

LINKED DATA AND THE DIGITAL HUMANITIES – A MATCH MADE IN HEAVEN OR HELL?

Krzysztof Janowicz and Yingjie Hu

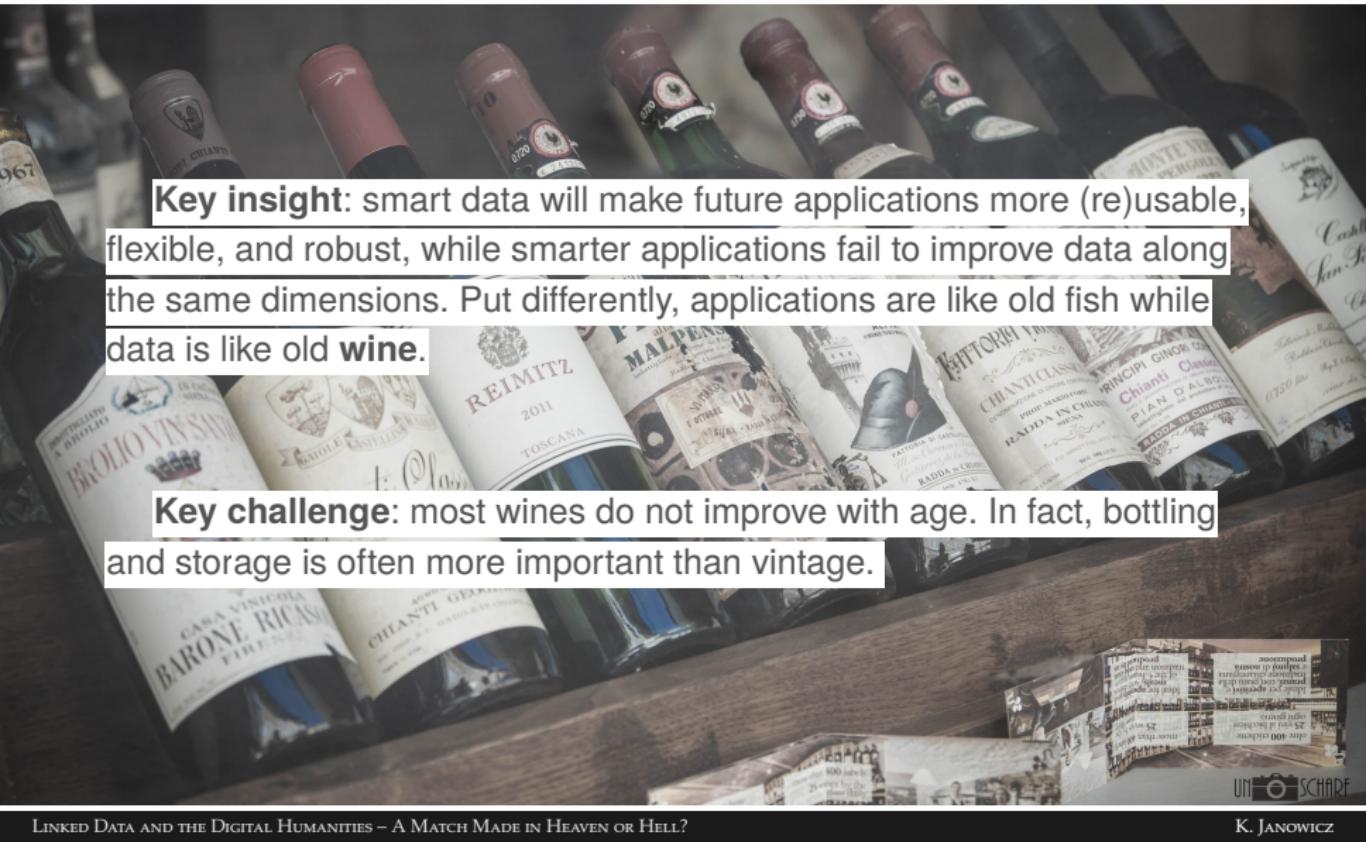
STKO Lab, University of California, Santa Barbara, USA

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.

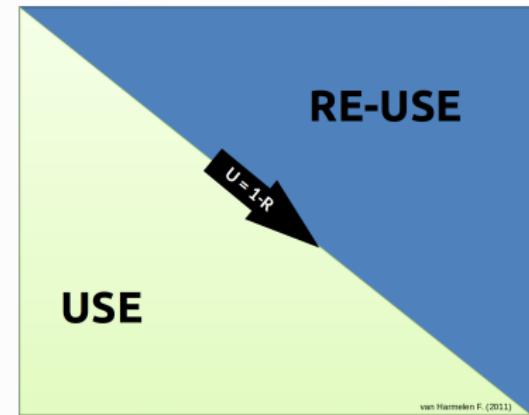
Key challenge: most wines do not improve with age. In fact, bottling

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]



van Harmelen F. (2011)

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**

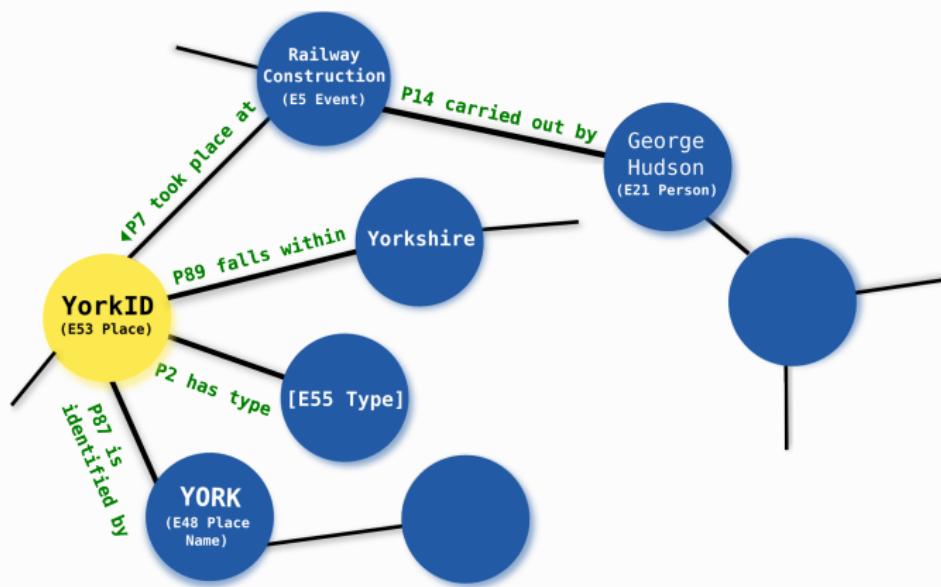
We will outline how one could accomplish these steps using the so-called **Linked Data paradigm** by showing examples centered around **places** and **trajectories**.

