Constructing a **Knowledge Graph** of **Historical Mining Data**

Basel Shbita, Namrata Sharma, Binh Vu,

Fandel Lin & Craig A. Knoblock

USC Information Sciences Institute

6th International Workshop on Geospatial Linked Data (GeoLD)

co-located with the 21st Extended Semantic Web Conference (ESWC 2024)

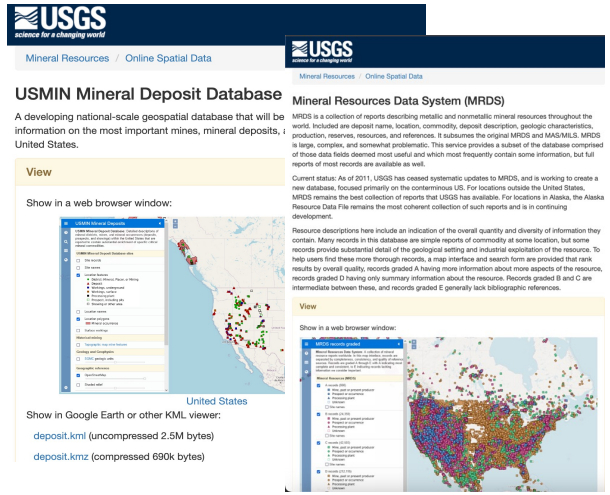
05/26/2024

Agenda

- Intro ⟸
- Problem
- Approach
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work
- Future Directions
- Conclusions

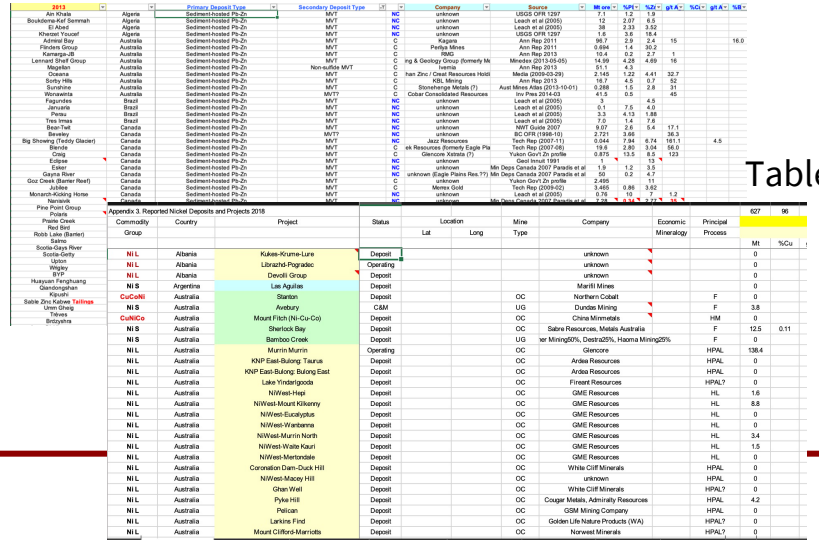
Historical & Geo Data

- Rich sources of information
  - understanding human & environmental systems
  - describing human & natural activities
- Labor-intensive to analyze across time & space
  - e.g., economic viability, physical changer, geo-related characteristics
- Often require grounding & additional contextual information
  - e.g., demographics, geology, stratigraphy, etc…

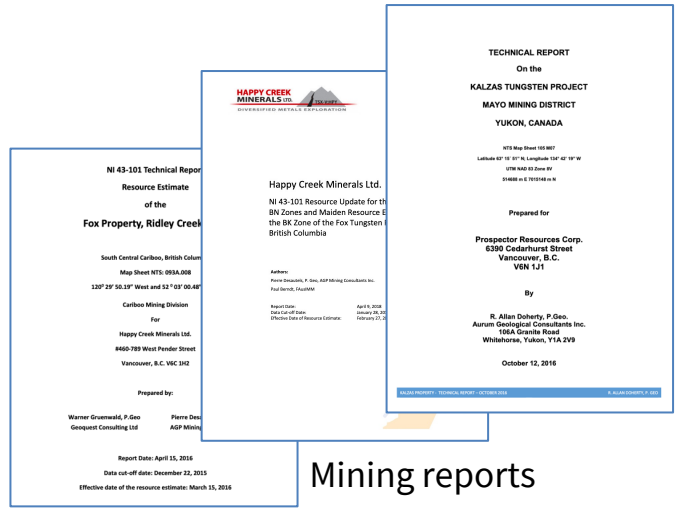


Geospatial databases (e.g., mining sites)

