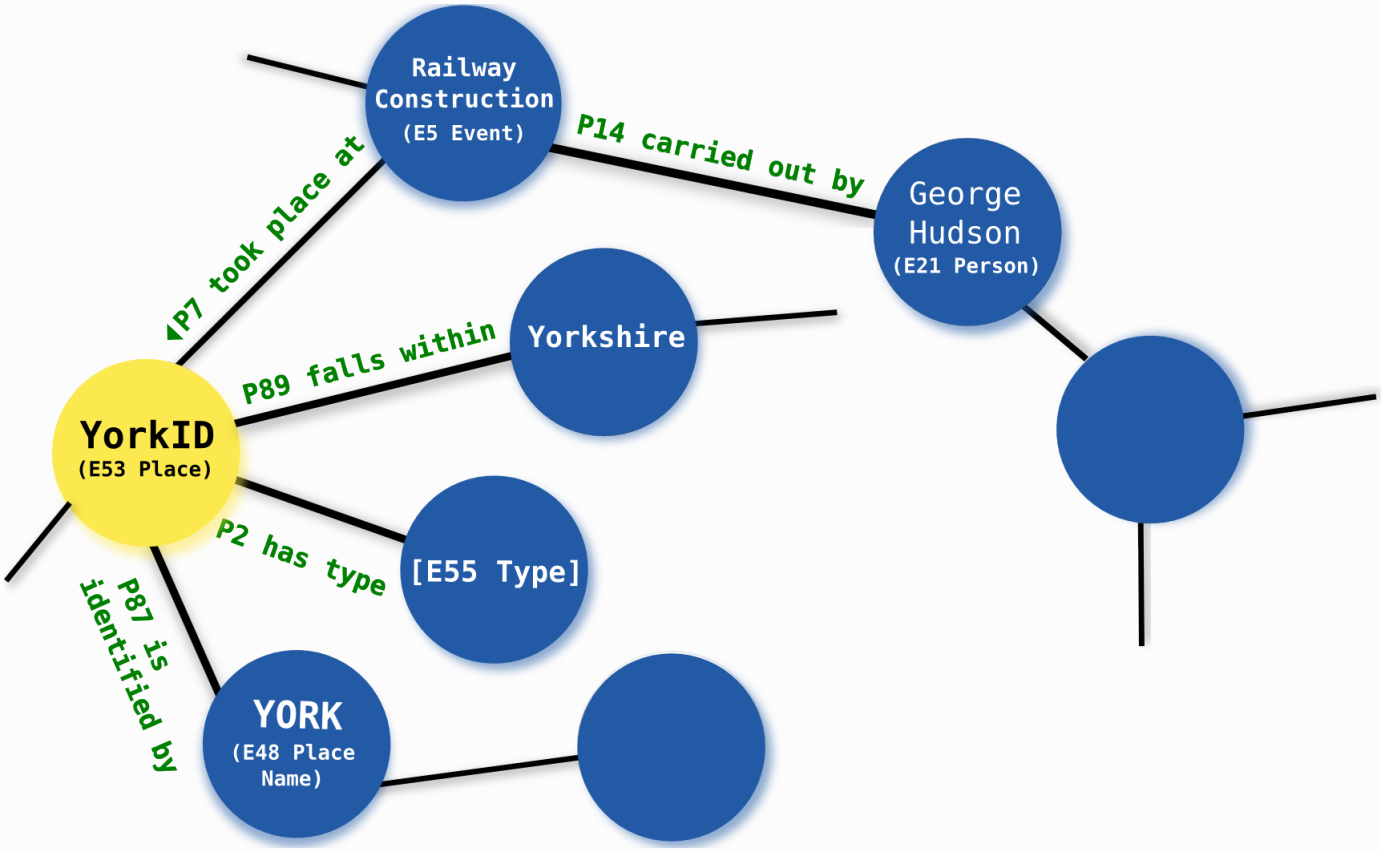
We will outline how one could accomplish these steps using the so-called **Linked Data paradigm** by showing examples centered around **places** and **trajectories**.

Places ...

- are **abstract** (social) entities used to structure knowledge and to ease communication
- are of interest to individual **communities** in a certain **region** and for a particular **time**
- can be referred to by **various names** (toponyms), narrative descriptions, or even by placeholders such as *Anyshire*
- are **categorized** according to some **commonly** agreed upon characteristics
- often **refer to** a physical region in space
- can be modeled as relations such that an object can become a **location** for another one.
- can **split, move, merged, or disappear** over time.
- In most modern gazetteers, places have **names**, **types**, and **footprints**

Clear **separation of the physical region and the social construct** called place. We experience places via a number of **perceivable characteristics** of the physical region they refer to.

Modeling knowledge about a Place (York, UK)

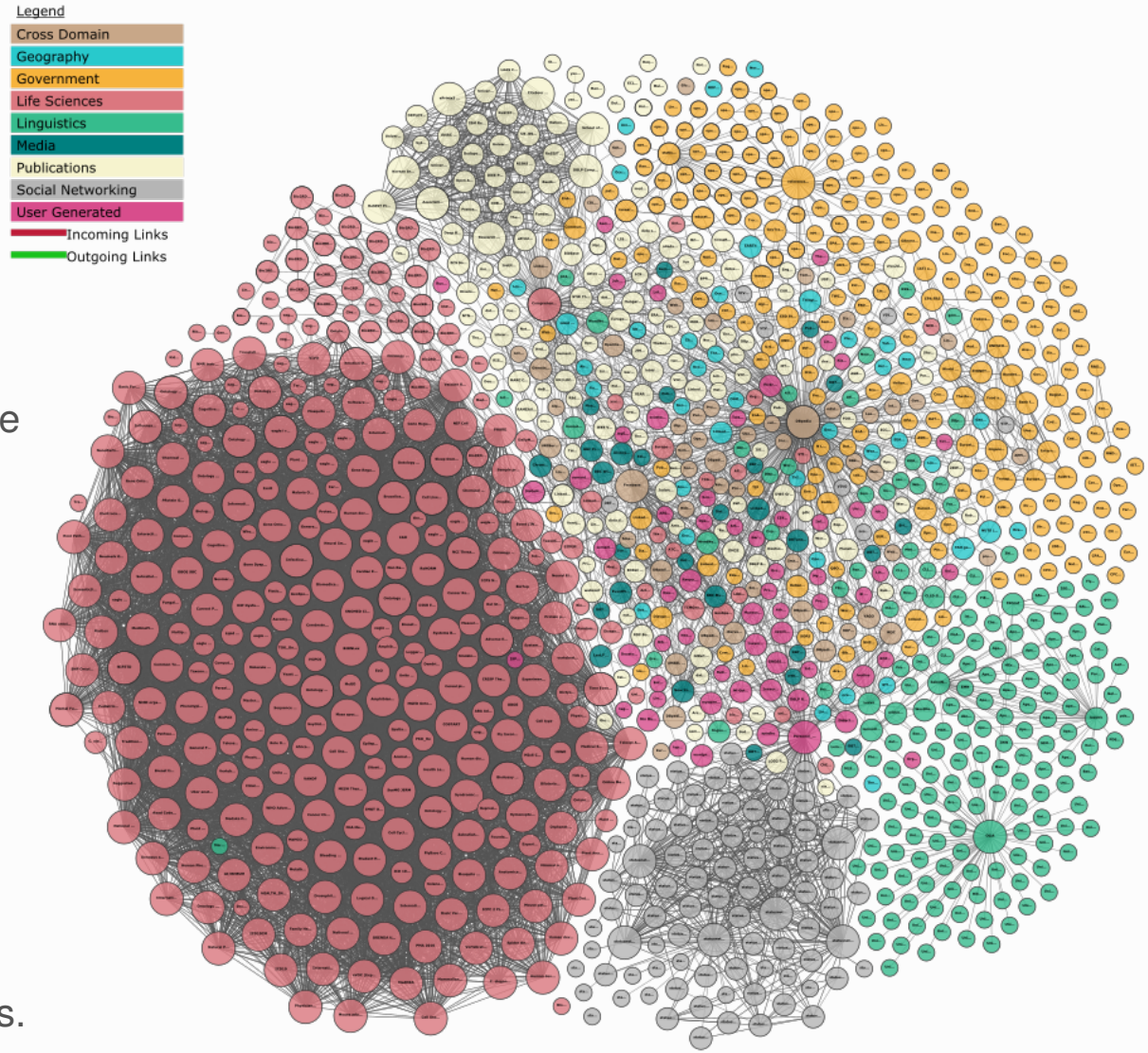


- *Place* as a **nexus** connecting events, objects, and actors.
- What do we **know** about these events, objects, and actors?
- How does York, UK **relate** to the Scandinavian city of Yorvik?