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)

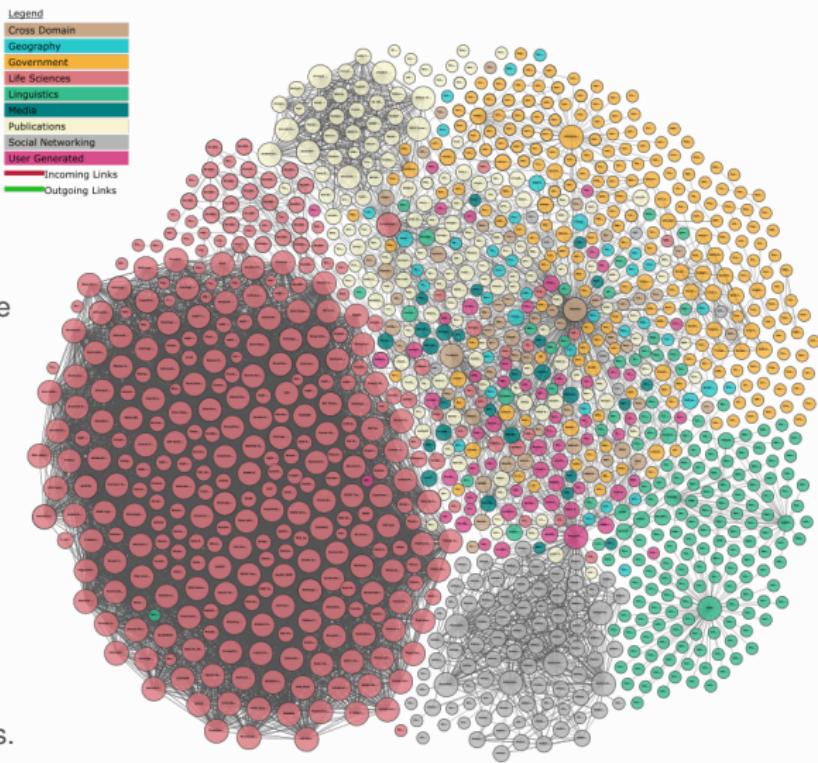


- Place as a **nexus** connecting events, objects, and actors.
- What do we **know** about these events, objects, and actors?
- How does York, UK **relate** to the Scandinavian city of Yorvik?

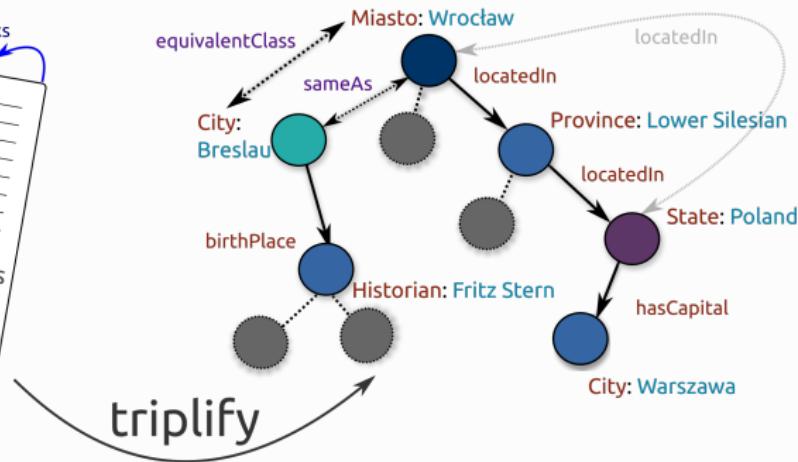
LINKED DATA: A DENSELY INTERCONNECTED GLOBAL KNOWLEDGE GRAPH

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.

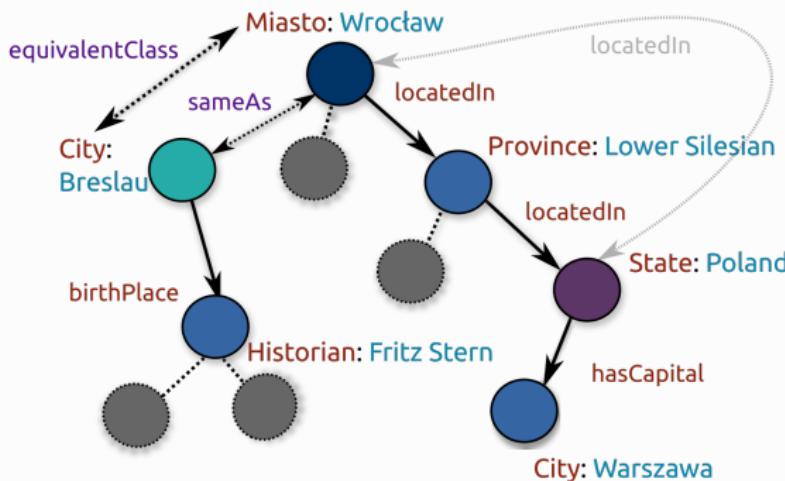


FROM LINKED DOCUMENTS TO LINKED DATA



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

phuzzy.link <http://phuzzy.link/browse/dbpedia.org/sparql?db=rdf:York>

Settings:
language: "en"
locals: "en-US"
limit: 128

Plugins:
phuzzy-xsd
phuzzy-language-filter
phuzzy-colored-prefixes: 1 arg
phuzzy-info: 1 arg
phuzzy-geo: 1 arg

