Deeply Integrating Linked Data with Geographic Information Systems

Esri UC 2019, July 9th 2019

Gengchen Mai [1] Krzysztof Janowicz [1] Bo Yan [1] Simon Scheider [2]

[1] STKO Lab, University of California, Santa Barbara

[2] Utrecht University, Netherlands

Linked Data (LD)

- A set of **design principles** for sharing machine-readable interlinked data on the Web[1]
- **Datasets** constructed by following these principles

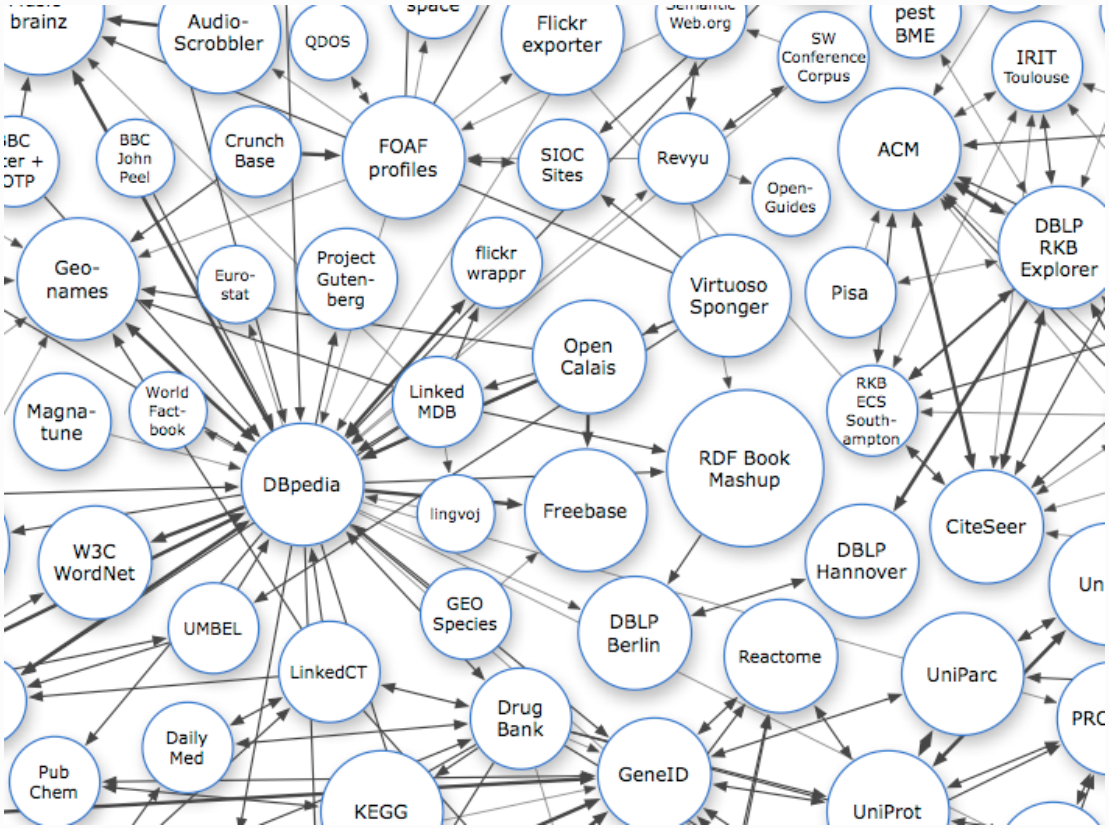
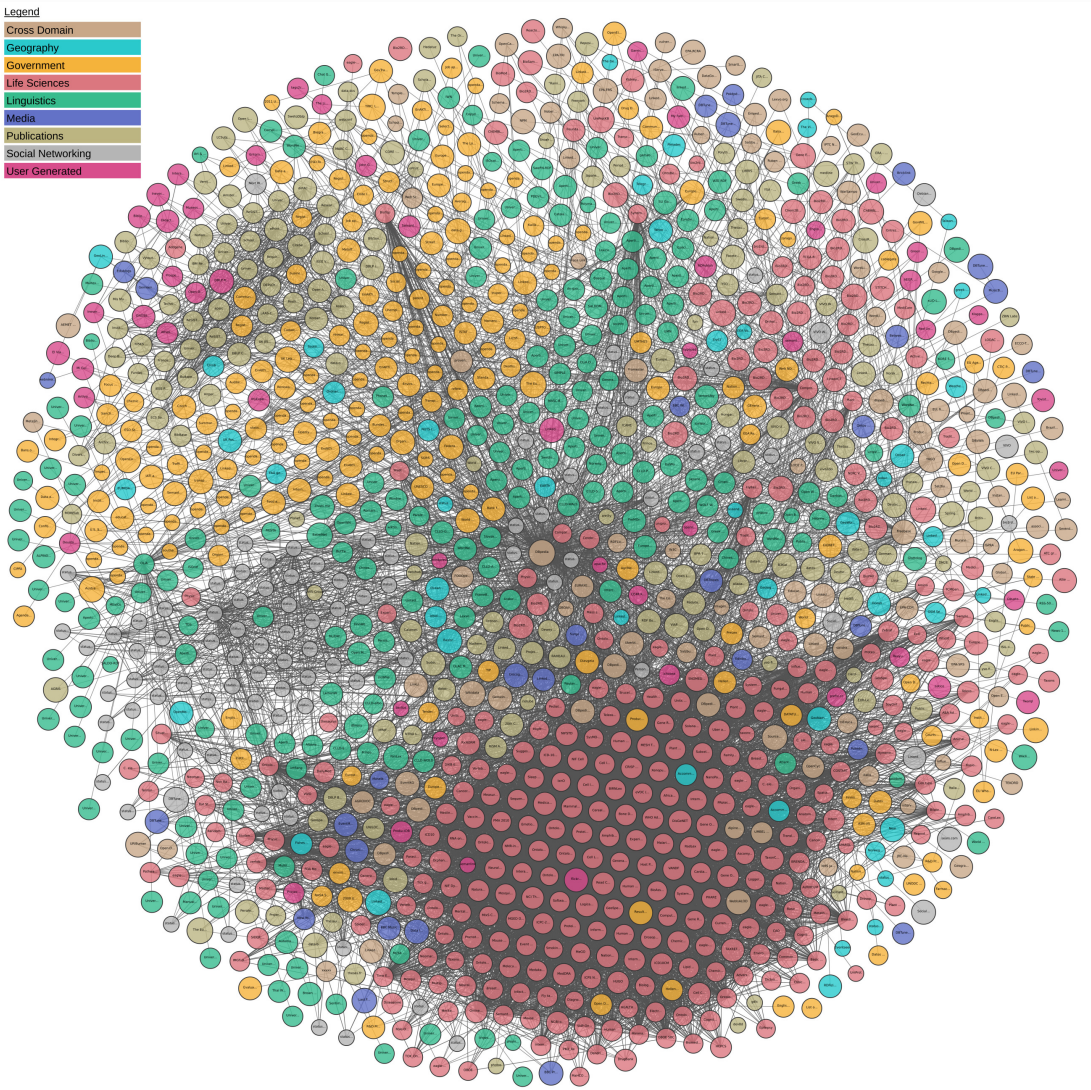


Figure From http://linkeddata.org/

[1] https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

Linked Open Cloud (LOD)



- **Geographic data** play a prominent role in the Linked Data cloud as places act as **central nexuses** that interconnect events, people, and objects.