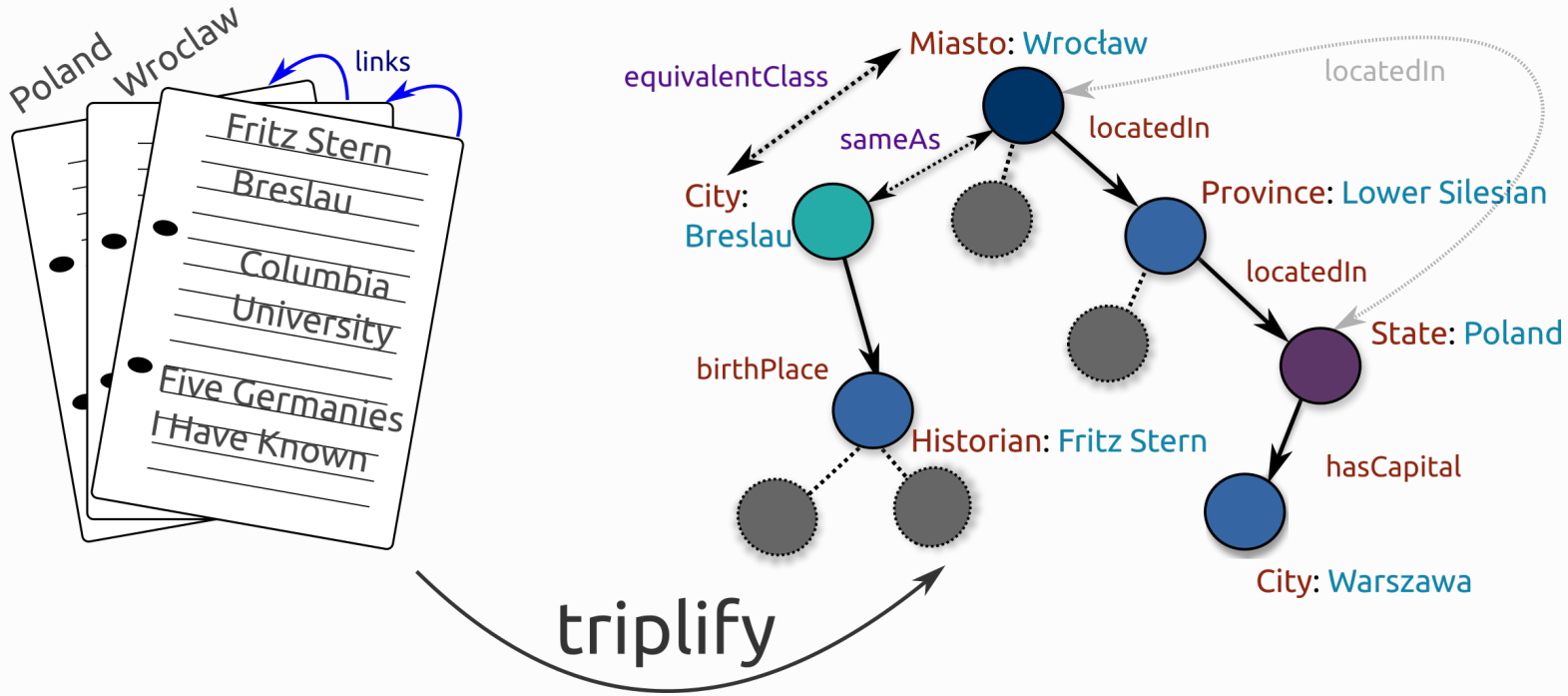
Linked Data: A Densely Interconnected Global Knowledge Graph

The publicly available part of the Linked Data cloud contains approximately **150 billion** triples distributed over **10000** diverse datasets and connected to each other by millions of links. The private part contributes to Google's new search engine, Apple's Siri, IBM's Watson, etc.

Many of the most **interconnected** hubs are **geographic data** hubs or cross-domain hubs containing millions of **places**. Also note that geo-hubs do not form their own cluster but sit between other clusters.

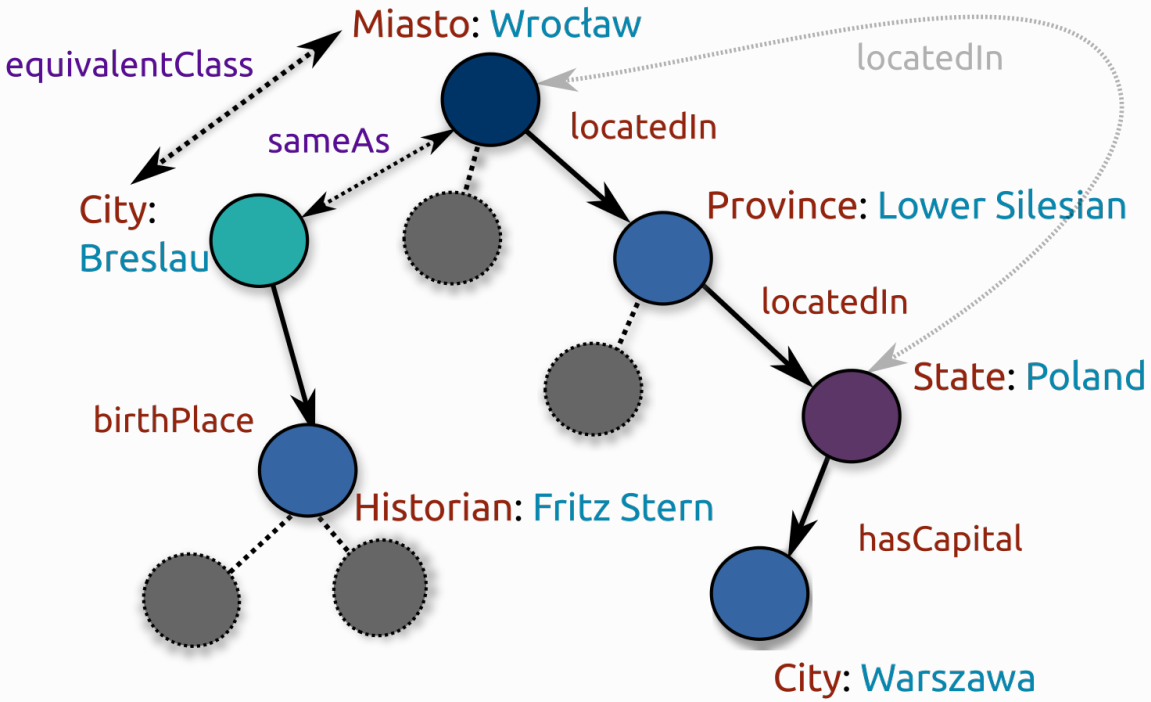


From Linked Documents to Linked Data



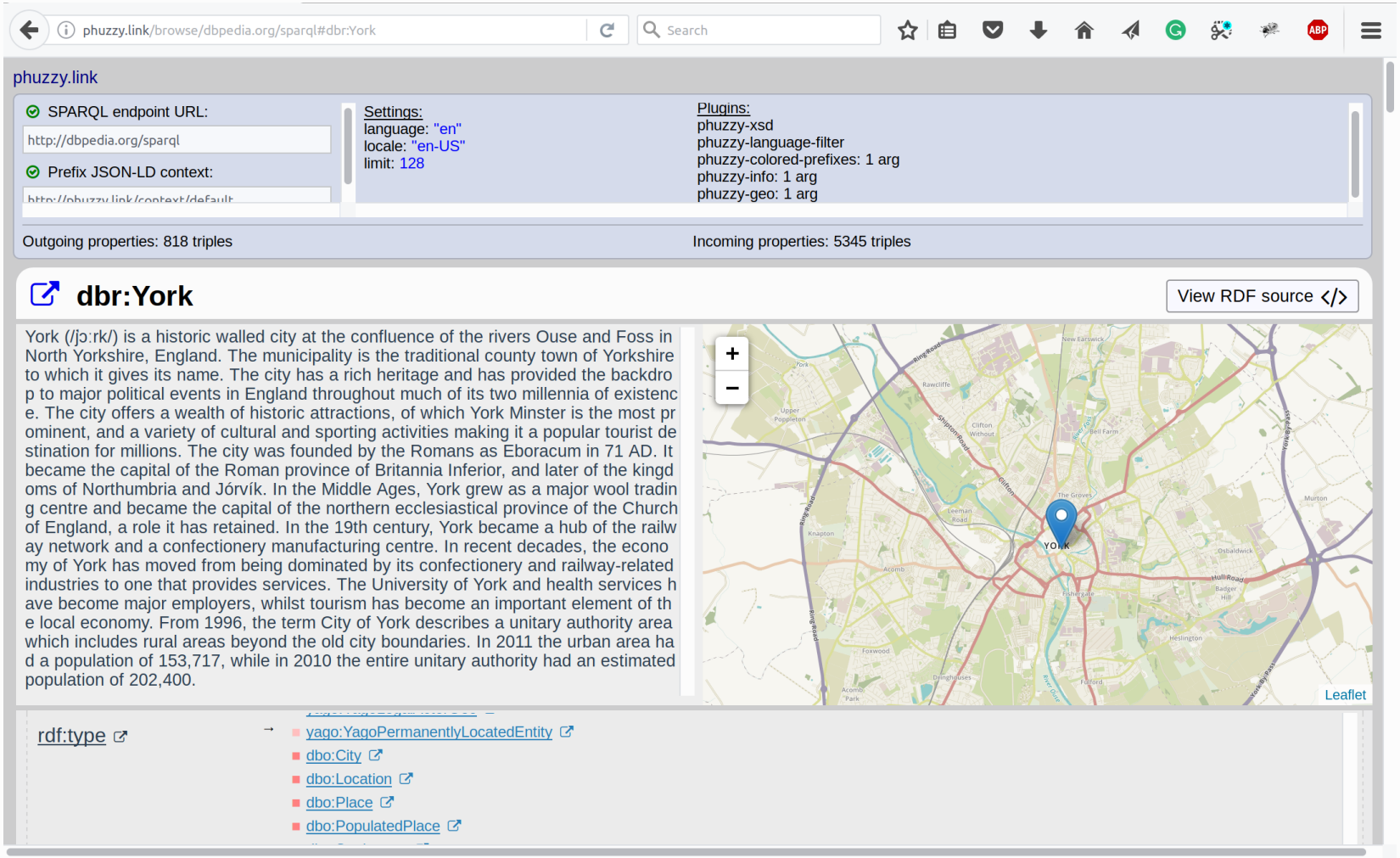
Use Uniform Resource Identifiers (**URI**) to identify **entities**, **link** them to other entities, encode information about these entities using the human and **machine-understandable RDF**, and make them available on the **Web**.