Web

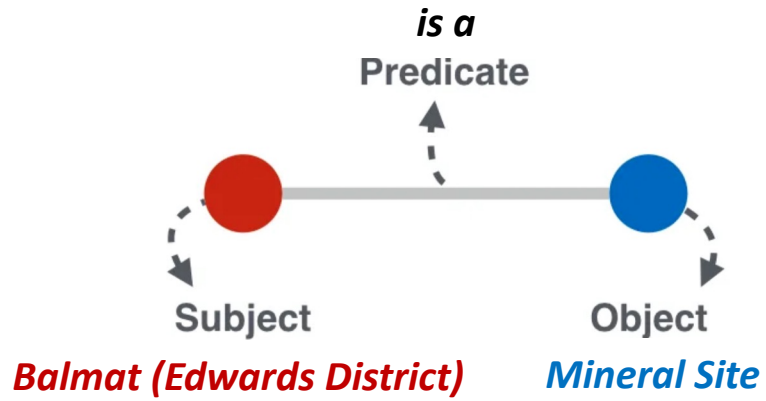


Tables with geo-data



Mining reports

KGs

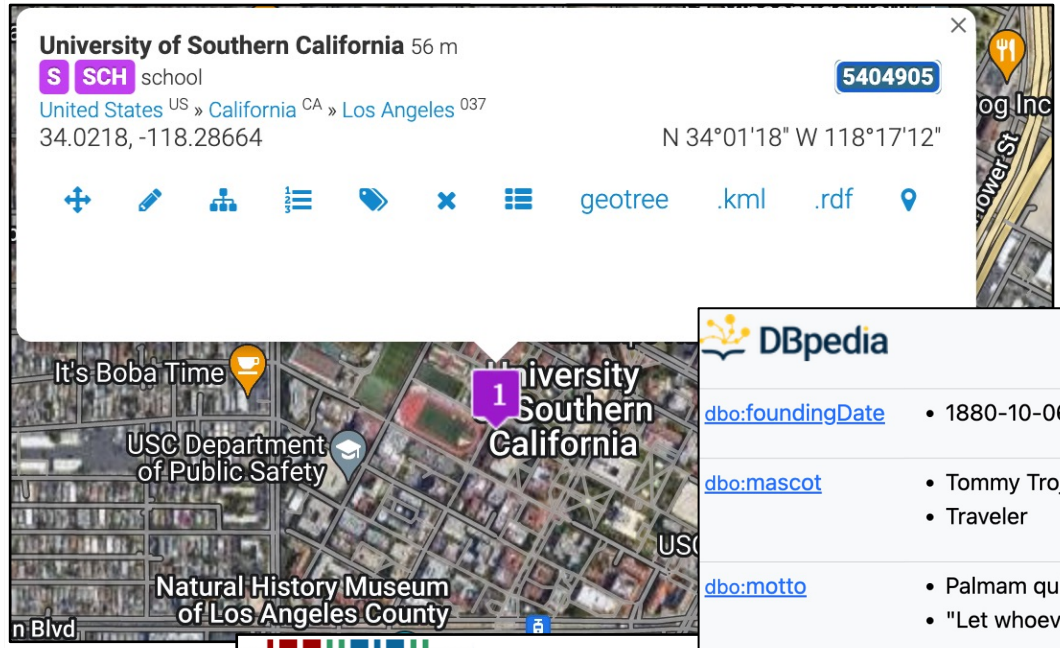


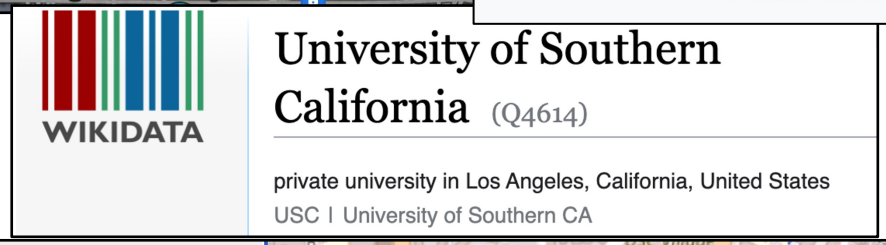
- Knowledge Graph (KG)?
  - Graphs are natural way to encode data
  - Using semantic concepts & relationships
    - Semantic Network = **Knowledge Graph**
- Why use KG?
  - Combine expressivity, interoperability, & standardization in the Semantic Web stack
- Semantic Web?
  - Extension of WWW, enabling the Web of Data (aka "Linked Data")
  - Encoding of semantics with the data
  - Linked Open Data principles // FAIR