Introduction

- **Linked Data** and **Knowledge Graphs** power many applications:
  - Intelligent assistance systems: Apple Siri
  - Search engines: Google Search
- The Linked Data paradigm addresses many key challenges of **GIScience** and cyber-data infrastructures (Kuhn et al., 2014).

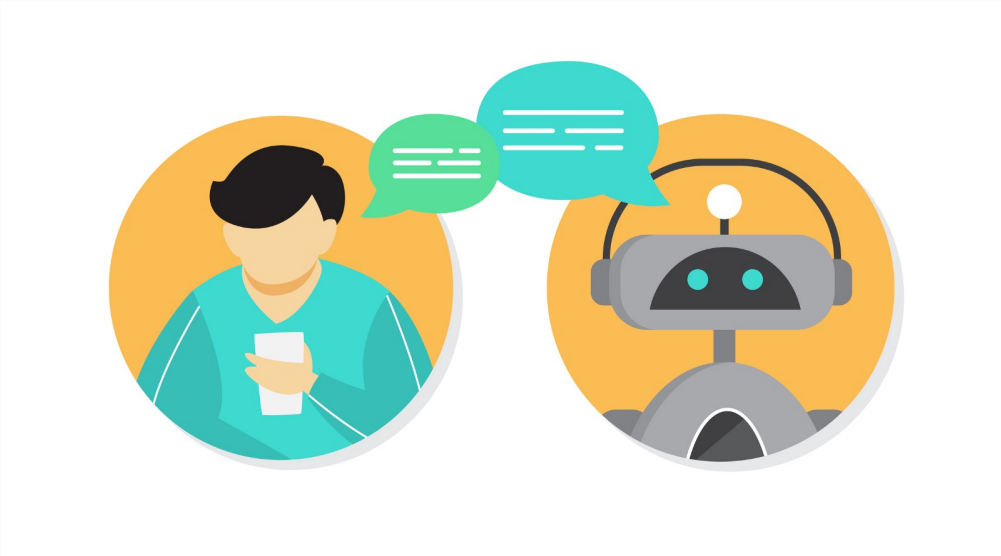
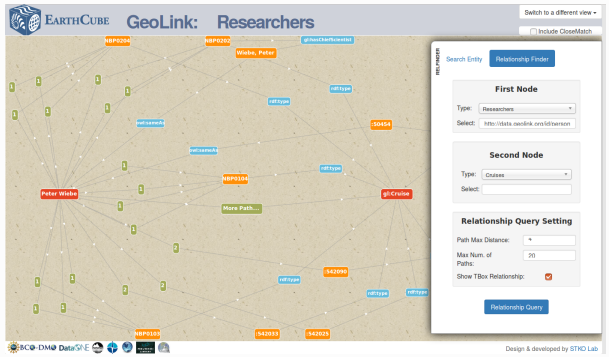


Figure From https://towardsdatascience.com/automatic-question-answering-ac7593432842

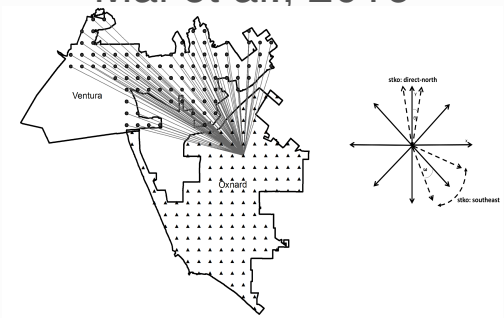
The plethora of Geographic Linked Data

Despite all these success stories, from a GIS perspective, **Linked Data seems almost like a one-way street**.

- It is now easier than ever to **publish and consume** geo-data on the (Semantic) Web:
  - Converting shapefiles to RDF
  - Fusing geometries from different sources (Giannopoulos et al., 2014)
  - Discovering links (Mai et al., 2016)
  - Querying remote endpoints (Battle et al., 2012)
  - Computing geospatial properties on-demand (Regalia et al., 2016).
- All this work focuses merely on how to get geo-data out of **data silos**.
- The question of **how to actually make use of this plethora of data** remains largely unanswered.



Mai et al., 2016



Regalia et al., 2016

The plethora of Geographic Linked Data

- Two main ways of accessing and utilizing Geographic Linked Data:
  - Directly use GeoSPARQL-enabled SPARQL endpoint: **very limited spatial analysis capability**
  - Convert Linked Data to CSV/JSON and Import them into a GIS: Data are flattened and the **link structure is largely lost**
- **Problem**: While we can semantically enrich geo-data and publish them as Linked Data, **consuming these data in a GIS** and applying the vast toolboxes of modern **spatial analysis** is more difficult, especially if we aim at maintaining the **link structure** while doing so.
- **Proposed Solution**: a **Linked Data connector framework** for the **deep integration of Linked Data and GIS**

Linked Data Connector

- At first glance - **a software engineering task**: add Linked Data as another data source of a GIS
- In reality - **a complicated research problem**:
  - Linked Data is **not a data format but a paradigm** for Web-scale, distributed data infrastructures
  - Linked Data does **not harmonize well** with how we conceptualize data (exchange) in GIS