From Linked Documents to Linked Data



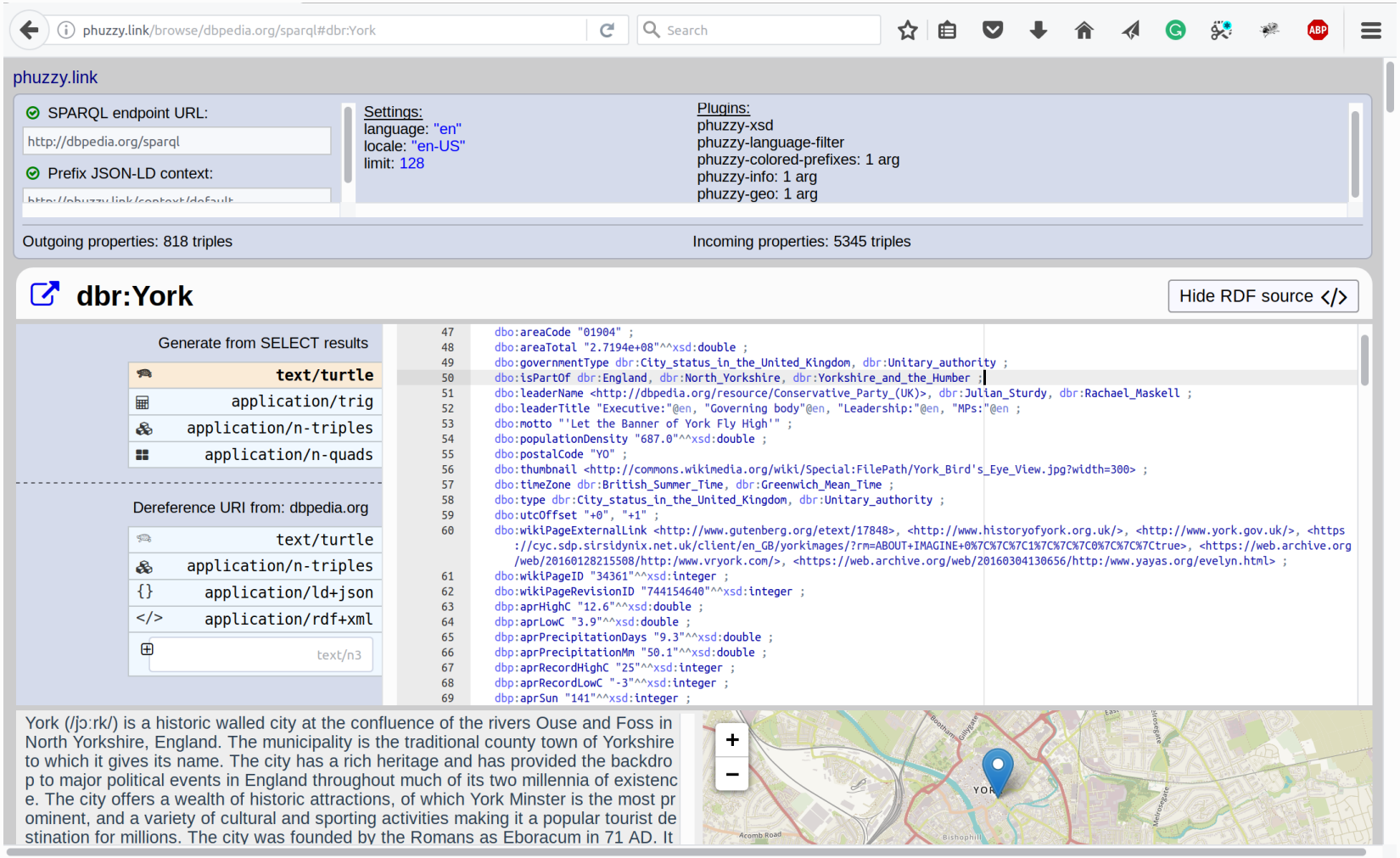
- **Knowledge as a graph** of globally unique, dereferenceable resources
- **No distinction** between data and metadata.
- **Ontologies** describe what terms such as *City* and *locatedIn* mean
- **Formal semantics** supports inferences, e.g., *equivalentClass* and *sameAs*

Human and Machine Readable and *Reason*-able



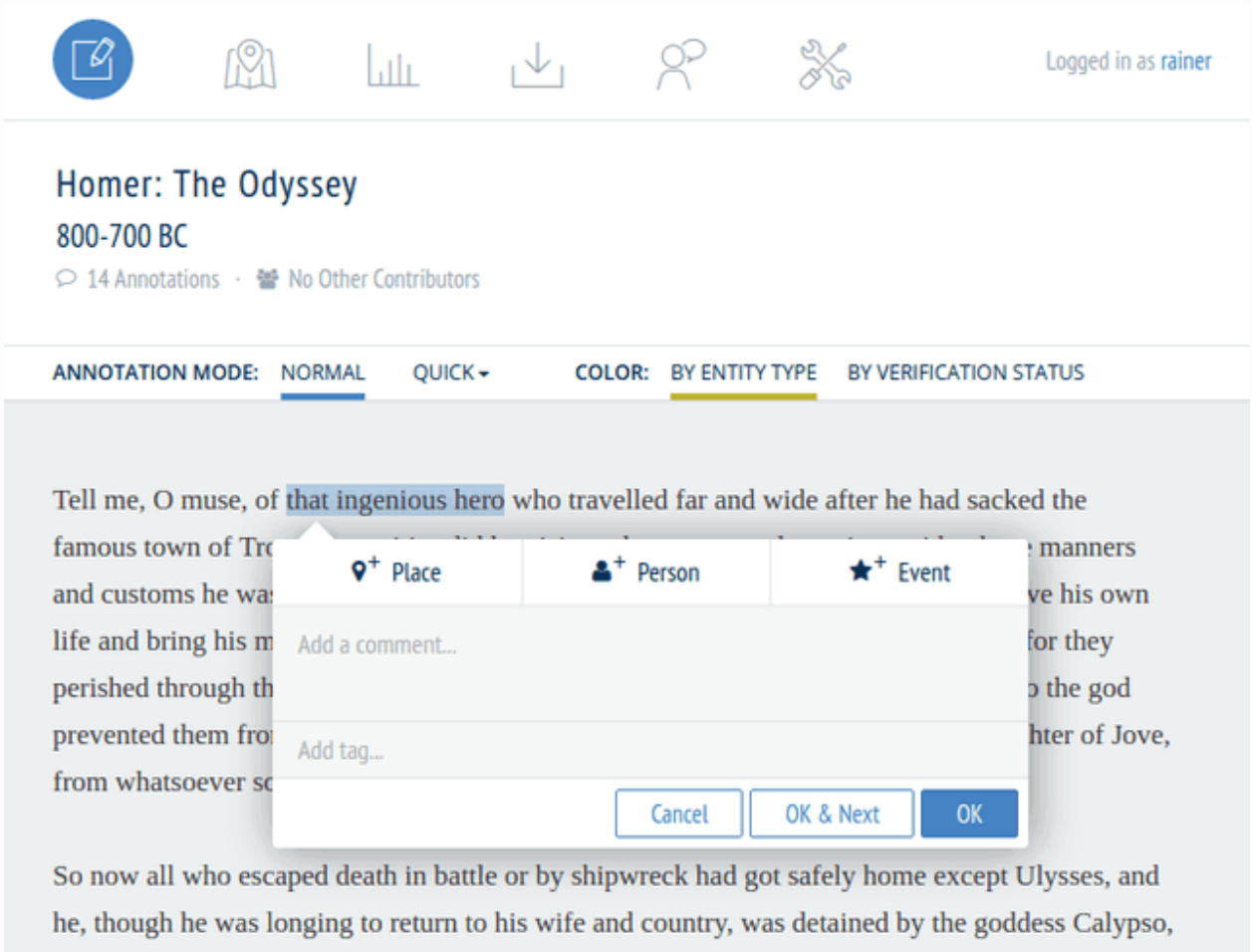
Note that the **place types** come from different **ontologies**

Human and Machine Readable and *Reason*-able



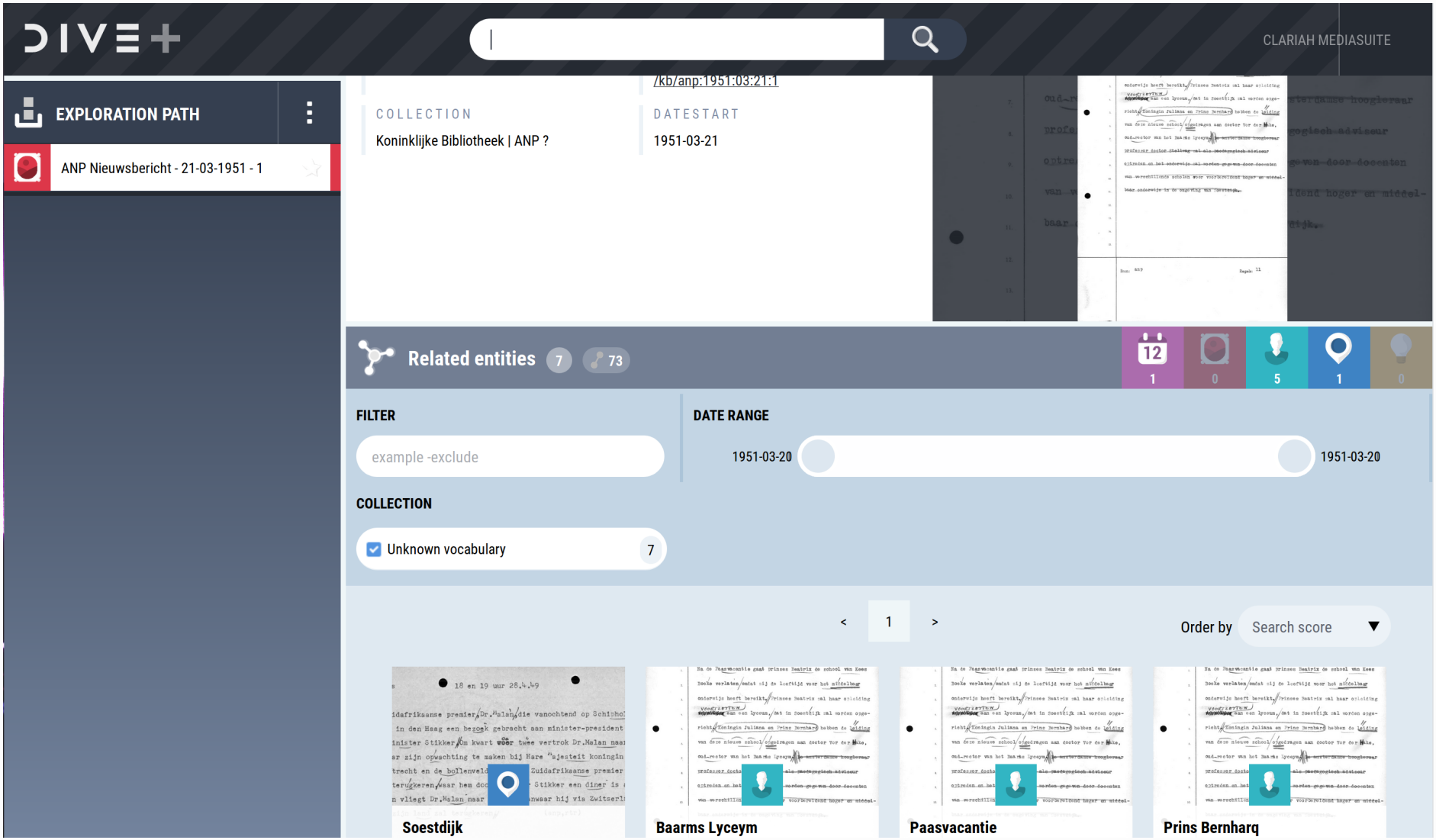
Note that even **population** and average temperature data are available as triples