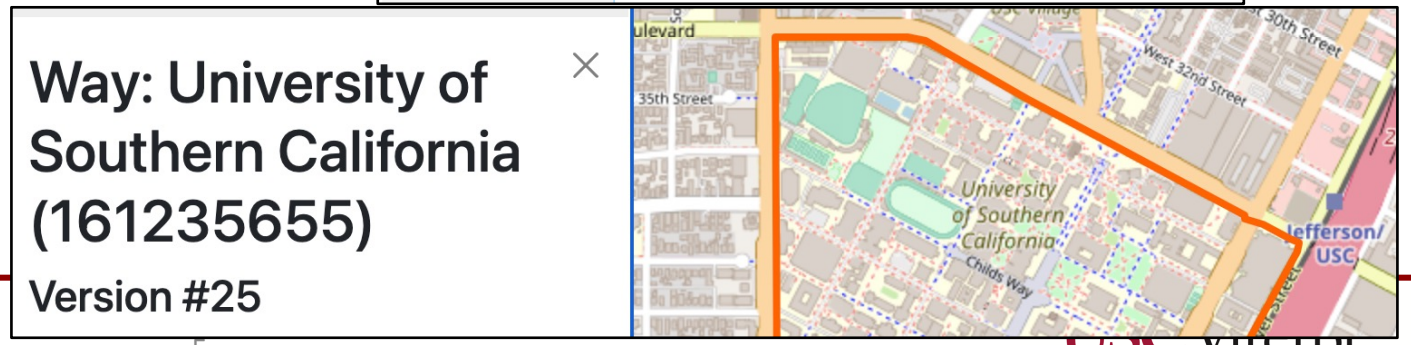
Geo & Spatio-Temporal KGs

- Spatio-Temporal KGs
  - Contextual (what)
  - **Spatial** (where)
  - **Temporal** (when)

- Geo-semantics
  - Representation, annotation, & reasoning
  - Modeling & ontology development
  - Integration & interoperability