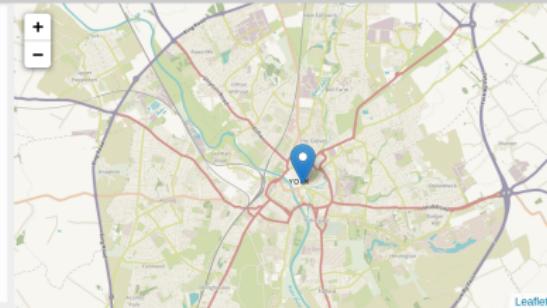
Outgoing properties: 818 triples Incoming properties: 5345 triples

dbr:York [View RDF source </>](#)

YORK (jorvik) is a historic walled city at the confluence of the rivers Ouse and Foss in North Yorkshire, England. The municipality is the traditional county town of Yorkshire to which it gives its name. The City has a rich heritage and has provided the backdrop to major political events in England throughout much of its two millennia of existence. The city offers a wealth of historic attractions, of which York Minster is the most prominent, and a variety of cultural and sporting activities making it a popular tourist destination for millions. The city was founded by the Romans as Eboracum in 71 AD. It became the capital of the Roman province of Britannia Inferior, and later of the kingdoms of Northumbria and Jórvík. In the Middle Ages, York grew as a major wool trading centre and became the capital of the northern ecclesiastical province of the Church of England, a role it has retained. In the 19th century, York became a hub of the railway network and a confectionery manufacturing centre. In recent decades, the economy of York has moved from being dominated by its confectionery and railway-related industries to one that provides services. The University of York and health services have become major employers, whilst tourism has become an important element of the local economy. From 1996, the term City of York describes a unitary authority area which includes rural areas beyond the old city boundaries. In 2011 the urban area had a population of 153,717, while in 2010 the entire unitary authority had an estimated population of 202,400.

[rdf:type](#) ↗

- [yago:YagoPermanentlyLocatedEntity](#) ↗
- [dbo:City](#) ↗
- [dbo:Location](#) ↗
- [dbo:Place](#) ↗
- [dbo:PopulatedPlace](#) ↗



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

HUMAN AND MACHINE READABLE AND REASON-ABLE

phuzzy.link

SPARQL endpoint URL: <http://dbpedia.org/sparql>

Prefix JSON-LD context: <https://phuzzy.link/context>

Settings: language: "en" locale: "en-US" limit: 128

Plugins: phuzzy-xsd phuzzy-language-filter phuzzy-colored-prefixes: 1 arg phuzzy-info: 1 arg phuzzy-geo: 1 arg

Outgoing properties: 818 triples Incoming properties: 5345 triples

dbr:York Hide RDF source </>

Generate from SELECT results

text/turtle application/trig application/n-triples application/n-quads

Dereference URI from: dbpedia.org

text/turtle application/n-triples application/ld+json application/rdf+xml

text/n3

47 dbr:areaCode "01904" ;
48 dbr:areaTotal "2.7194e+08"^^xsd:double ;
49 dbr:governmentType dbr:City_Status_in_the_United_Kingdom, dbr:Unitary_authority ;
50 dbe:isPartOf dbr:England, dbr:North_Yorkshire, dbr:Yorkshire_and_the_Humber ;
51 dbe:leaderName <http://dbpedia.org/resource/Conservative_Party_(UK)>, dbr:Julian_Sturdy, dbr:Rachael_Maskell ;
52 dbe:leaderTitle "Executive";@en, "Governing body"@en, "Leadership"@en, "MPs"@en ;
53 dbe:motto "Let the Banner of York Fly High" ;
54 dbe:populationDensity "687.0"^^xsd:double ;
55 dbe:postalcodes "YO" ;
56 dbo:thumbnail "https://commons.wikimedia.org/w/index.php?title=File:York_Bird's_Eye_View.jpg&width=300" ;
57 dbo:timeZone dbr:British_Summer_Time, dbr:Greenwich_Mean_Time ;
58 dbo:type dbr:City_Status_in_the_United_Kingdom, dbr:Unitary_authority ;
59 dbo:useOffset "+0", "+1" ;
60 dbo:wikipediaExternalLink <[https://www.historyoffyork.org.uk/](https://www.gutenberg.org/etext/1748#>, <, <<https://www.york.gov.uk/>>, <<https://web.archive.org/web/20160304130656/http://www.yayays.org/evelyn.html>>;
61 <<https://web.archive.org/web/20160304130656/http://www.yayays.org/evelyn.html>> .
62 dbo:wikiPageWikiLink "7445464#>"^^xsd:integer ;
63 dbo:wikiPageWikiLink "7445464#>"^^xsd:integer ;
64 dbo:apricHigh "12.5"^^xsd:double ;
65 dbo:apricLow "3.9"^^xsd:double ;
66 dbo:apricPrecipitationDays "9.3"^^xsd:double ;
67 dbo:apricPrecipitationMm "50.1"^^xsd:double ;
68 dbo:apricRecordHigh "25"^^xsd:integer ;
69 dbo:apricRecordLow "-3"^^xsd:integer ;
70 dbo:aprisun "141"^^xsd:integer ;

York (Jōrk) is a historic walled city at the confluence of the rivers Ouse and Foss in North Yorkshire, England. The municipality is the traditional county town of Yorkshire to which it gives its name. The city has a rich heritage and has provided the backdrop to major political events in England throughout much of its two millennia of existence. The city offers a wealth of historic attractions, of which York Minster is the most prominent, and a variety of cultural and sporting activities making it a popular tourist destination for millions. The city was founded by the Romans as Eboracum in 71 AD. It

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

DIGITAL HUMANITIES HEAVEN: RECOGITO BY PELAGIOS

The screenshot shows a digital humanities annotation interface. At the top, there are icons for edit, user profile, download, and search, followed by the text "Logged in as rainer". Below this, the title "Homer: The Odyssey" and author "800-700 BC" are displayed, along with a note about 14 annotations and no other contributors.