Digital Humanities Heaven: Recogito by Pelagios



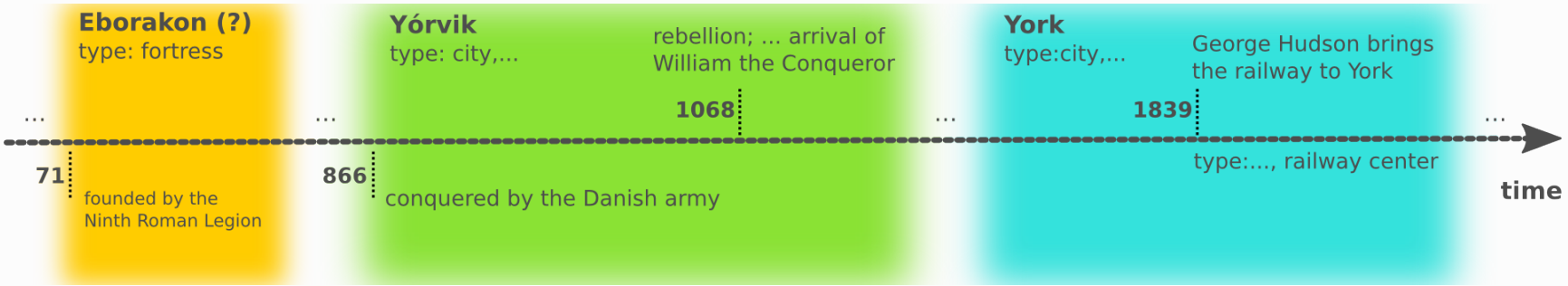
Annotate text and historic maps using places, events, objects, and so forth using global identifiers (**URIs**) to identify them **across different** texts/maps.

Digital Humanities Heaven: DIVE+



Temporal Scoping Hell

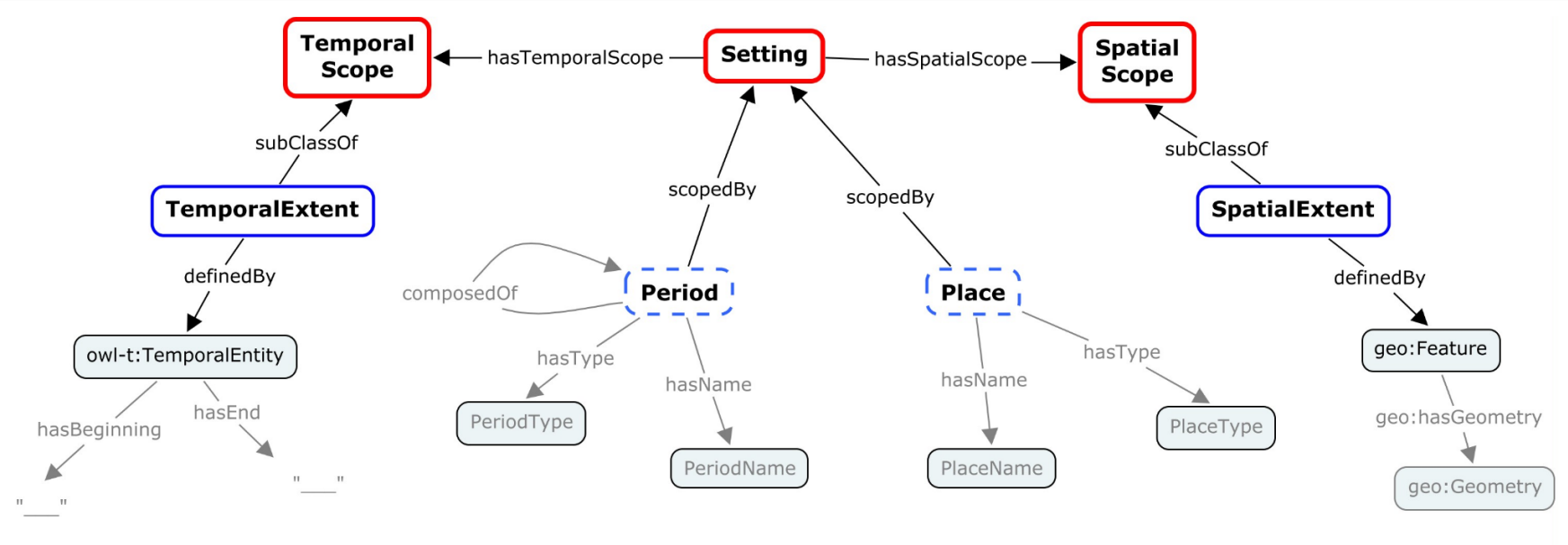
- Today, many **gazetteers** are available as Linked Data
- A **time-line** view of the York, UK



- *Place* as its **history** (and therefore as perdurant).
- York is not a **Roman fortress (T)** anymore, and it does not go by the **name** Jórvík **(N)** anymore. It is also not **part (F)** of the Kingdom of Northumbria anymore.
- Should there by **one URI** or many for the 'same' place?

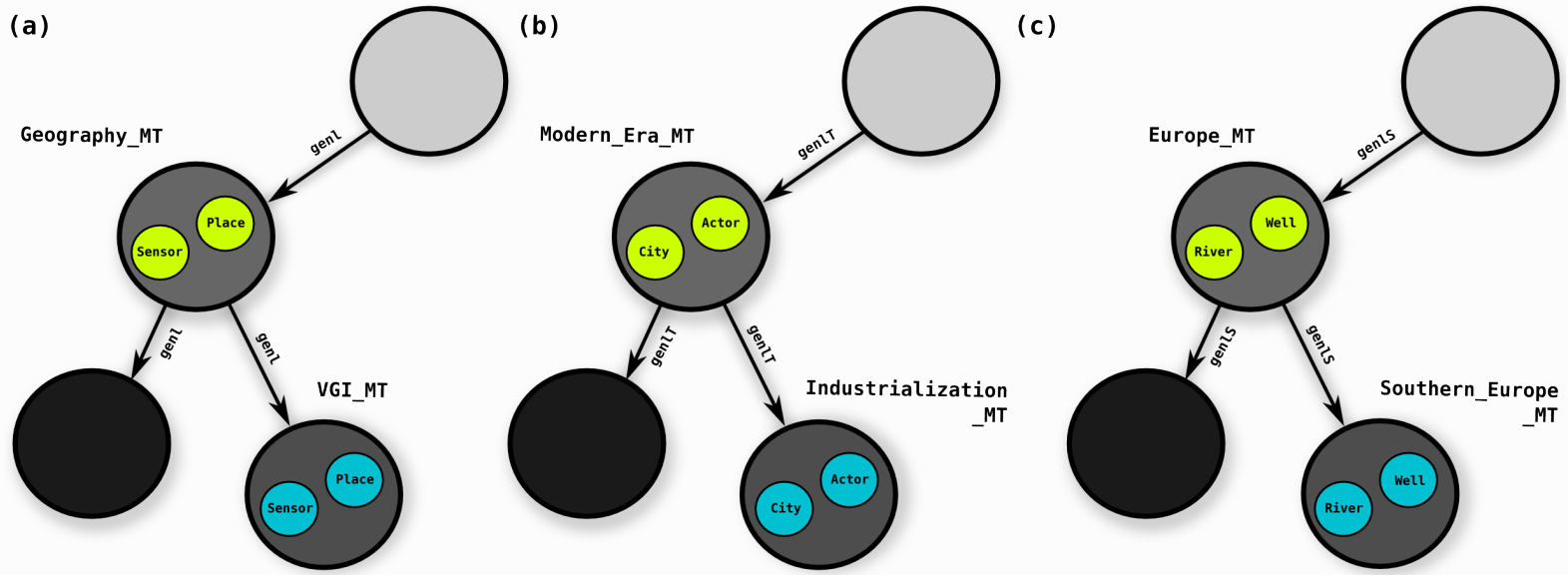
- Technical Linked Data solutions include **reification**, named graphs, blank nodes