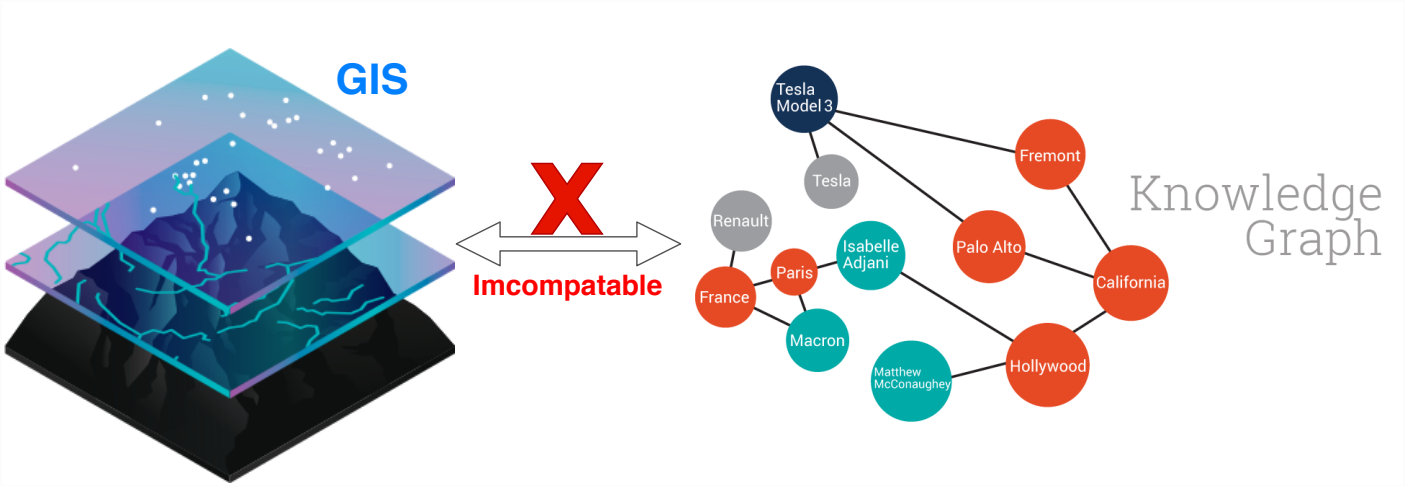


Figure made based on Esri and Ontotext images

Linked Data Connector

Considerations:

- How should GIS and its users **interact** with Linked Data?
- Which **concrete benefits** does Linked Data bring to the table with respect to spatial analysis?
- How can these key benefits of Linked Data be **maintained** during conversion and analysis **without having to flatten the data back to a tabular format**?
- How to utilize the **ontologies** used to semantically lift Linked Data?

Prototypical Implementation: we implement **individual toolboxes** for **ArcGIS 10.4** as proof-of-concept by using **ArcPy**.

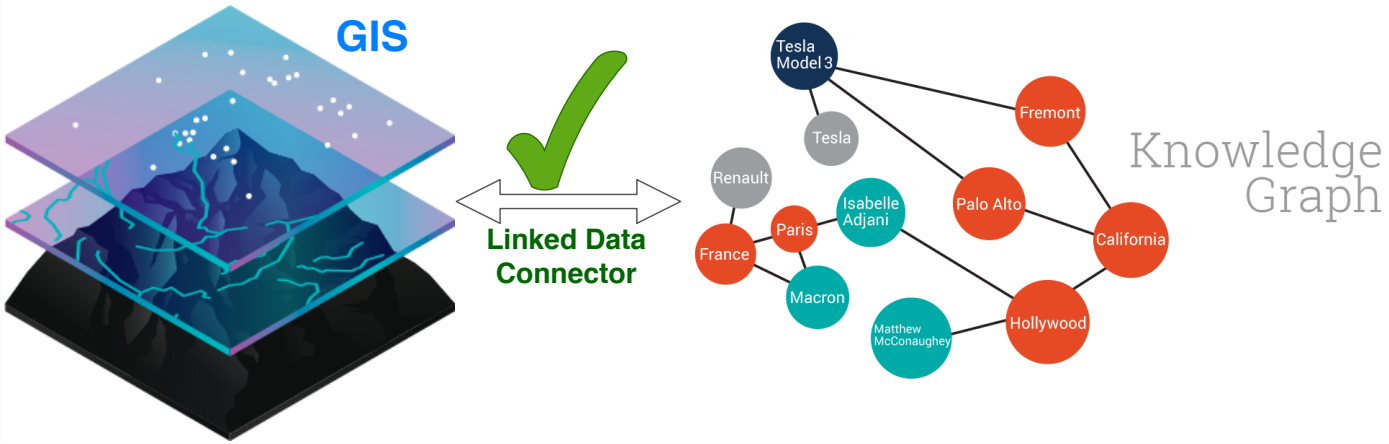
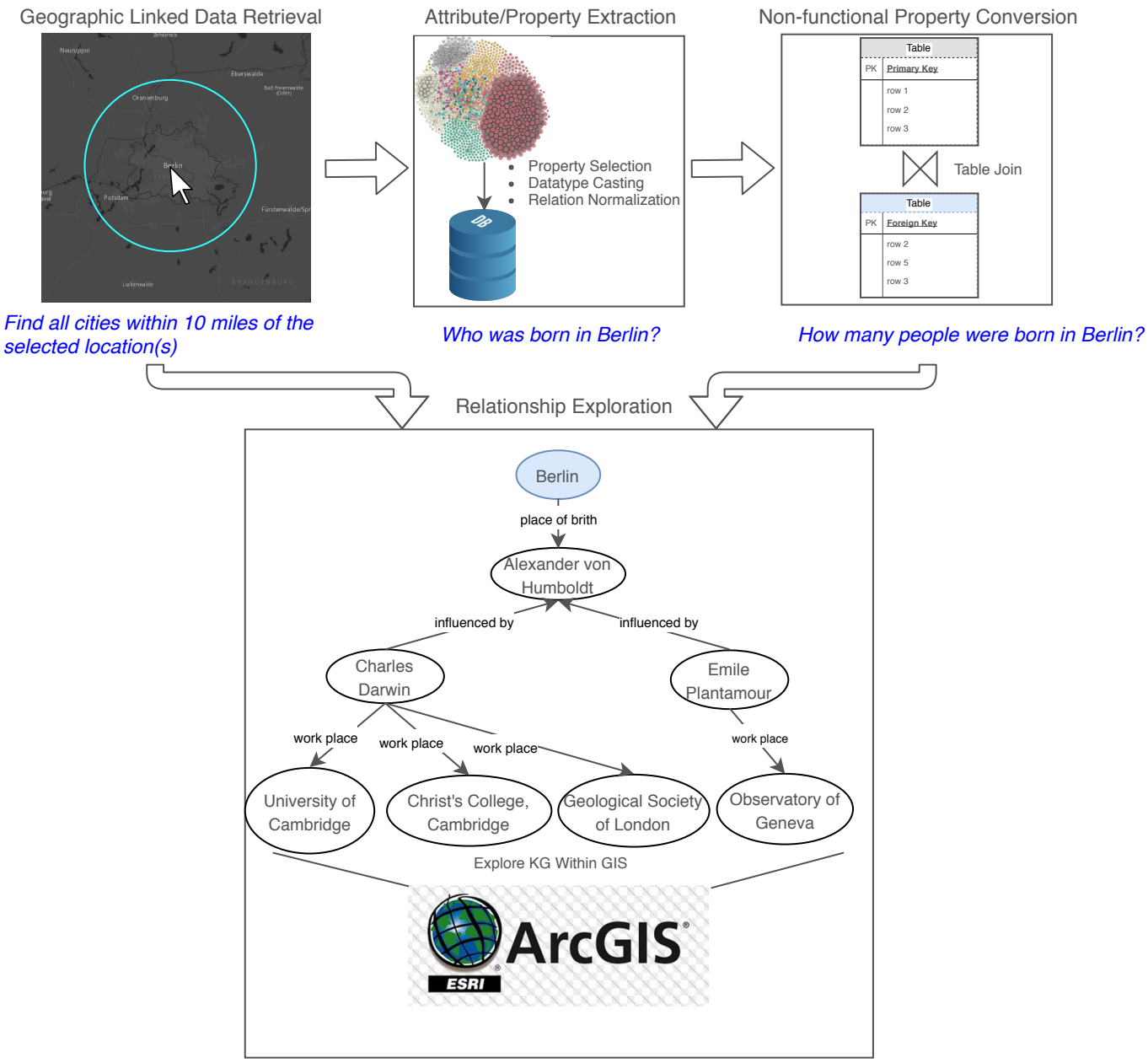