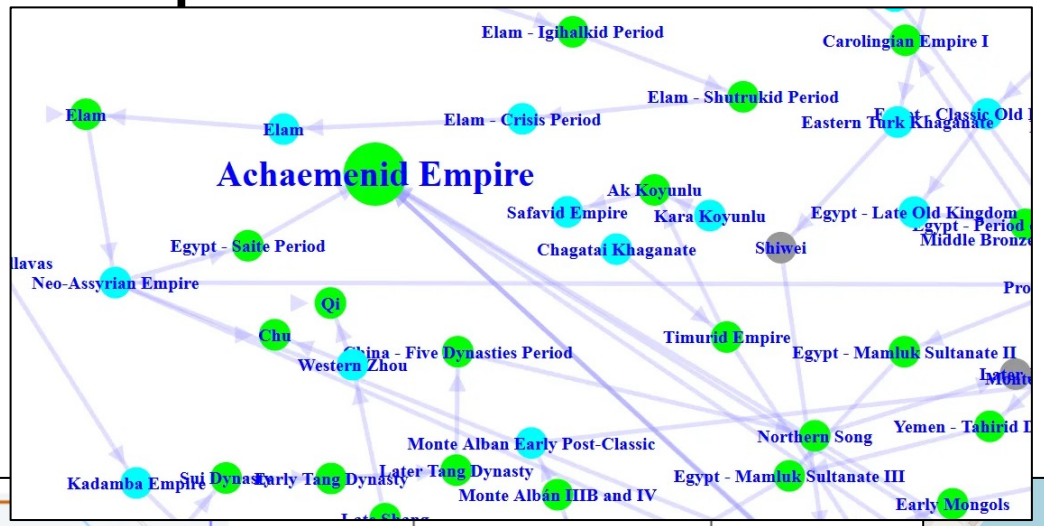


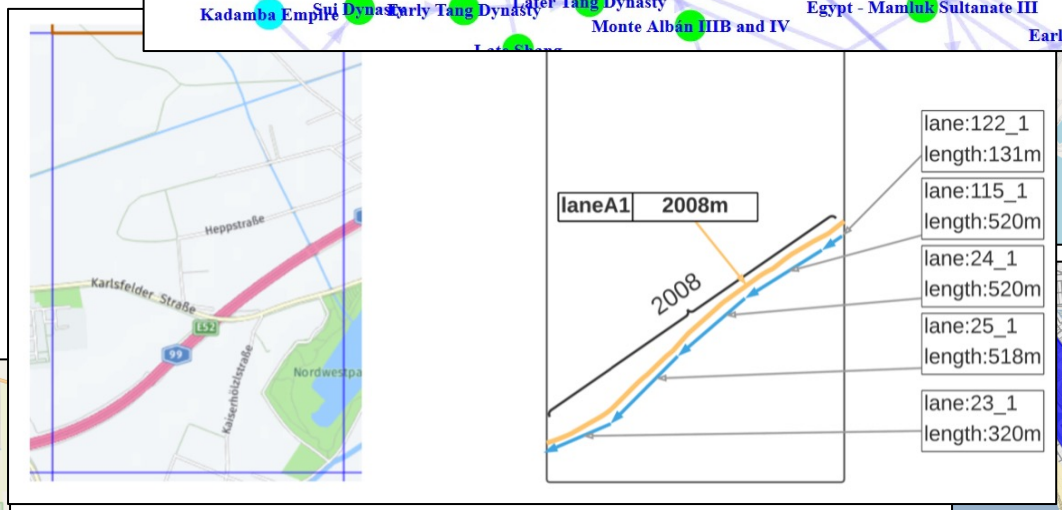




Intro: Spatio-Temporal KGs

- So, what's so special about them?
  - Spatial analysis
  - Temporal analysis
  - Spatio-temporal aggregations
  - Geographic QA
  - Environmental & social science
  - Urban planning
  - Transportation
  - etc...





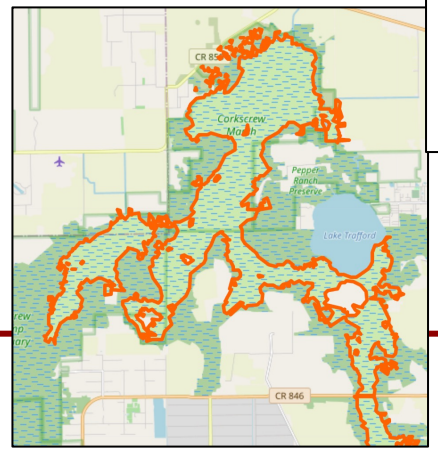


figure from *Szwoch, G. (2019). Combining road network data from openstreetmap with an authoritative database. Journal of Transportation Engineering, Part A: Systems, 145(2), 04018085.*

figure from *https://terminusdb.com/blog/human-history-knowledge-graph/*

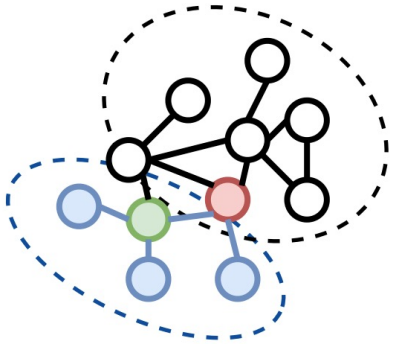
Agenda

- Intro
- Problem ⟸
- Approach
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work
- Future Directions
- Conclusions

Research Problem

- How can we transform & link unstructured digitized & historical geo-data into structured, semantic, & queryable spatio-temporal KGs?