Place, Period, and Setting for Linked Data Gazetteers



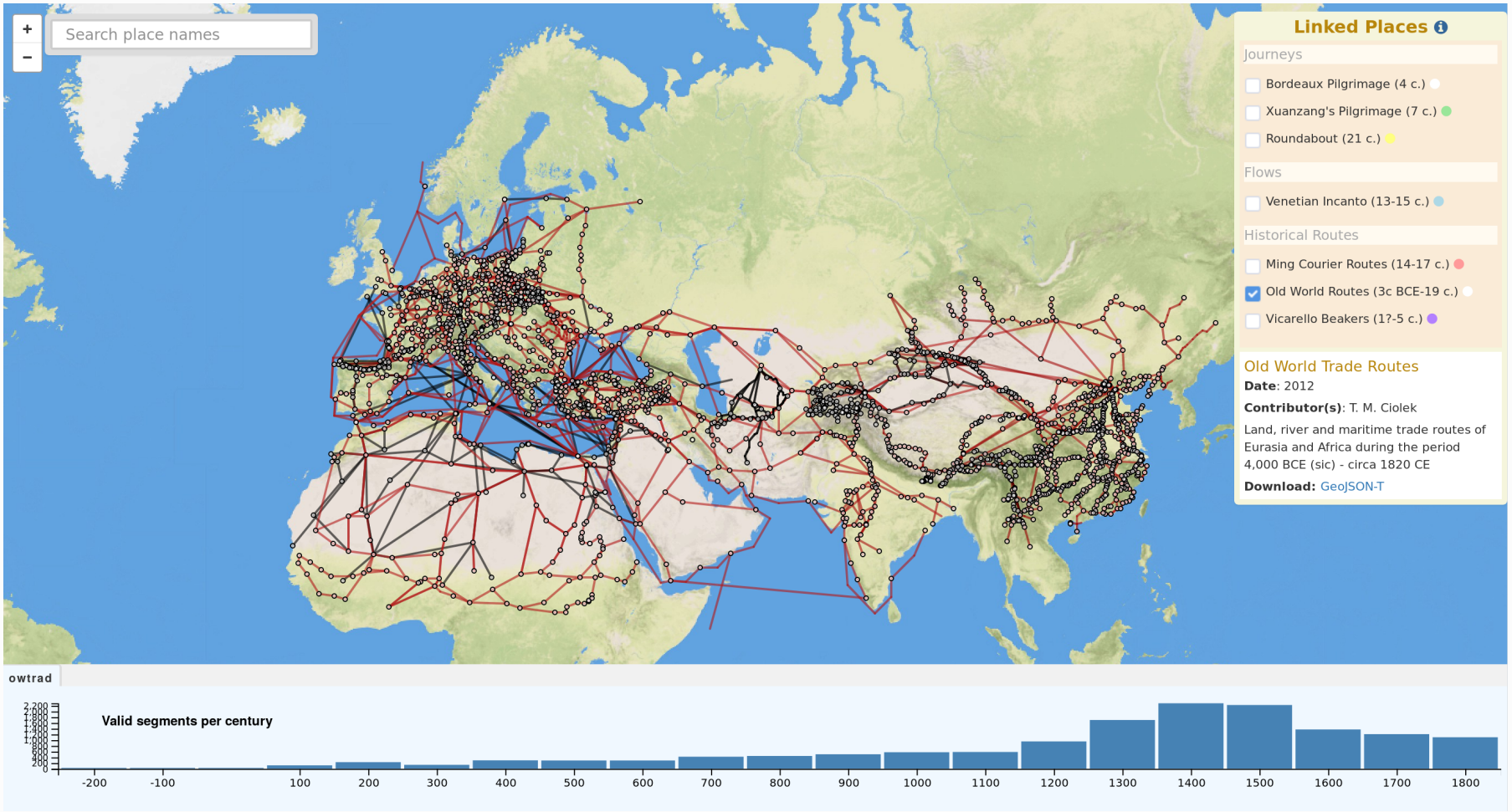
A modeling solution: **places** are also **temporal** and **events/periods** are also **spatial**, e.g., Levant Bronze Age and the Age of Enlightenment.

Digital Humanities Hell: Scoping Terminology by Space and Time



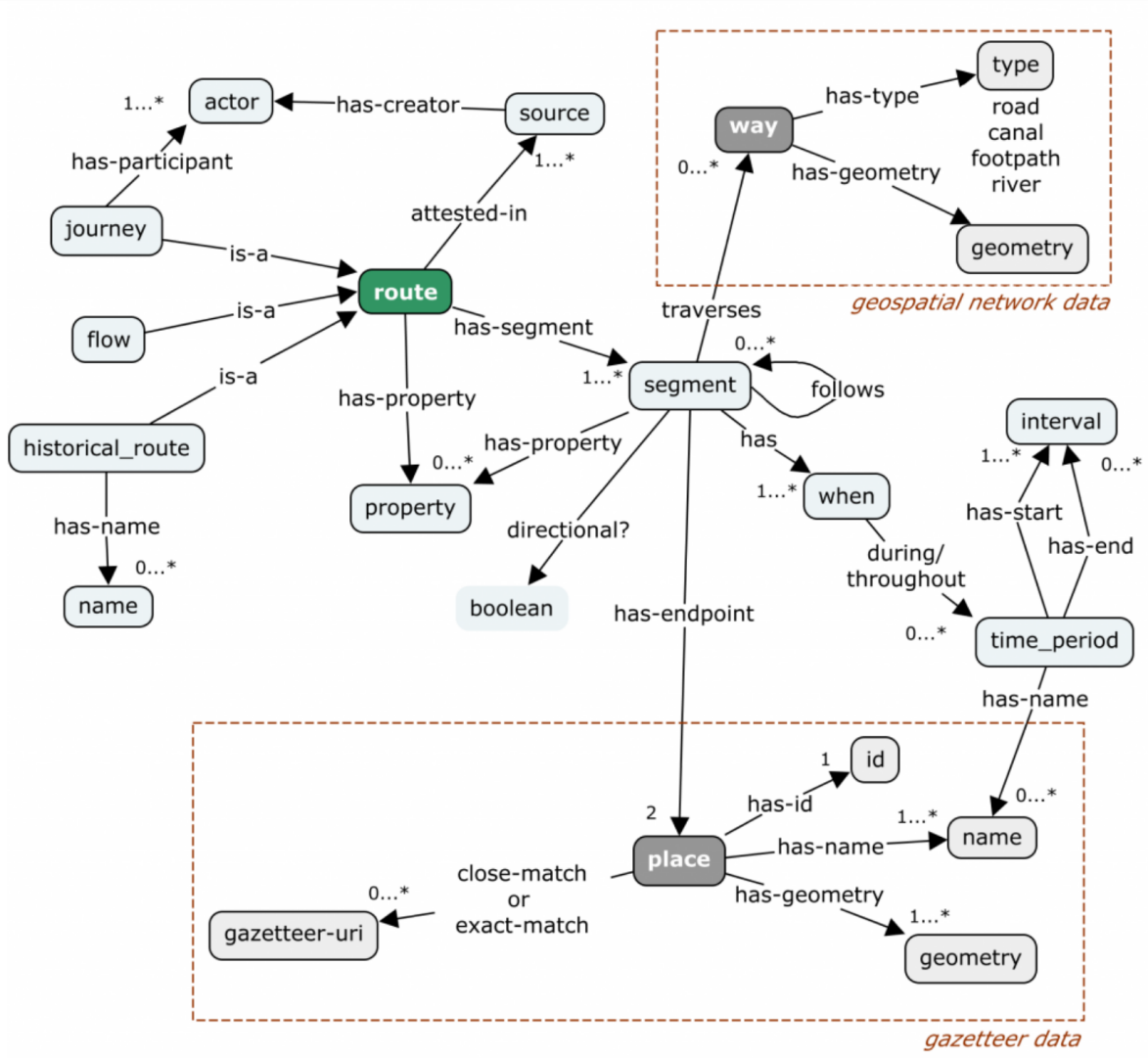
- Different domains, e.g., historic periods, require **different conceptualizations**
- 'Is one man's terrorist another man's freedom fighter?'
- **Local and consistent** theories structured by subsumption, space, and time within a **global and inconsistent** theory.

From Individual Places to Routes, Flows, and Pilgrimages



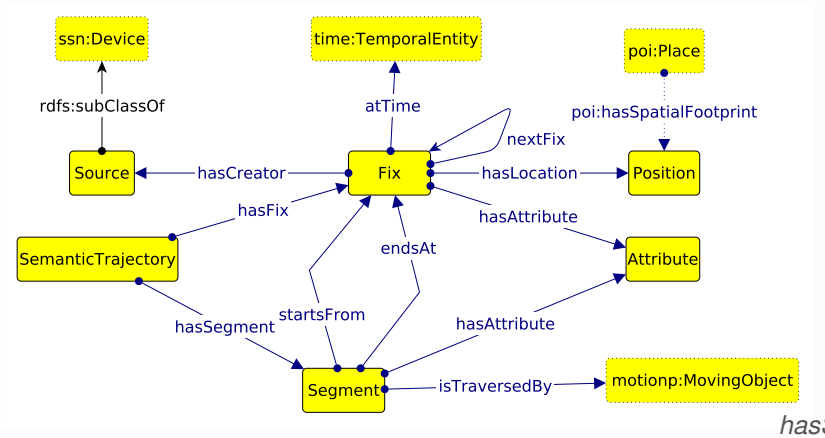
Is there a **common vocabulary** (ontology) for describing trajectories of goods, people, vessels, wildlife?

A Conceptual Model of Historical Geographical Movement



Early draft by Karl Grossner.