The interface includes a "ANNOTATION MODE" dropdown set to "NORMAL", a "COLOR" dropdown set to "BY ENTITY TYPE" (highlighted in yellow), and buttons for "BY VERIFICATION STATUS".

The main text area contains a passage from Homer's Odyssey: "Tell me, O muse, of that ingenious hero who travelled far and wide after he had sacked the famous town of Troy. He suffered much at the hands of the sons of Atreus, because of his friend Agamemnon, and customs he was forced to make. To end his life and bring his mortal bones down to Hades, he was driven by Poseidon to the land of the Phaeacians, where he perished through the treachery of the Phaeacians, who prevented them from getting him back home again, from whatsoever source he might have come."

A tooltip is overlaid on the word "hero", showing three categories: "Place" (marked with a location pin icon), "Person" (marked with a person icon), and "Event" (marked with a star icon). The "Person" category is selected. The tooltip also contains fields for "Add a comment..." and "Add tag...", and buttons for "Cancel", "OK & Next", and "OK".

Below the main text, another passage is visible: "So now all who escaped death in battle or by shipwreck had got safely home except Ulysses, and he, though he was longing to return to his wife and country, was detained by the goddess Calypso,"

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+

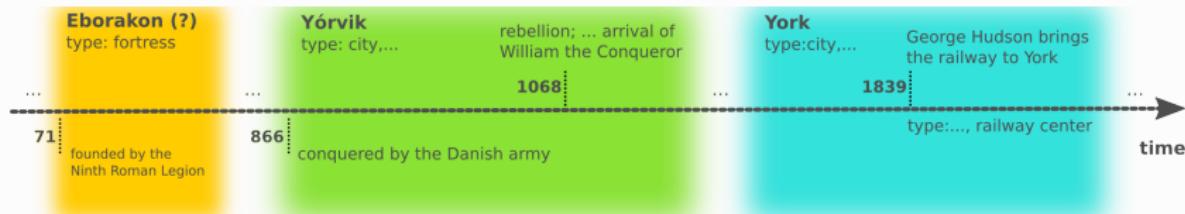
The screenshot displays the DIVE+ web application interface. At the top, there is a navigation bar with sections for 'EXPLORATION PATH', 'COLLECTION', 'DATESTART', and 'FILTER'. A search bar contains the identifier '/kb/ann:1951-03-21.1'. Below the search bar, a large panel shows a detailed view of a document entry, including its title 'ANP Nieuwsbericht - 21-03-1951 - 1' and various text snippets. To the right of this panel is a vertical sidebar labeled 'CLARIAH MEDIASUITE'.

Below the main search bar, there is a section titled 'Related entities' with a count of 73. It includes filters for 'example-exclude' and 'Unknown vocabulary' (selected). A date range from '1951-03-20' to '1951-03-20' is also present. The results list several entities, each with a thumbnail, name, and a snippet of text:

- Soestdijk**: 'De Heer minister-president Soestdijk is vandaag op koninklijk commando van koningin Juliana ontvangen door de burgemeester van Den Haag. De koningin heeft hem een uitvoerige ontvangst gegeven en een geschenk overhandigd. De koningin heeft er ook rekening mee gehouden dat Soestdijk vandaag een belangrijke dag ten doop heeft. (Littoral een dag na zijn vijftigste verjaardag).
- Baarms Lyceym**: 'De Heer minister-president Soestdijk is vandaag op koninklijk commando van koningin Juliana ontvangen door de burgemeester van Den Haag. De koningin heeft hem een uitvoerige ontvangst gegeven en een geschenk overhandigd. De koningin heeft er ook rekening mee gehouden dat Soestdijk vandaag een belangrijke dag ten doop heeft. (Littoral een dag na zijn vijftigste verjaardag).
- Paasvacantie**: 'De Heer minister-president Soestdijk is vandaag op koninklijk commando van koningin Juliana ontvangen door de burgemeester van Den Haag. De koningin heeft hem een uitvoerige ontvangst gegeven en een geschenk overhandigd. De koningin heeft er ook rekening mee gehouden dat Soestdijk vandaag een belangrijke dag ten doop heeft. (Littoral een dag na zijn vijftigste verjaardag).
- Prins Bernhard**: 'De Heer minister-president Soestdijk is vandaag op koninklijk commando van koningin Juliana ontvangen door de burgemeester van Den Haag. De koningin heeft hem een uitvoerige ontvangst gegeven en een geschenk overhandigd. De koningin heeft er ook rekening mee gehouden dat Soestdijk vandaag een belangrijke dag ten doop heeft. (Littoral een dag na zijn vijftigste verjaardag).