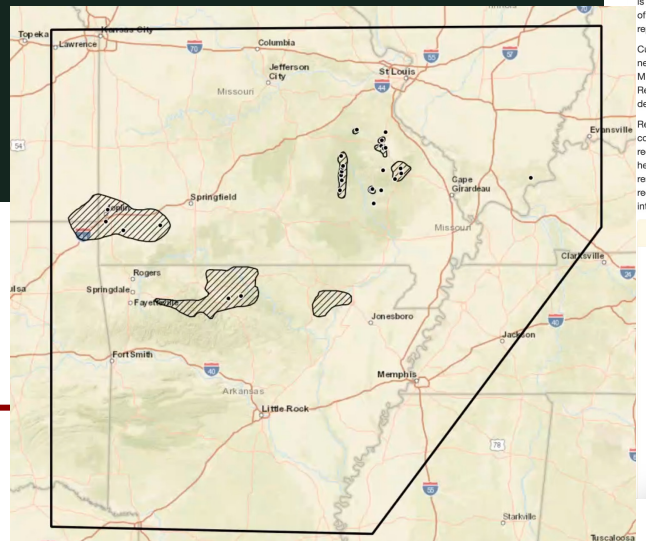
- Objectives:
  - Auto KG construction & entity resolution (ER) from various historical geo-data sources
  - Semantic enrichment by linking (EL) to additional sources on the web
  - Adherence to Semantic Web principles
    - shared, accessible, visualized, standardized across-domains, & scalable for easy use by downstream tasks for easy analysis & expressive integration

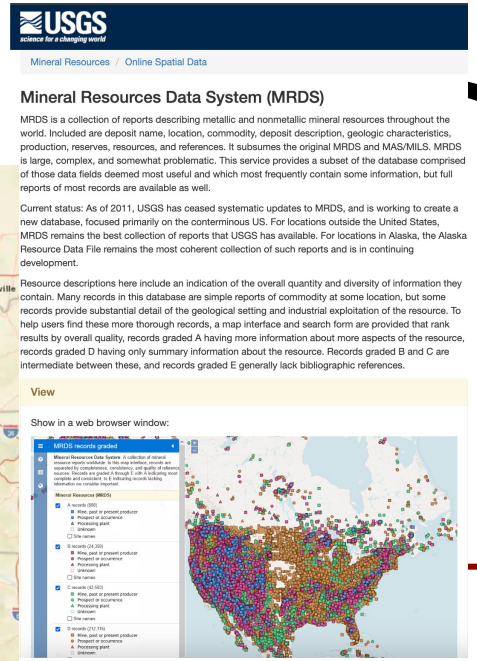


Goals

- Integrating geo-referenced textual & historical data with quantitative information into a comprehensive, dynamic, & spatio-temporal KG
  - capture data & entity semantics, entity resolution, & accurate data modeling
- Demonstrate via a KG of historical mining data
  - Historically takes months of work; geologists describe it as a “soul crushing exercise”

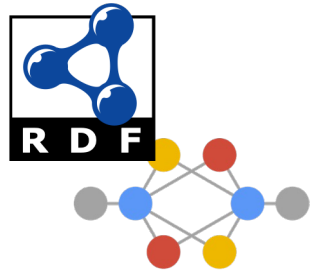

```
{
  "MineralSite": [
    {
      "source_id": "https://w3id.org/usgs/z/4530692/2P29BJHV",
      "record_id": 1,
      "name": "NI 43-101 Technical Report for the Lantinen Koillismaa Project in Europe, Finland dated
March 2017",
      "location_info": {
        "location": "POINT(28.17472 65.94611)",
        "crs": "WGS84",
        "country": "Finland",
        "state_or_province": "Central Finland"
      },
```