Figure made based on Esri and Ontotext images

Contribution

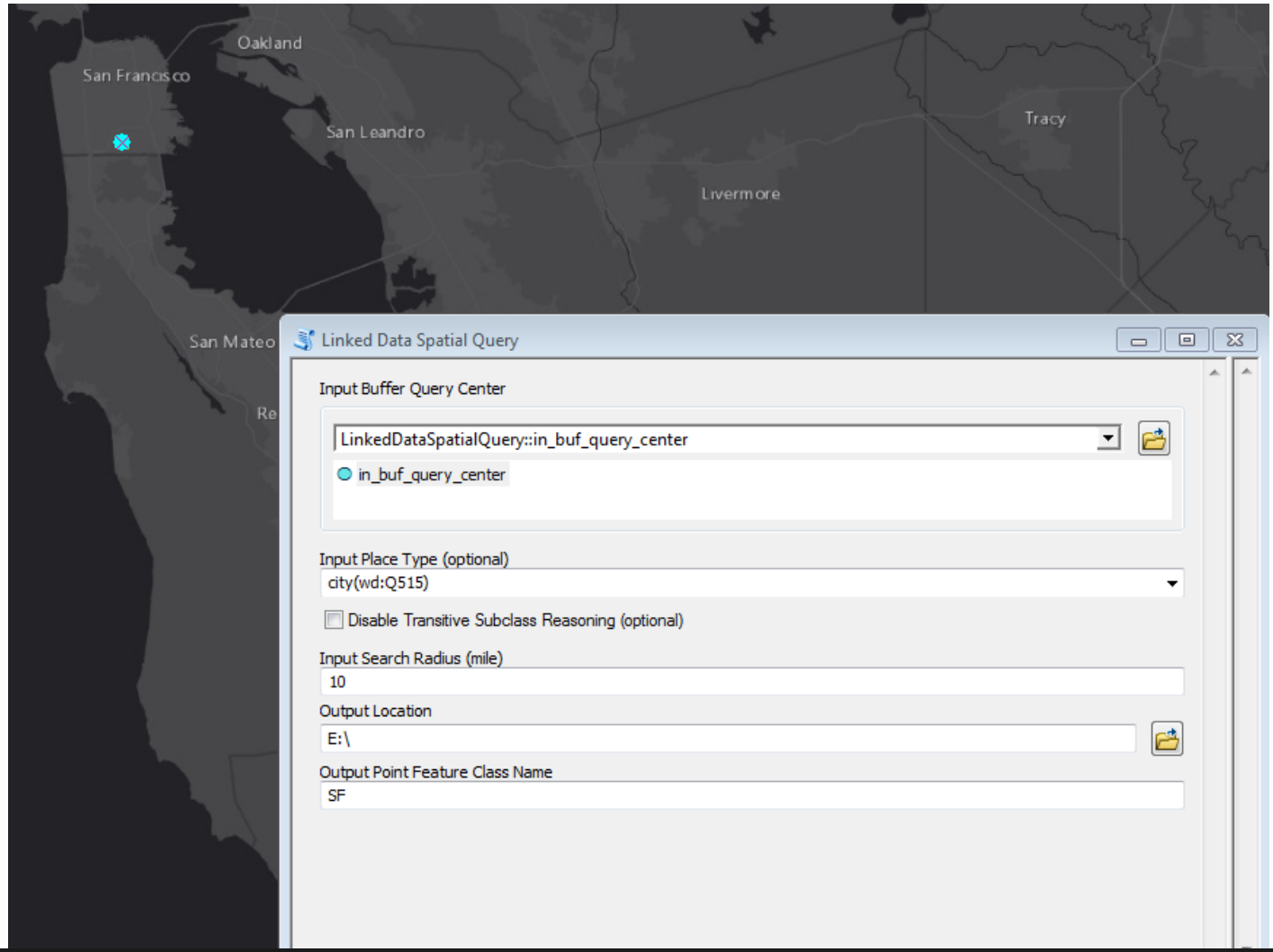
- We proposes **Linked Data Connector framework** - a deep integration of Linked Data into GIS.
- We demonstrate how to utilize **Semantic Web reasoning and ontologies** to extract additional properties (e.g., subsumption reasoning and (inverse) partonomical relations).
- Our deep integration supports **exploratory search** via **n-degree property path queries**, a feature that is not typically found in a GIS environment.
- We demonstrate how the queried Linked Data can be seamlessly used to perform **GIS analysis**.

Workflow


Geographic Linked Data Retrieval
Attribute/Property Extraction
Non-functional Property Conversion
Property Selection
Datatype Casting
Relation Normalization
DB
Table
PK Primary Key
row 1
row 2
row 3
Table Join
Table
PK Foreign Key
row 2
row 5
row 3
Find all cities within 10 miles of the selected location(s)
Who was born in Berlin?
How many people were born in Berlin?
Relationship Exploration
Berlin
place of brith
Alexander von Humboldt
influenced by
influenced by
Charles Darwin
Emile Plantamour
work place
work place
work place
work place
University of Cambridge
Christ's College, Cambridge
Geological Society of London
Observatory of Geneva
Explore KG Within GIS
ArcGIS
ESRI
Find the affiliations of people influenced by Alexandar von Humboldt?