A Common Core Ontology for Trajectories Reusable Across Domains



$$Fix \sqsubseteq \exists atTime.time{:}TemporalEntity \sqcap \exists hasLocation.Position \sqcap \exists hasFix^{-}.SemanticTrajectory \quad (1)$$

$$Segment \sqsubseteq \exists startsFrom.Fix \sqcap \exists endsAt.Fix \quad (2)$$

$$\top \sqsubseteq \leq 1 startsFrom.\top \quad (3)$$

$$\top \sqsubseteq \leq 1 endsAt.\top \quad (4)$$

$$Segment \sqsubseteq \exists hasSegment^{-}.SemanticTrajectory \quad (5)$$

$$hasNext \sqsubseteq hasSuccessor \quad (6)$$

$$hasNext^{-} \sqsubseteq hasPrevious \quad (7)$$

$$hasSuccessor^{-} \sqsubseteq hasPredecesor \quad (8)$$

$$startsFrom^{-} \circ endsAt \sqsubseteq hasNext \quad (9)$$

$$hasSuccessor \circ hasSuccessor \sqsubseteq hasSuccessor \quad (10)$$

$$Fix \sqcap \neg \exists endsAt^{-}.Segment \sqsubseteq StartingFix \quad (11)$$

$$Fix \sqcap \neg \exists startsFrom^{-}.Segment \sqsubseteq EndingFix \quad (12)$$

$$Segment \sqcap \exists startsFrom.StartingFix \sqsubseteq StartingSegment \quad (13)$$

$$Segment \sqcap \exists endsAt.EndingFix \sqsubseteq EndingSegment \quad (14)$$

$$SemanticTrajectory \sqsubseteq \exists hasSegment.Segment \quad (15)$$

$$hasSegment \circ startsFrom \sqsubseteq hasFix \quad (16)$$

$$hasSegment \circ endsAt \sqsubseteq hasFix \quad (17)$$

$$\exists hasSegment.Segment \sqsubseteq SemanticTrajectory \quad (18)$$

$$\exists hasSegment^{-}.SemanticTrajectory \sqsubseteq Segment \quad (19)$$

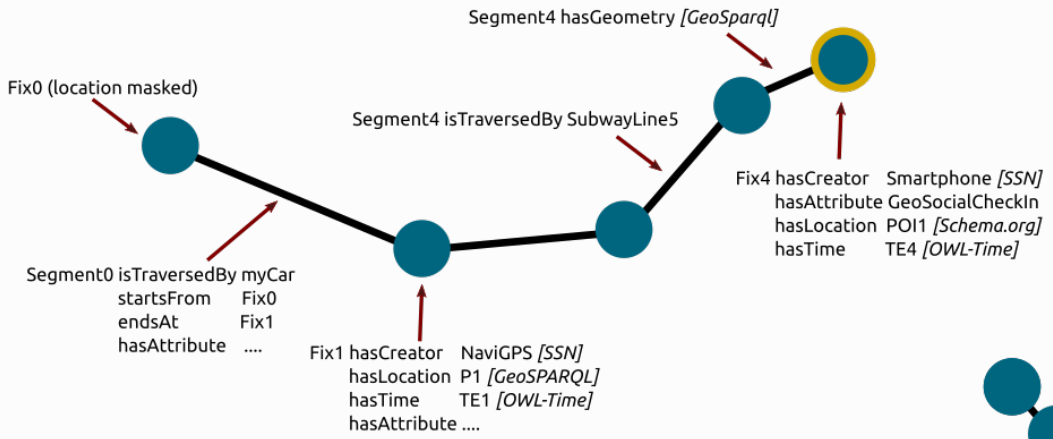
$$\exists hasFix.Fix \sqsubseteq SemanticTrajectory \quad (20)$$

$$\exists hasFix^{-}.SemanticTrajectory \sqsubseteq Fix \quad (21)$$

$$\ldots \quad (22)$$

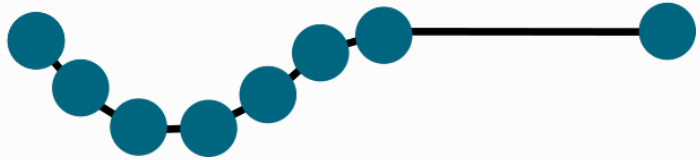
Rationale Behind the Trajectory Pattern

Human travel trajectory



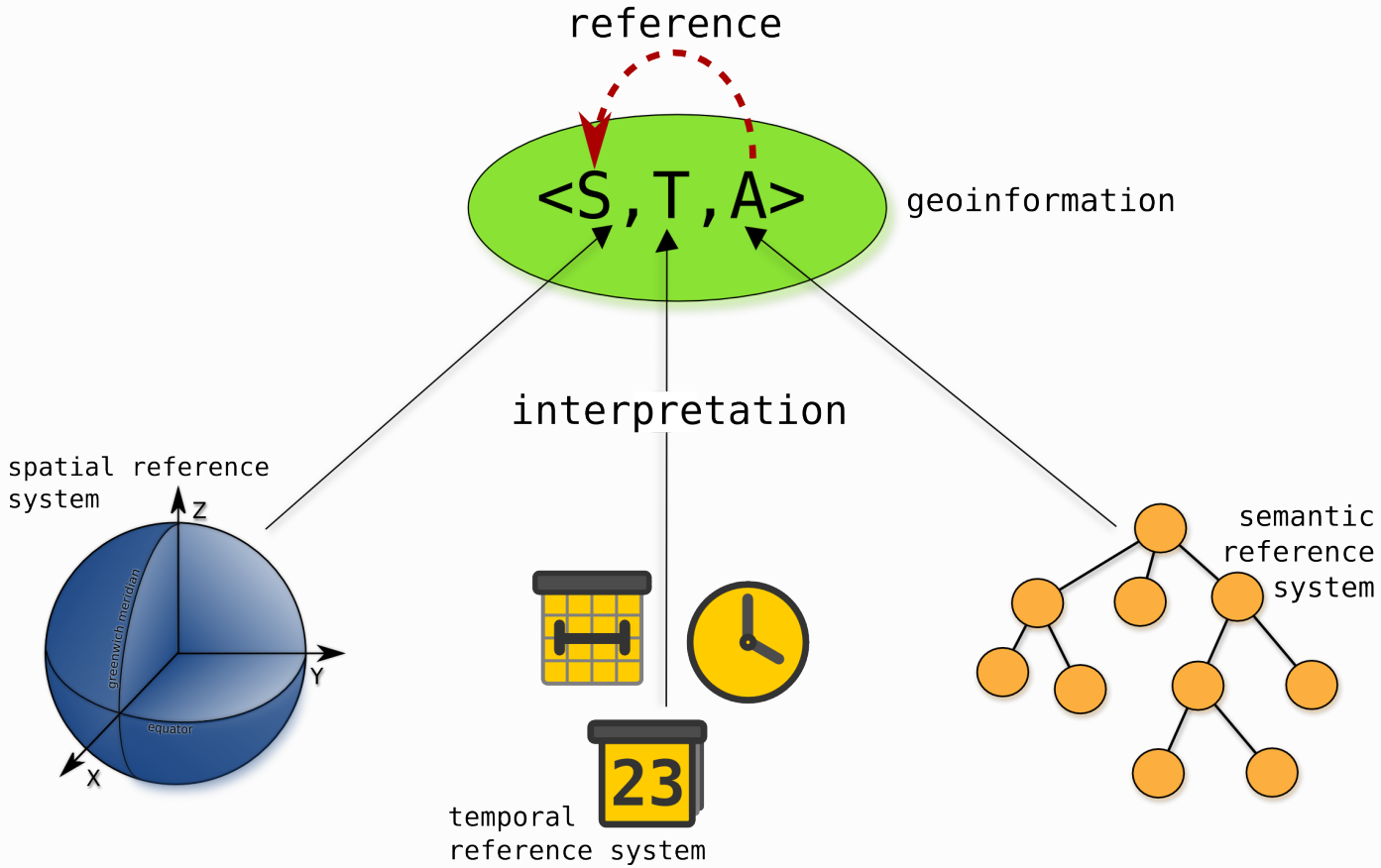
Abstraction

Discretization



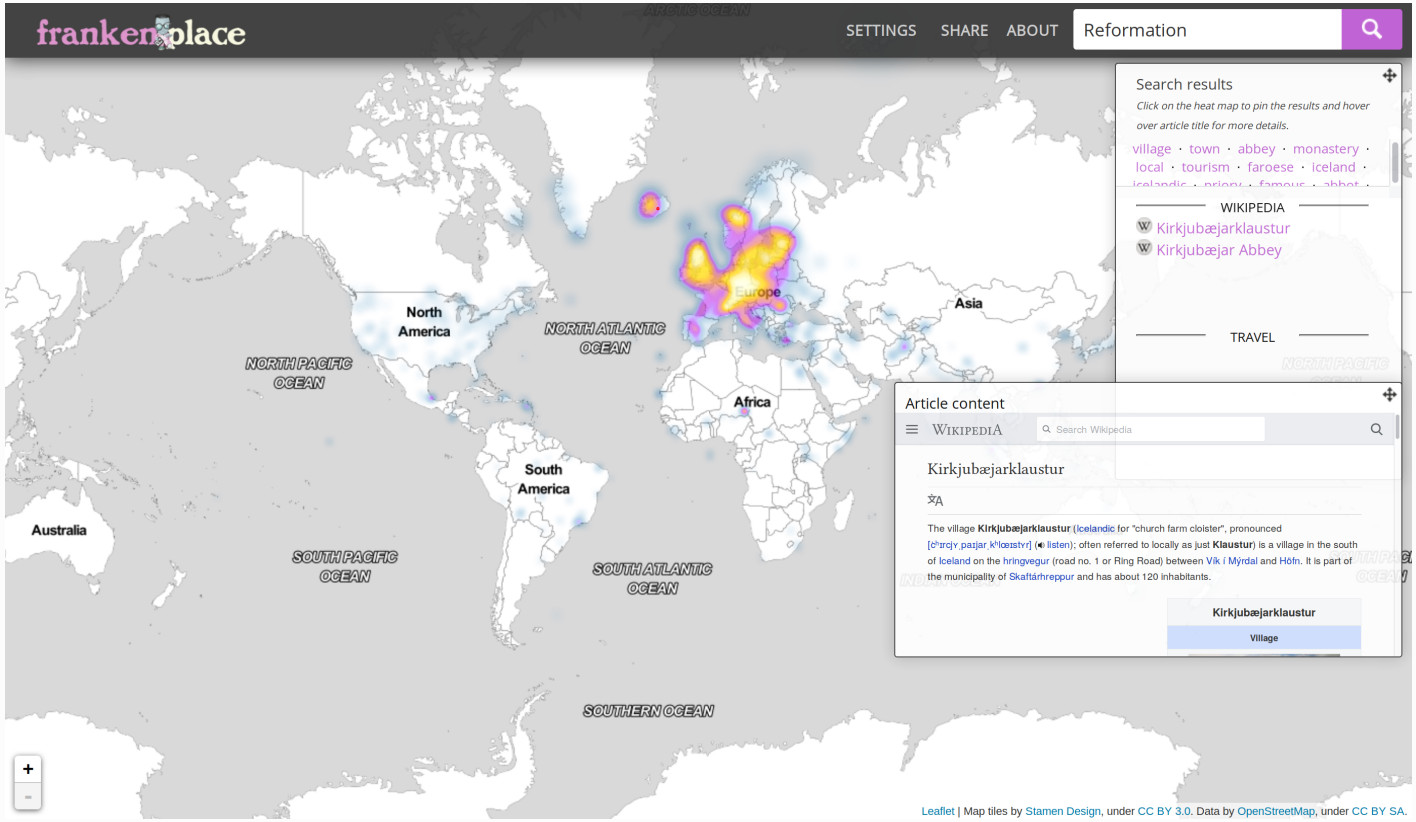
- Can be easily **extended** (see previous slide)
- Can cover a wide range of **domains**, e.g., wildlife tracking, historic routes
- Supports multiple **granularities**
- **Axiomatization** beyond mere surface semantics
- Has **hooks** to well-known ontologies, e.g., **SSN**.
- **Key question:** Where do we **stop** modelling? What is a place, source, etc?