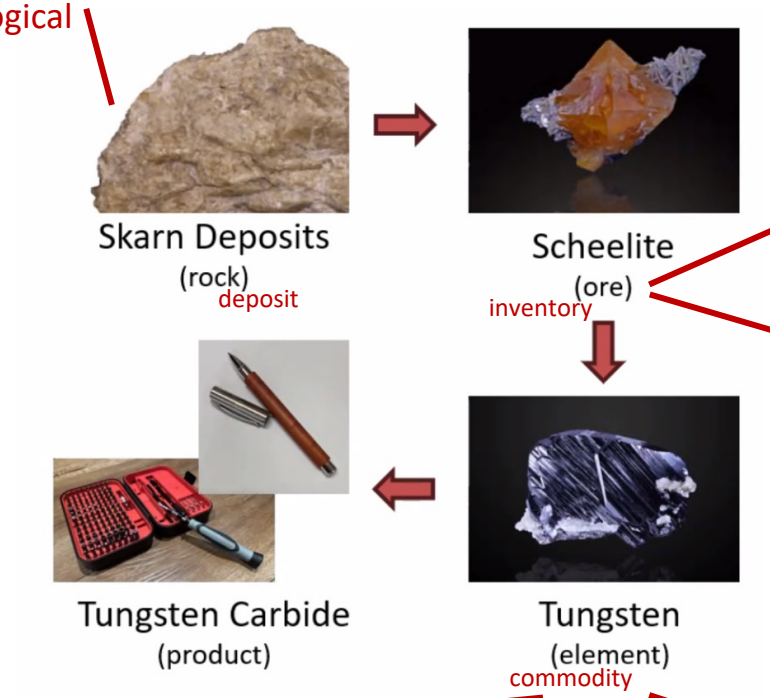
Extracted Textual & Vector Data

Data understanding & KG construction

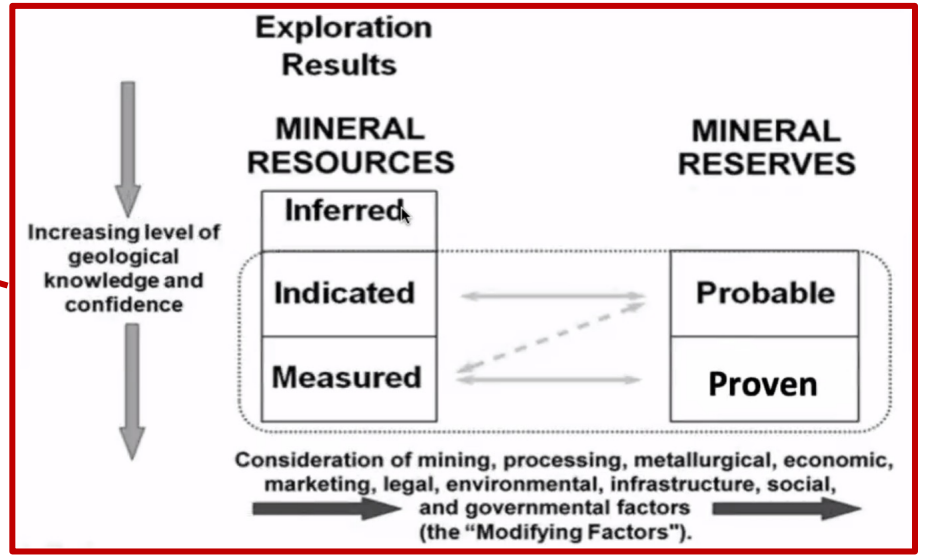


Transformation & Aggregation

natural occurrence of minerals (geological formations)



naturally occurring solid material
usually mixtures of minerals and other materials
(bauxite = ore of aluminum; hematite = ore of iron)