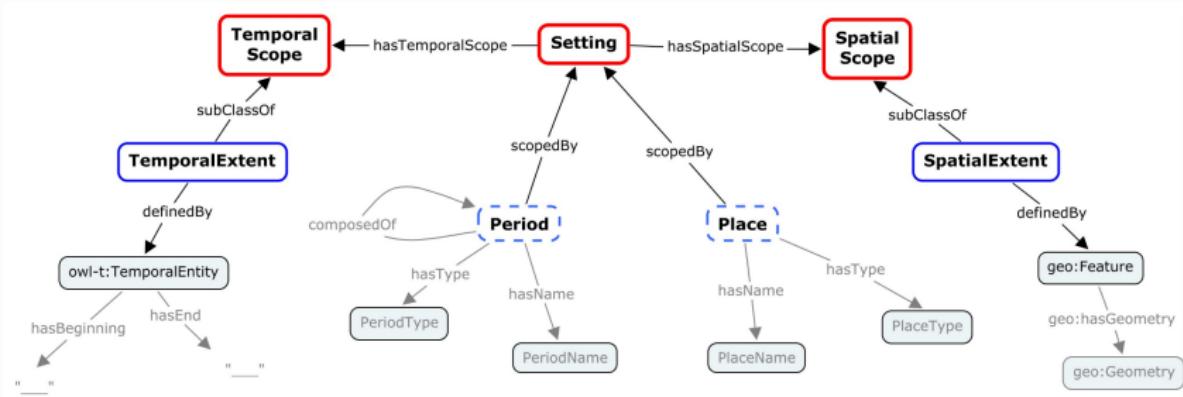
TEMPORAL SCOPING HELL

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



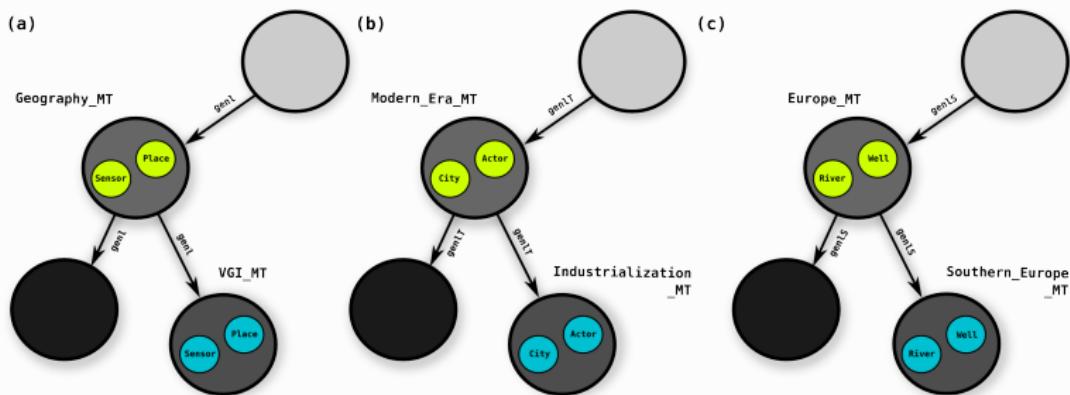
- Place as its **history** (and therefore as perdurant).
 - York is not a **Roman fortress (T)** anymore, and it does not go by the **name Jórvík (N)** anymore. It is also not **part (F)** of the Kingdom of Northumbria anymore.
 - Should there be **one URI** or many for the 'same' place?
-
- Technical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

PLACE, PERIOD, AND SETTING FOR LINKED DATA GAZETTEERS



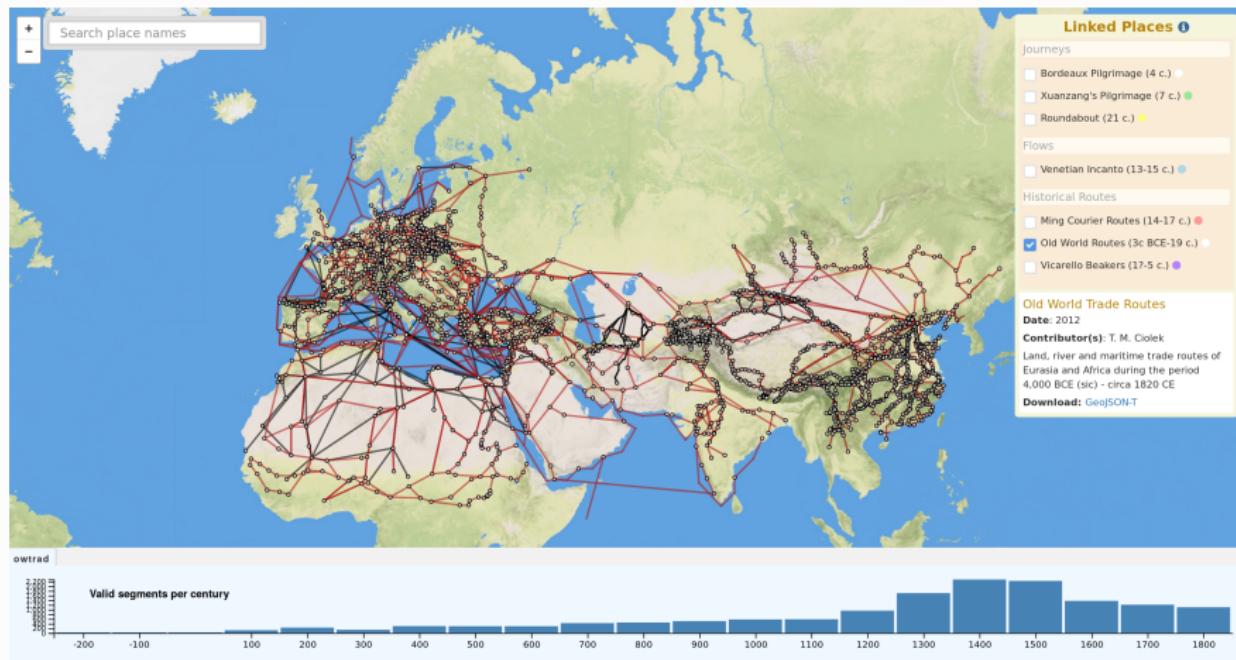
A modeling solution: **places** are also **temporal** and **events/periods** are also **spatial**, e.g., Levant Bronze Age and the Age of Enlightenment.