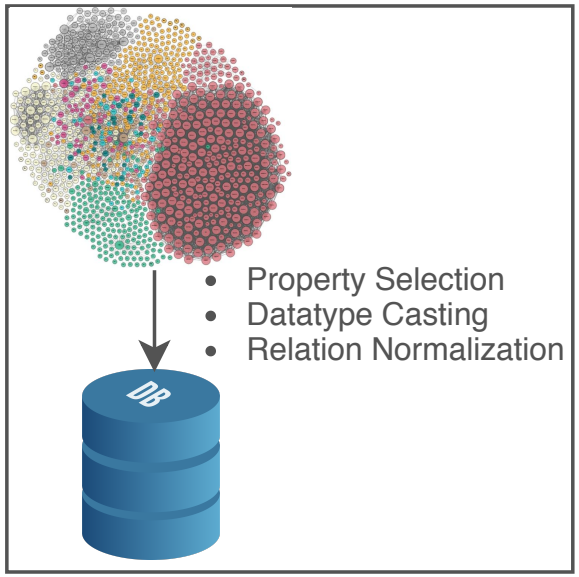
Geographic Linked Data Retrieval

Buffer Query: Directly interact with Geographic Linked Data through a SPARQL endpoint



Property Enrichment

- **Properties Selection:** Identify meaningful properties
- **Datatype Casting of Datatype Properties:** Linked Data typically do not restrict datatype properties to a specific XSD data type
- **Spatial Relation Normalization:** handling N-to-N relation/property

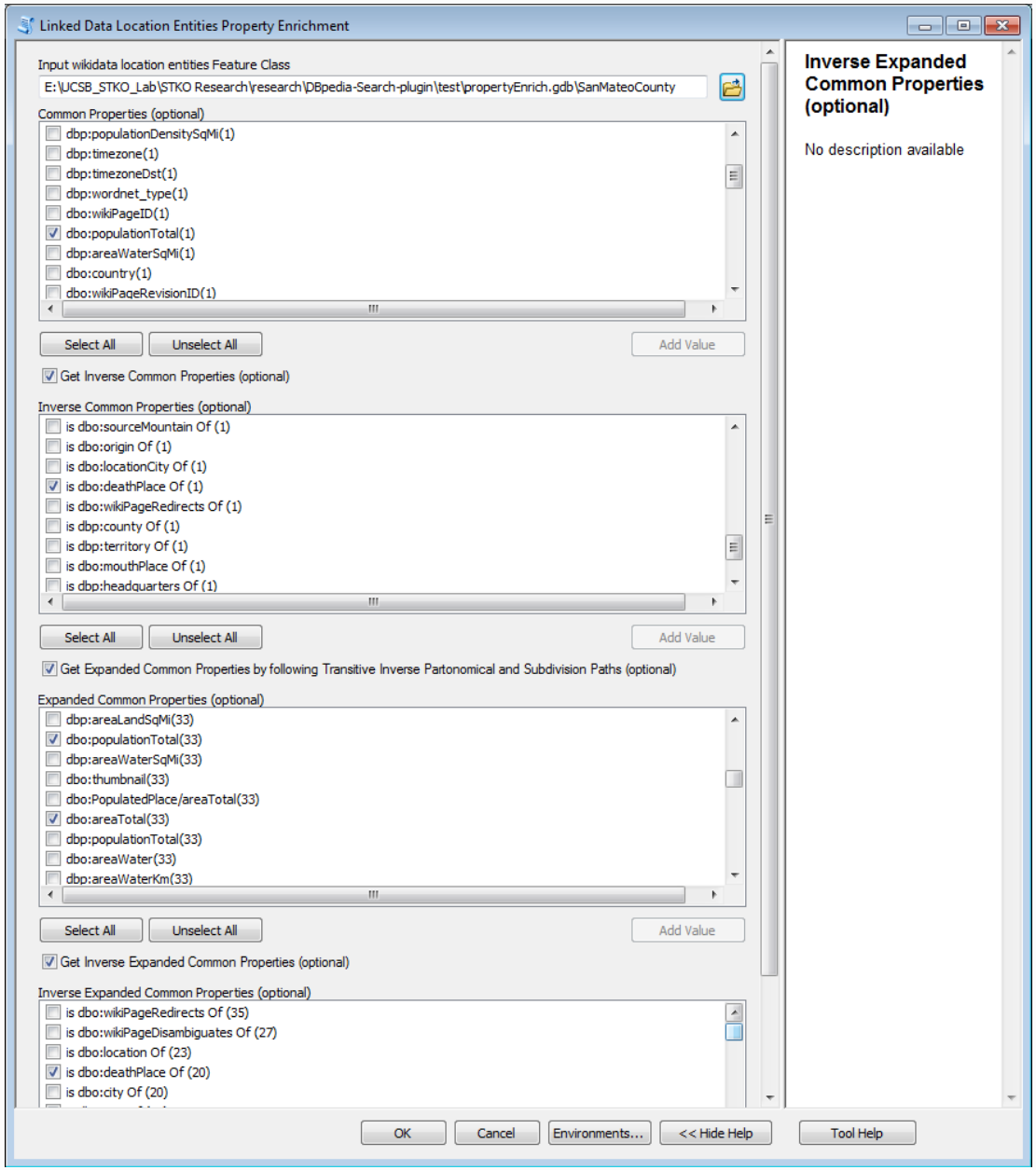


Who was born in Berlin?

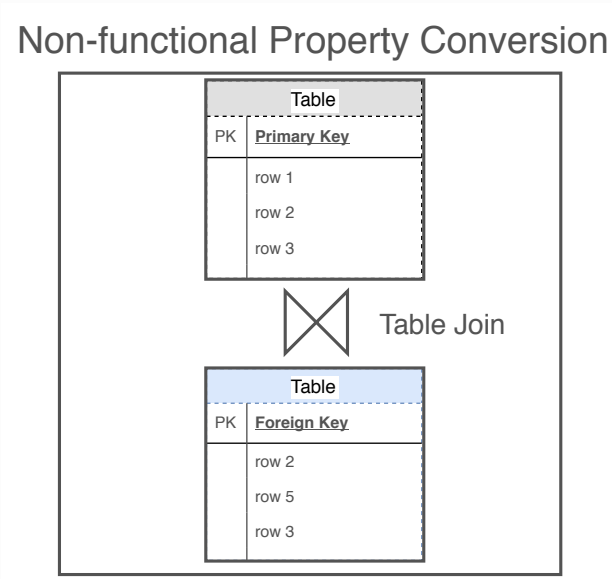
Property Enrichment

Get the **common properties** of these spatial entities and their **subdivisions**:

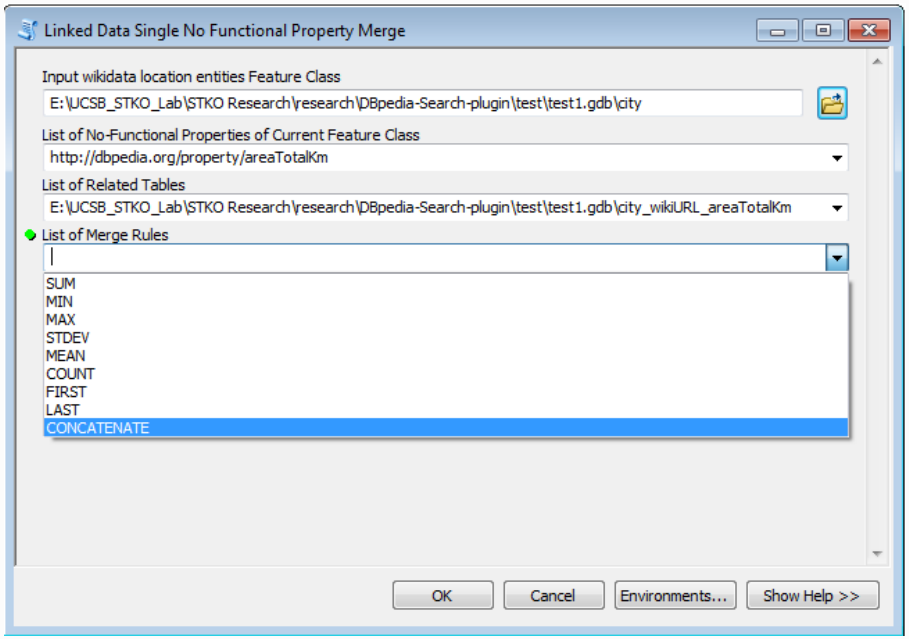
- Common properties
- Inverse common properties
- Expanded common properties
- Inverse expanded common properties



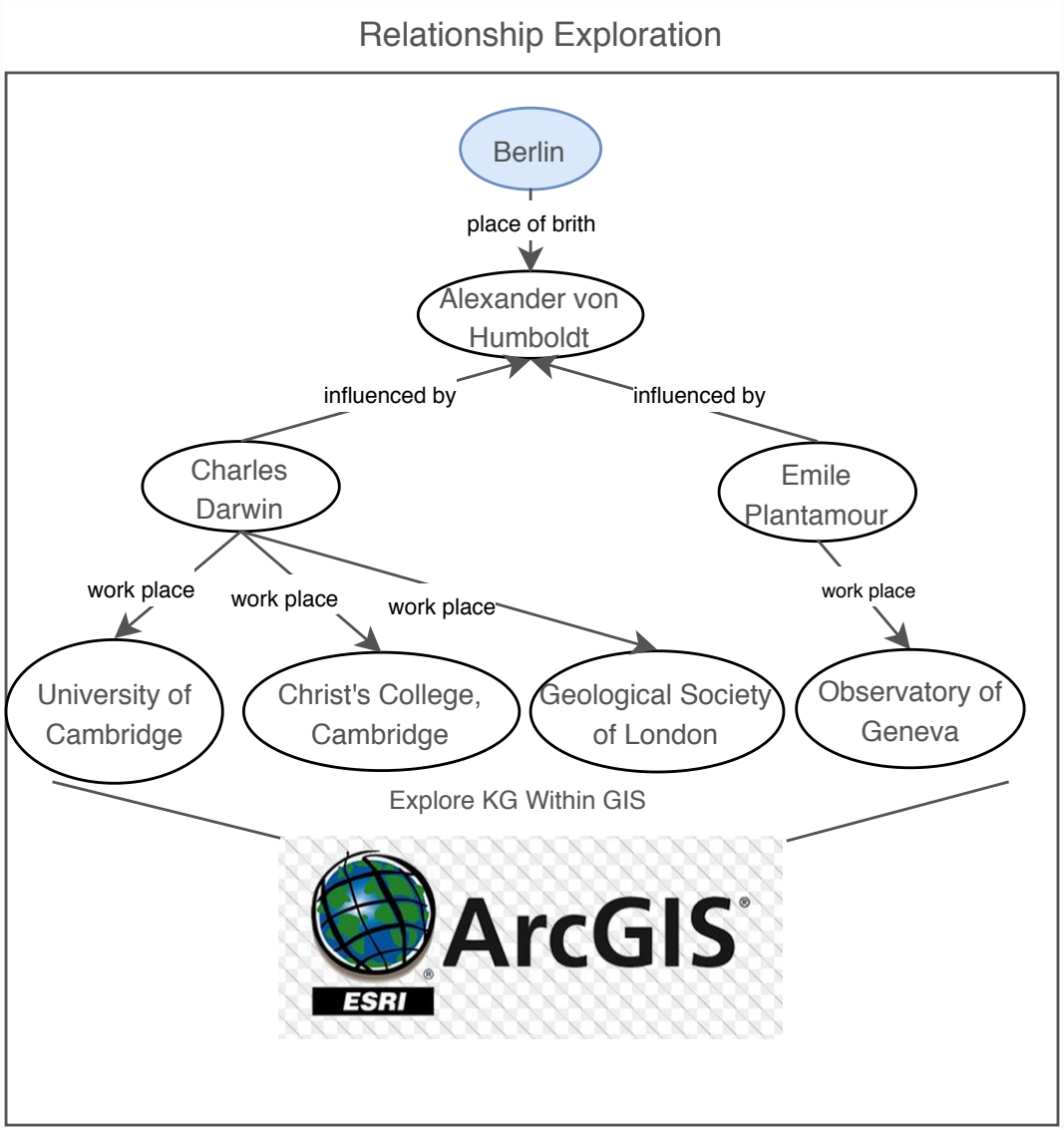
Non-functional Property Conversion



How many people were born in Berlin?