Thematic Data as Additional Reference



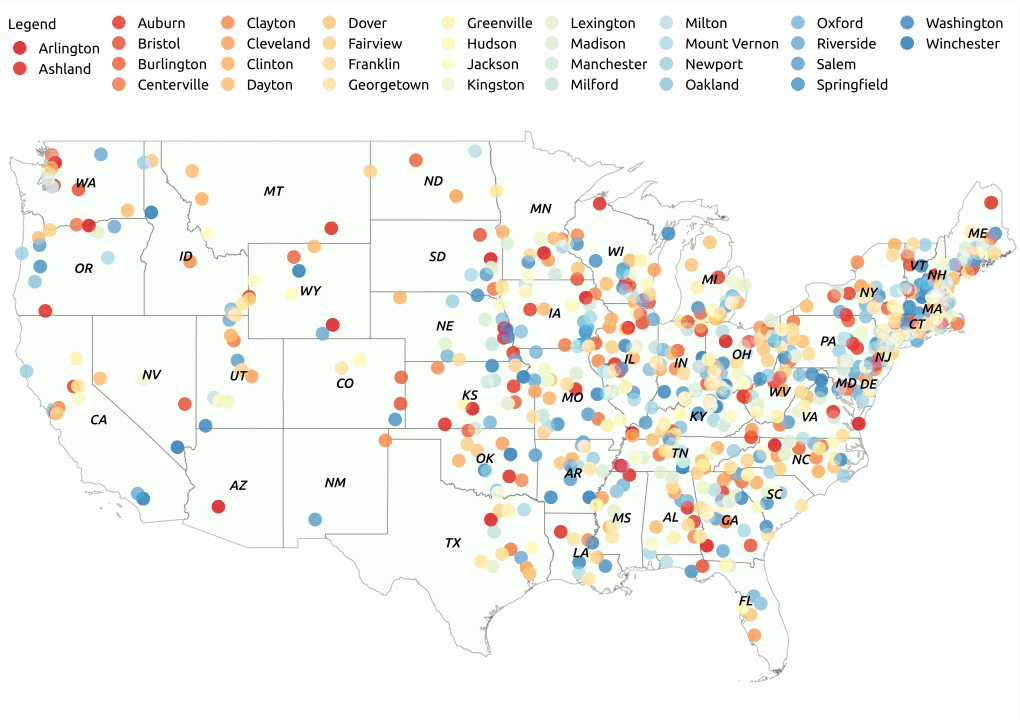
- Thematic data can be used as additional source of **reference** if spatial (or temporal) data is vague or insufficient.

Text is Geo-Indicative



- So far, we largely focused on **top-down** knowledge engineering (to make data reusable), what about **bottom-up** learning, e.g., for place name disambiguation?
- It is possible to learn the **geo-indicativeness** of non-georeferenced text

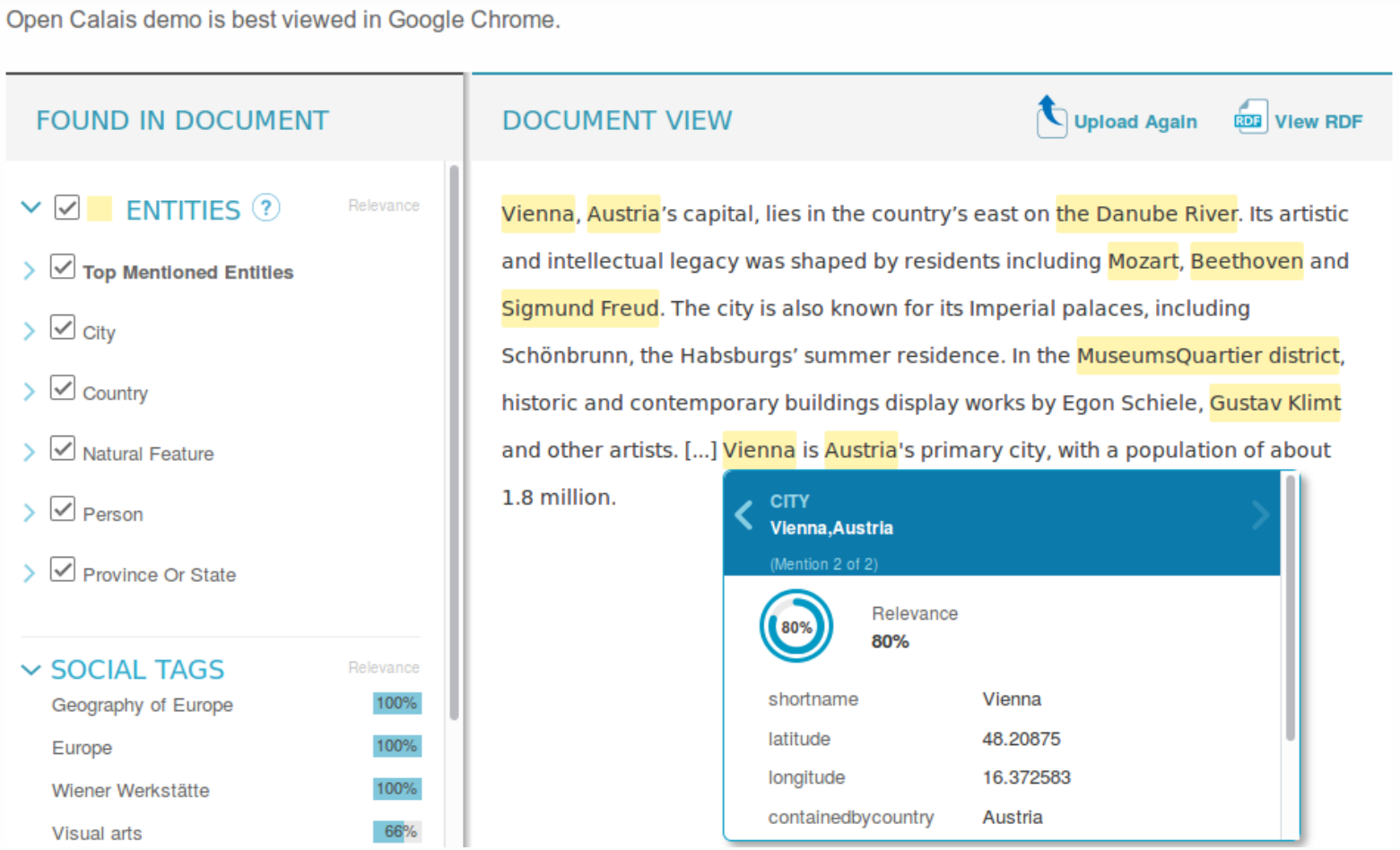
Place Name Disambiguation (from Short Texts)



- **Goal:** Given a short text which contains a place name and given a list of candidate places that share this name, determine to which specific place the text refers.