DIGITAL HUMANITIES HELL: SCOPING TERMINOLOGY BY SPACE AND TIME



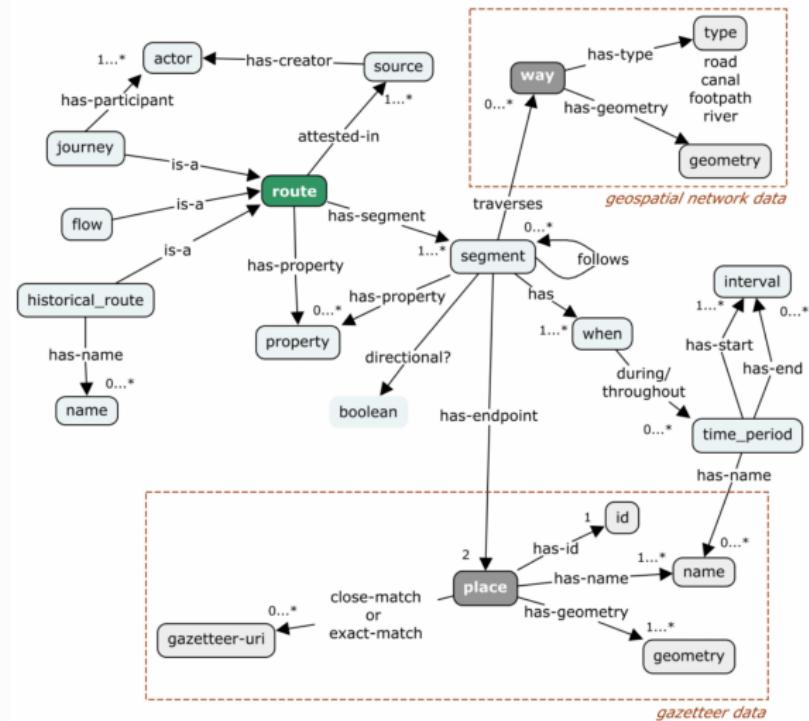
- Different domains, e.g., historic periods, require **different conceptualizations**
- ‘Is one man’s terrorist another man’s freedom fighter?’
- **Local and consistent** theories structured by subsumption, space, and time within a **global and inconsistent** theory.

FROM INDIVIDUAL PLACES TO ROUTES, FLOWS, AND PILGRIMAGES

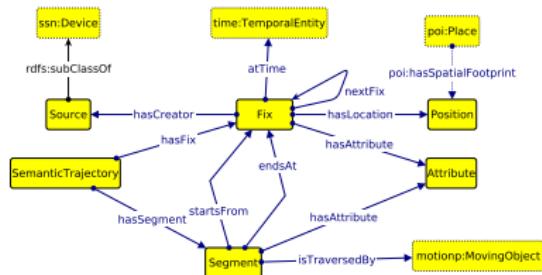


Is there a **common vocabulary** (ontology) for describing trajectories of goods, people, vessels, wildlife?

A CONCEPTUAL MODEL OF HISTORICAL GEOGRAPHICAL MOVEMENT



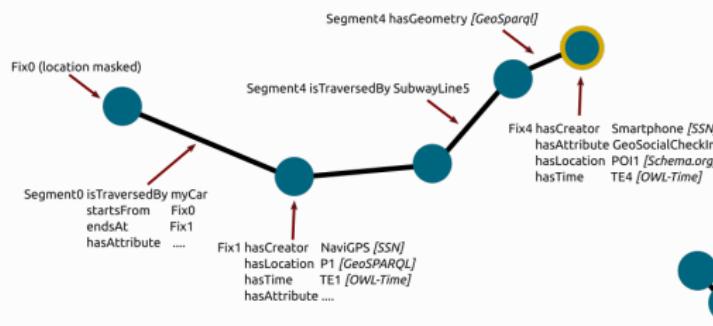
A COMMON CORE ONTOLOGY FOR TRAJECTORIES REUSABLE ACROSS DOMAINS



- $\text{Fix} \sqsubseteq \exists \text{atTime}.\text{time:TemporalEntity} \sqcap \exists \text{hasLocation}.\text{Position}$ (1)
- $\sqcap \exists \text{hasFix}^-.\text{SemanticTrajectory}$ (2)
- $\text{Segment} \sqsubseteq \exists \text{startsFrom}.\text{Fix} \sqcap \exists \text{endsAt}.\text{Fix}$ (3)
- $\text{T} \sqsubseteq \exists \text{startsFrom}.\text{T}$ (4)
- $\text{T} \sqsubseteq \exists \text{endsAt}.\text{T}$ (5)
- $\text{Segment} \sqsubseteq \exists \text{hasSegment}^-.\text{SemanticTrajectory}$ (6)
- $\text{hasNext} \sqsubseteq \text{hasSuccessor}$ (7)
- $\text{hasNext}^- \sqsubseteq \text{hasPrevious}$ (8)
- $\text{hasSuccessor}^- \sqsubseteq \text{hasPredecessor}$ (9)
- $\text{startsFrom}^- \circ \text{endsAt} \sqsubseteq \text{hasNext}$ (10)
- $\text{hasSuccessor} \circ \text{hasSuccessor} \sqsubseteq \text{hasSuccessor}$ (11)
- $\text{Fix} \sqcap \neg \exists \text{endsAt}^-.\text{Segment} \sqsubseteq \text{StartingFix}$ (12)
- $\text{Fix} \sqcap \neg \exists \text{startsFrom}^-.\text{Segment} \sqsubseteq \text{EndingFix}$ (13)
- $\text{Segment} \sqcap \exists \text{startsFrom}.\text{StartingFix} \sqsubseteq \text{StartingSegment}$ (14)
- $\text{Segment} \sqcap \exists \text{endsAt}.\text{EndingFix} \sqsubseteq \text{EndingSegment}$ (15)
- $\text{SemanticTrajectory} \sqsubseteq \exists \text{hasSegment}.\text{Segment}$ (16)
- $\text{hasSegment} \circ \text{startsFrom} \sqsubseteq \text{hasFix}$ (17)
- $\text{hasSegment} \circ \text{endsAt} \sqsubseteq \text{hasFix}$ (18)
- $\exists \text{hasSegment}.\text{Segment} \sqsubseteq \text{SemanticTrajectory}$ (19)
- $\exists \text{hasSegment}^-.\text{SemanticTrajectory} \sqsubseteq \text{Segment}$ (20)
- $\exists \text{hasFix}.\text{Fix} \sqsubseteq \text{SemanticTrajectory}$ (21)
- $\exists \text{hasFix}^-.\text{SemanticTrajectory} \sqsubseteq \text{Fix}$ (22)
- ...