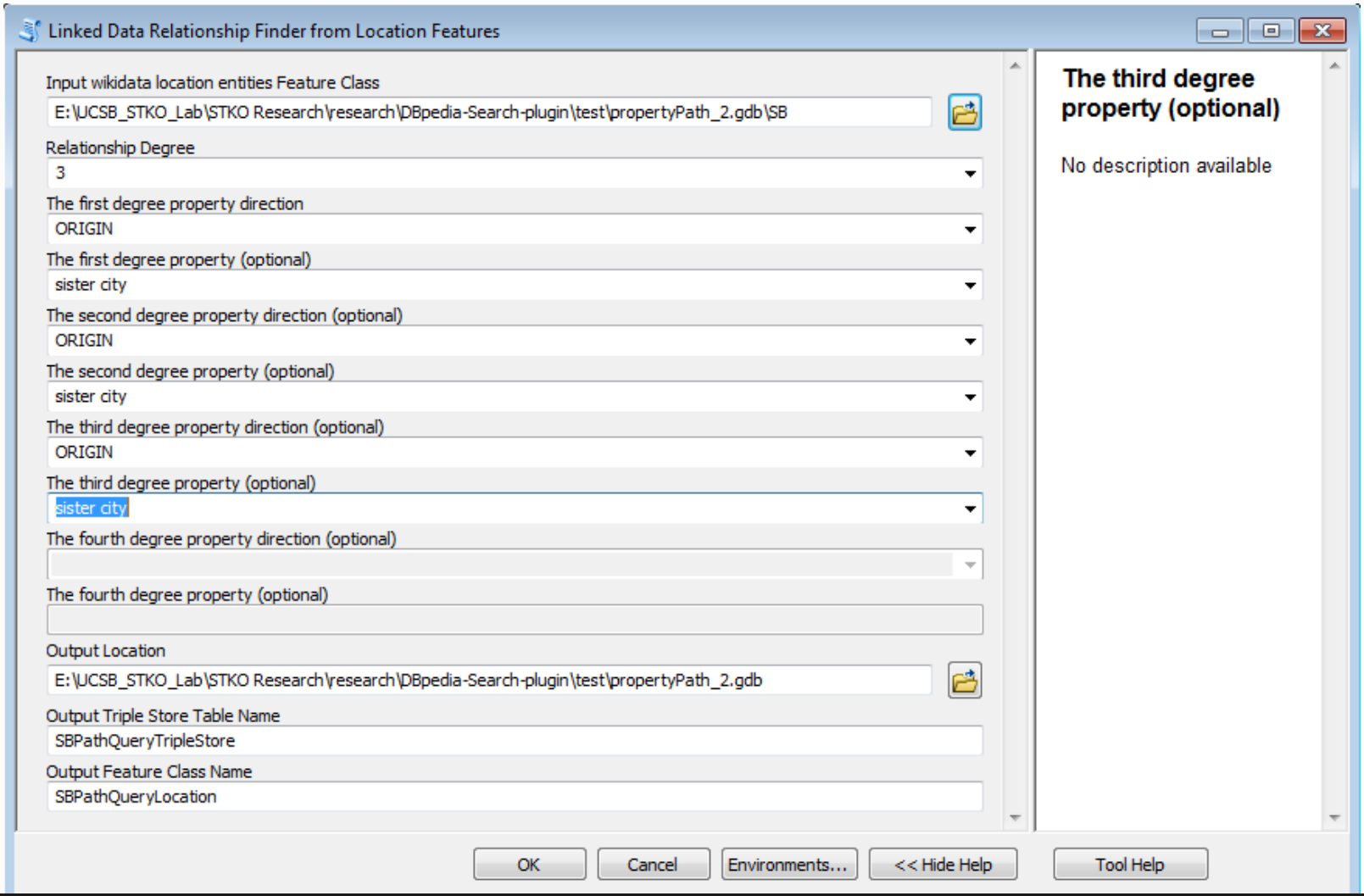
Relation Exploration



Find the affiliations of people influenced by Alexandar von Humboldt?

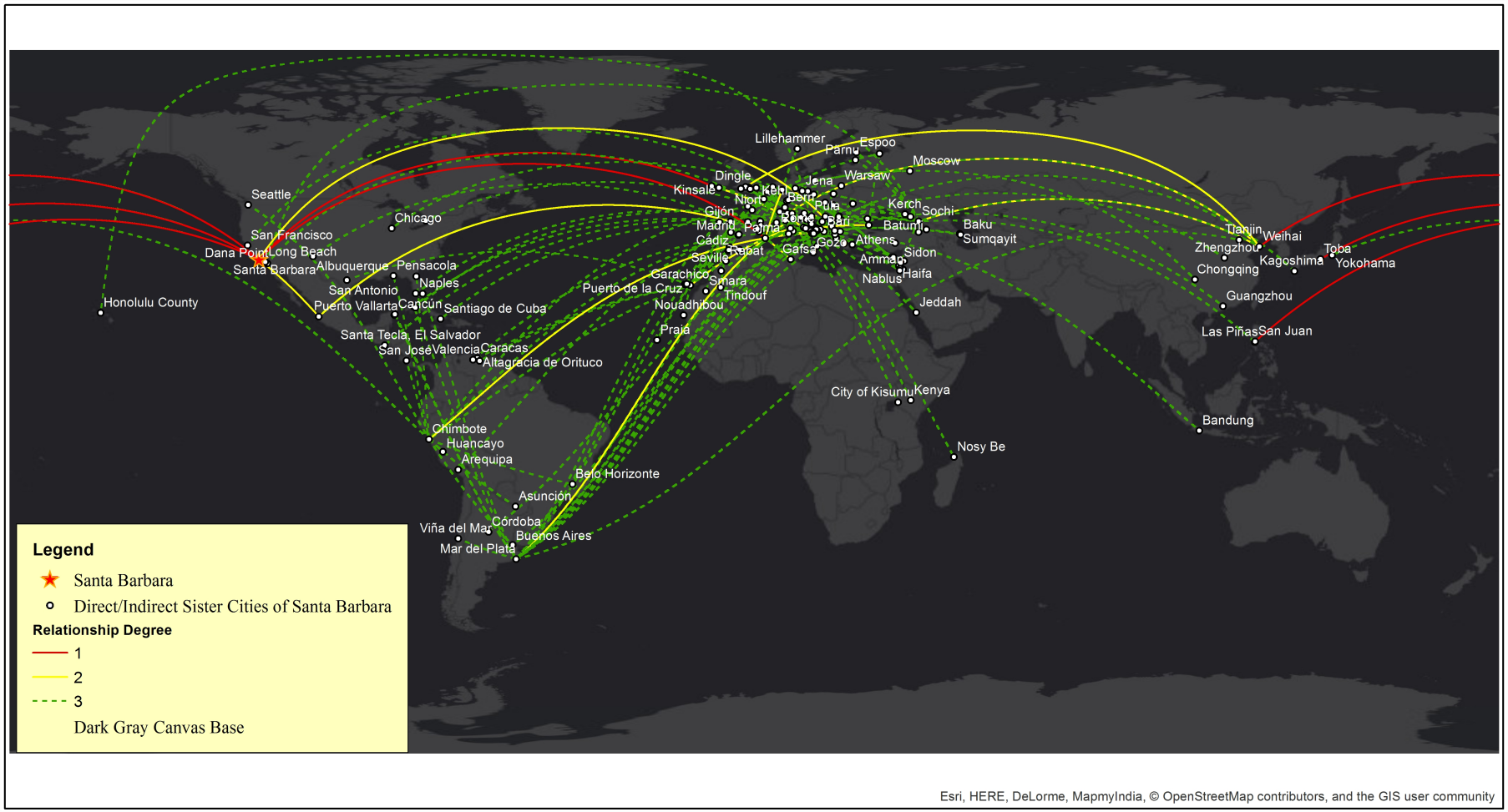
Relation Exploration

Query for 1-, 2-, and 3-degree sister city from the input spatial entities (Santa Barbara).



Relation Exploration

Query for 1-, 2-, and 3-degree sister city from the input spatial entities (Santa Barbara).