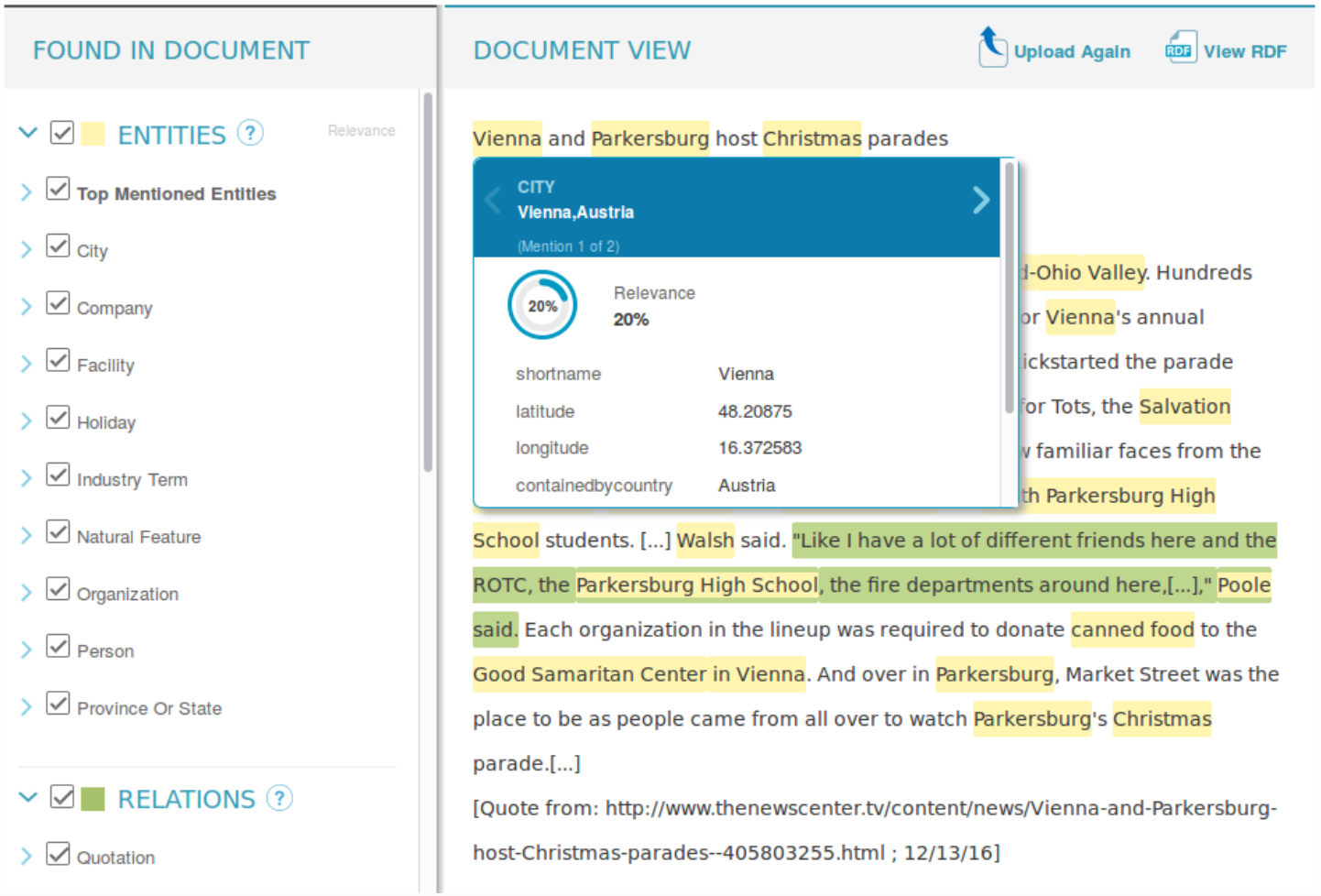
- **Challenge:** Disambiguation accuracy depends on how well we can capture the contexts in which a place name is uttered, but short texts only provide a few contextual clues.
- **Relevance:** Identifying places in unstructured text and semantically turning these texts into semi-structured data is important for a variety of tasks such as geographic information retrieval, geo-knowledge extraction and management, question answering, and exploratory search.

OpenCalais and DBpedia Spotlight (Linked Data-based Enrichment)



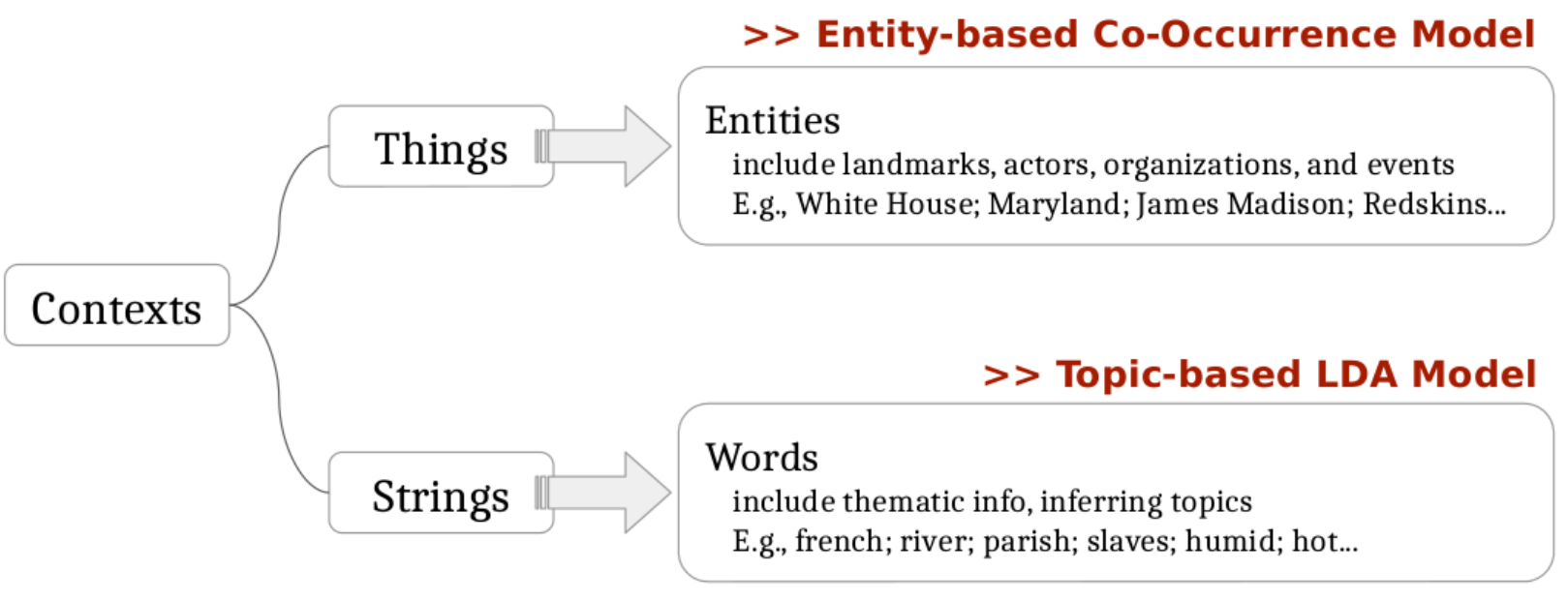
Core idea: try to extract *named* places, people, organizations, and themes.

OpenCalais and DBpedia Spotlight (Linked Data-based Enrichment)



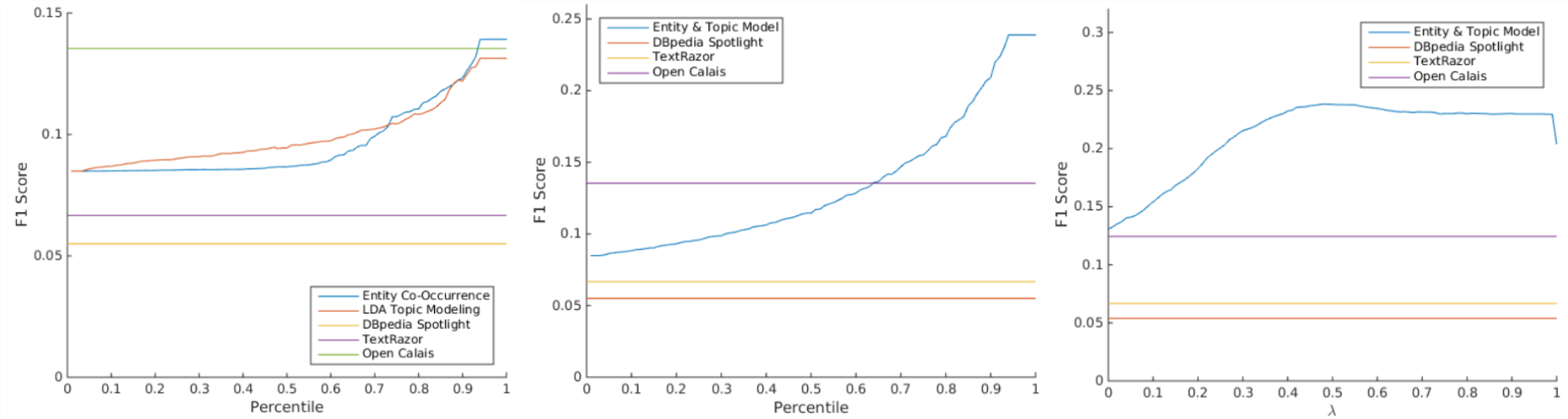
Disambiguation accuracy largely depends on interpreting **contextual clues** making the task very hard for **short texts**.

Things and Strings Model



- **Core idea:** bring together structured and unstructured content
  - Detect related **entities** (not merely places) from **knowledge graphs** and relate them to the list of potential places.
  - Use **thematic signatures** to geo-located non-geographic text, e.g., summer residence, canned food, parade, imperial,...
  - Learn **weights** between entity co-occurrence model and LDA topic model from training data and adjust for common entities and terms (TF-IDF).

Things and Strings Model



Model	Parameters	F1-Score
DBpedia Spotlight	*Confidence* = 0.2; *Support* = 0	0.055
TextRazor	n/a	0.067
Open Calais	n/a	0.135
Our **ETM**	λ = 0.48; 94th percentile	**0.239**

Example: Dayton, Nevada: *As the Native-American tribes in the area were nomadic, this made Dayton the first and oldest permanent non-native settlement in ~~Nevada~~.*

Conclusions

- We are increasingly **re-using** and **combining** data from **heterogeneous** sources
- Finding and successfully **integrating** these data is far from trivial
- Ideally, data would be **self-describing**
- **Spatiotemporally scoping** statements and terminology remains a major challenge
- **Semantic aging** is part of the overall replication crisis in science.
- The **Linked Data paradigm** seems to be among the best candidates we have so far, but it is not without problems
- Most interesting work is happening at the **intersection** of top-down knowledge engineering and bottom-up learning