RATIONALE BEHIND THE TRAJECTORY PATTERN

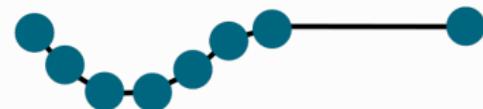
Human travel trajectory



Abstraction

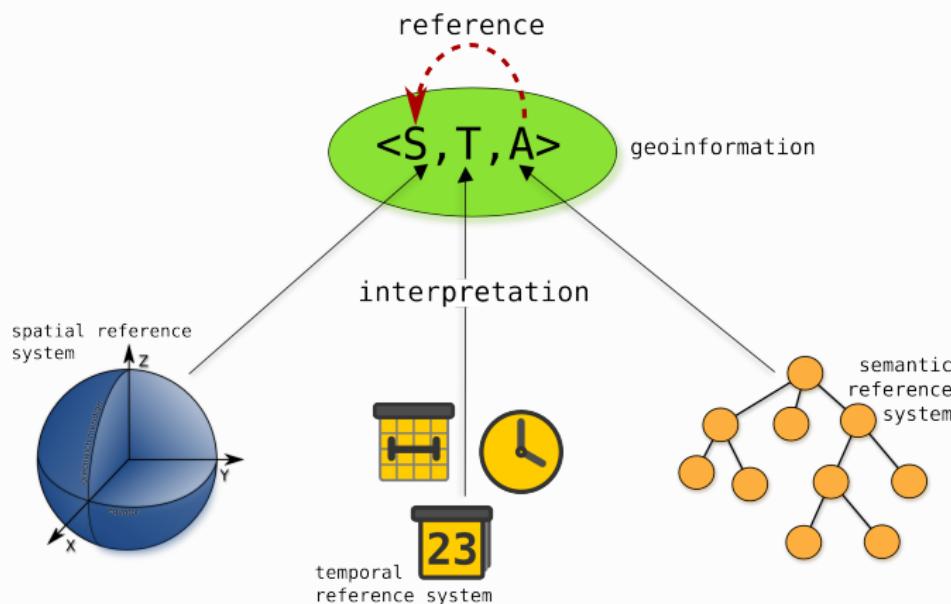


Discretization



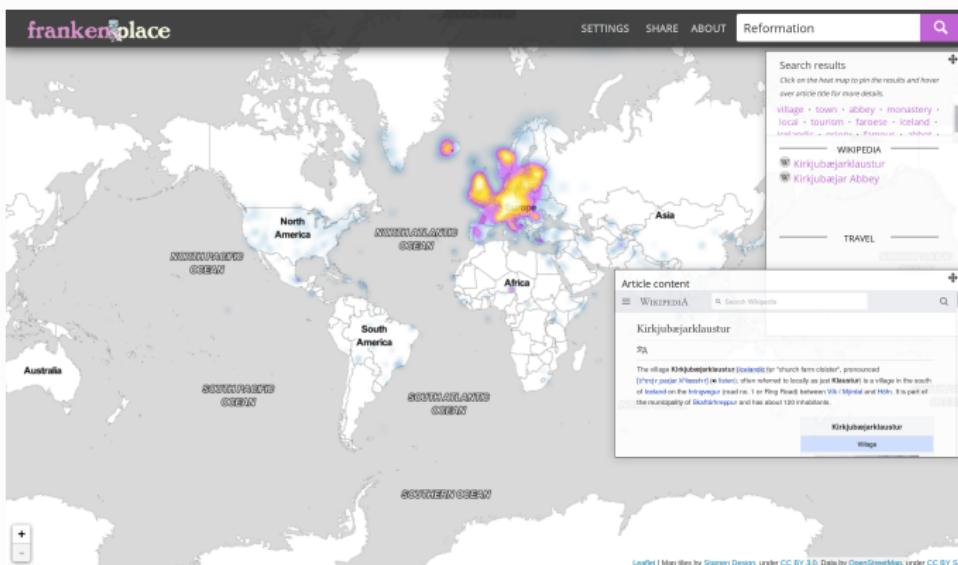
- Can be easily **extended** (see previous slide)
- Can cover a wide range of **domains**, e.g., wildlife tracking, historic routes
- Supports multiple **granularities**
- **Axiomatization** beyond mere surface semantics
- Has **hooks** to well-known ontologies, e.g., **SSN**.
- **Key question:** Where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

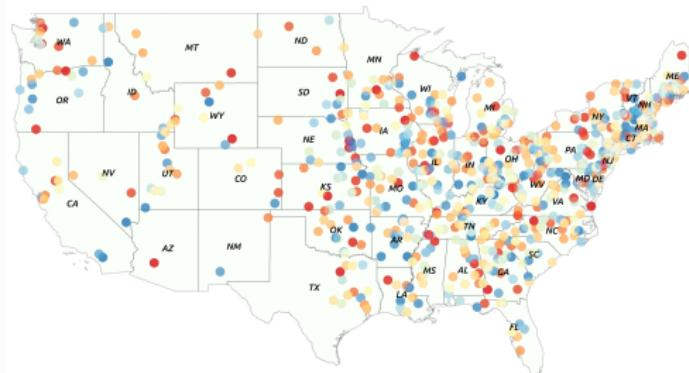


- So far, we largely focused on **top-down** knowledge engineering (to make data reusable), what about **bottom-up** learning, e.g., for place name disambiguation?
- It is possible to learn the **geo-indicativeness** of non-georeferenced text

PLACE NAME DISAMBIGUATION (FROM SHORT TEXTS)

Legend

Arlington	Auburn	Clayton	Dover	Greenville	Lexington	Milton	Oxford	Washington
Ashland	Bristol	Cleveland	Fairview	Hudson	Madison	Mount Vernon	Riverside	Winchester
Burlington	Clinton	Dayton	Georgetown	Jackson	Manchester	Newport	Salem	Springfield
Centerville				Kingston	Milford			



- **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)

Open Calais demo is best viewed in Google Chrome.