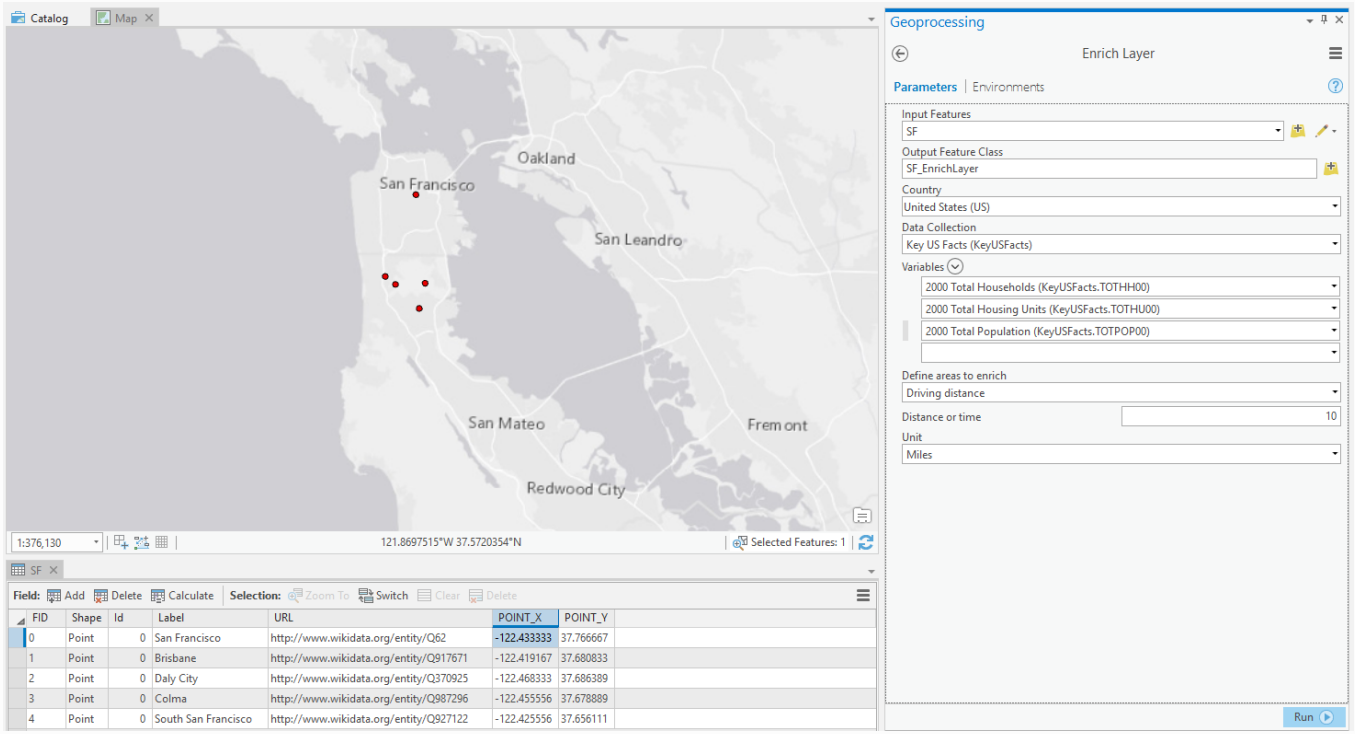


Linked Data Connector

- The presented Linked Data connector demonstrates a workflow for **geographic Linked Data retrieval**, **attribute enrichment & conversion**, and **linkage exploration** within a GIS.
- As far as we know, this is **the first work** about integrating Linked Data back to a GIS and making them ready for **spatial analysis** that does not simply flatten the data.
- The only system that supports a subset of the presented capabilities is the **GeoEnrichment**[2] service recently developed by Esri.

[2] https://developers.arcgis.com/rest/geoenrichment/api-reference/geoenrichment-service-overview.htm

Esri's GeoEnrichment



FID	Shape	Id	Label	URL	POINT_X	POINT_Y
0	Point	0	San Francisco	http://www.wikidata.org/entity/Q62	-122.433333	37.766667
1	Point	0	Brisbane	http://www.wikidata.org/entity/Q917671	-122.419167	37.680833
2	Point	0	Daly City	http://www.wikidata.org/entity/Q370925	-122.468333	37.686389
3	Point	0	Colma	http://www.wikidata.org/entity/Q987296	-122.455556	37.678889
4	Point	0	South San Francisco	http://www.wikidata.org/entity/Q927122	-122.425556	37.656111

The Enrich Layer toolbox in ArcGIS Pro.

OBJECTID	Shape	Id	Label	URL	POINT_X	POINT_Y	sourceCountry	ORIG_ID	areaType	bufferU	bufferUn	bufferRadii	2000 Total Households	2000 Total Housing Units	2000 Total Population
1	Point	0	San Francisco	http://www.wikidata.org/entity/Q62	-122.433333	37.766667	US	1	DriveTimeBuffer	esriMiles	Miles	10	388675	407068	958993
2	Point	0	Brisbane	http://www.wikidata.org/entity/Q917671	-122.419167	37.680833	US	2	DriveTimeBuffer	esriMiles	Miles	10	403067	420446	1009438
3	Point	0	Daly City	http://www.wikidata.org/entity/Q370925	-122.468333	37.686389	US	3	DriveTimeBuffer	esriMiles	Miles	10	425186	443492	1061784
4	Point	0	Colma	http://www.wikidata.org/entity/Q987296	-122.455556	37.678889	US	4	DriveTimeBuffer	esriMiles	Miles	10	416617	433722	1046568
5	Point	0	South San Francisco	http://www.wikidata.org/entity/Q927122	-122.425556	37.656111	US	5	DriveTimeBuffer	esriMiles	Miles	10	275948	283756	777603

The enriched attribute table after executing the Enrich Layer toolbox.

Comparison

Comparing GeoEnrichment Service with Linked Data connector:

- **Well-curated dataset** v.s. distributed, Web-scale, real-time **knowledge graph**
- **Points-driving time zones-attributes** v.s. **points-spatial entities-properties**
- The GeoEnrichment service supports datatype properties but not **object type properties** such as those linking actors, places, events, and objects together.
- From a question answering perspective, our Linked Data connector framework is more suitable to **answer geographic queries** such as about the oldest mission along California's coast.

Conclusion

- We proposed and implemented a **Linked Data connector workflow** to deeply integrate Linked Data and GIS without simply flattening the retrieved data.
- Our system constantly creates new tables and schema for them, thereby enabling users to **truly navigate the link structure** of knowledge graphs and to query each node's **datatype and object-type properties** on-the-fly.
- We compare our framework with GeoEnrichment and show that both approaches are **complementary**.

Future Work

- We see the presented research as a starting point towards a more **question answering oriented view** on GIS.
- We have not addressed issues of **scalability**, **user feedback and exception handling**, **provenance records**, and so forth, but believe that they will be important steps towards a deployable add-on to GIS systems.