FOUND IN DOCUMENT

▼ **ENTITIES** [?](#) Relevance

- > Top Mentioned Entities
- > City
- > Country
- > Natural Feature
- > Person
- > Province Or State

▼ **SOCIAL TAGS** Relevance

Geography of Europe	100%
Europe	100%
Wiener Werkstätte	100%
Visual arts	66%

DOCUMENT VIEW

[Upload Again](#) [View RDF](#)

Vienna, Austria's capital, lies in the country's east on the Danube River. Its artistic and intellectual legacy was shaped by residents including Mozart, Beethoven and Sigmund Freud. The city is also known for its Imperial palaces, including Schönbrunn, the Habsburgs' summer residence. In the MuseumsQuartier district, historic and contemporary buildings display works by Egon Schiele, Gustav Klimt and other artists. [...] Vienna is Austria's primary city, with a population of about 1.8 million.

CITY
Vienna,Austria
(Mention 2 of 2)

	Relevance 80%
shortname	Vienna
latitude	48.20875
longitude	16.372583
containedbycountry	Austria

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

OPENCALAIS AND DBPEDIA SPOTLIGHT (LINKED DATA-BASED ENRICHMENT)

FOUND IN DOCUMENT

- ▼ ENTITIES ?** Relevance
- > Top Mentioned Entities
- > City
- > Company
- > Facility
- > Holiday
- > Industry Term
- > Natural Feature
- > Organization
- > Person
- > Province Or State

- ▼ RELATIONS ?**
- > Quotation

DOCUMENT VIEW

Upload Again

View RDF

Vienna and Parkersburg host Christmas parades

CITY
Vienna,Austria

(Mention 1 of 2)

Relevance
20%

shortname Vienna

latitude 48.20875

longitude 16.372583

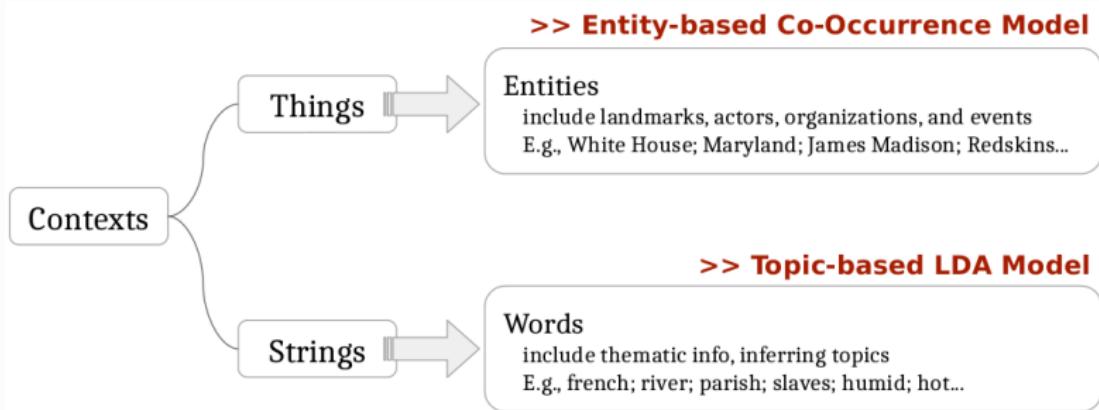
containedbycountry Austria

School students. [...] Walsh said. "Like I have a lot of different friends here and the ROTC, the Parkersburg High School, the fire departments around here.[...]," Poole said. Each organization in the lineup was required to donate canned food to the Good Samaritan Center in Vienna. And over in Parkersburg, Market Street was the place to be as people came from all over to watch Parkersburg's Christmas parade. [...]

[Quote from: <http://www.thenewscenter.tv/content/news/Vienna-and-Parkersburg-host-Christmas-parades--405803255.html> ; 12/13/16]

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

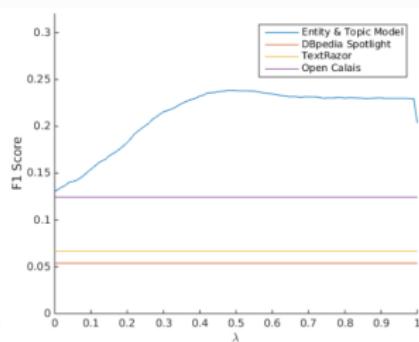
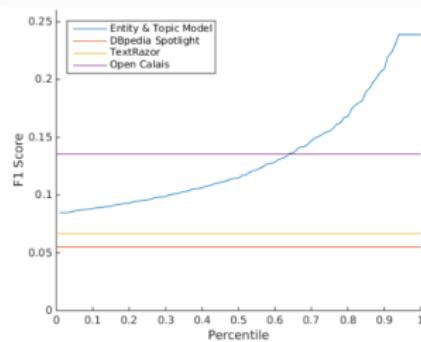
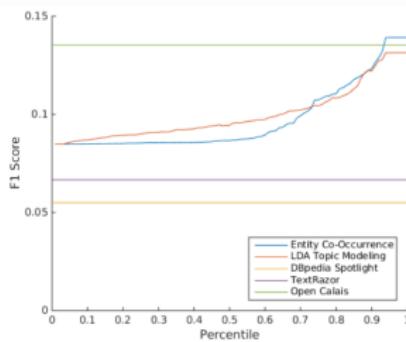
THINGS AND STRINGS MODEL



■ 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	<i>Confidence = 0.2; Support = 0</i>	0.055
TextRazor	n/a	0.067
Open Calais	n/a	0.135
Our ETM	$\lambda = 0.48$; 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