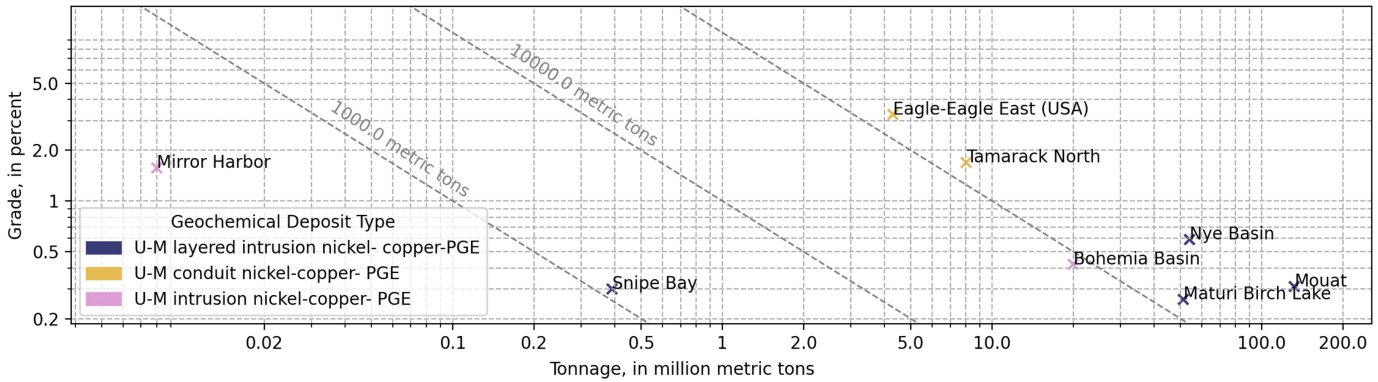


COMMODITY	AS-REPORTED FO	AS-REPORTED AI	CONVERSION FA	CONVERTED FORM
Aluminum	Al		1.8895	Al2O3
Antimony	Sb		1.1971	Sb2O3
Barium	Ba		1.6995	BaSo4
Borates	B	Boron	3.2198	B2O3
Borates	H3BO3	Boric Acid	0.5629	B2O3
Cesium	Cs2O	Cesium Oxide, C	0.9432	Cs
Chromium	Cr		1.4615	Cr2O3
Cobalt	CoOH	Cobalt Hydroxid	0.776	Co
Iron Ore	Fe2O3	Hematite, Haem	0.6994	Fe

material or primary agricultural product

One desired output: Grade-Tonnage model

- For a given commodity/deposit type/location/time-range:
  - Construct grade and tonnage models from the data on existing mines
  - Compile rich mineral site data



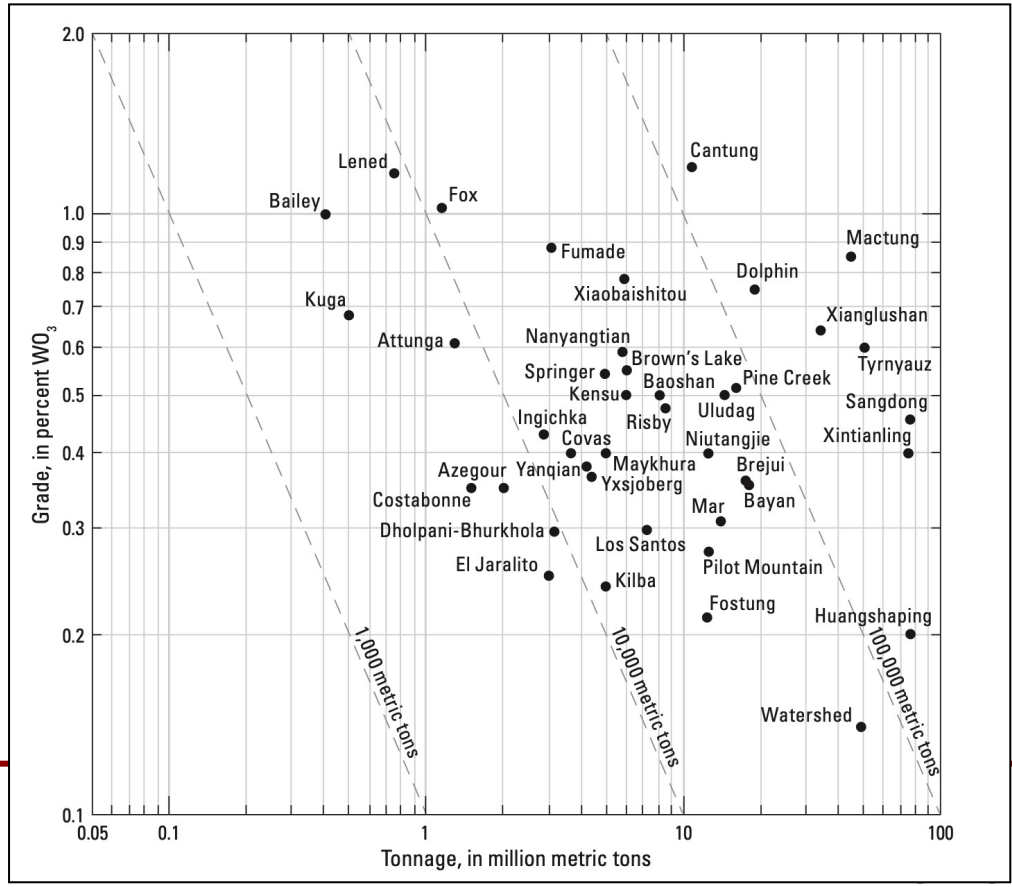


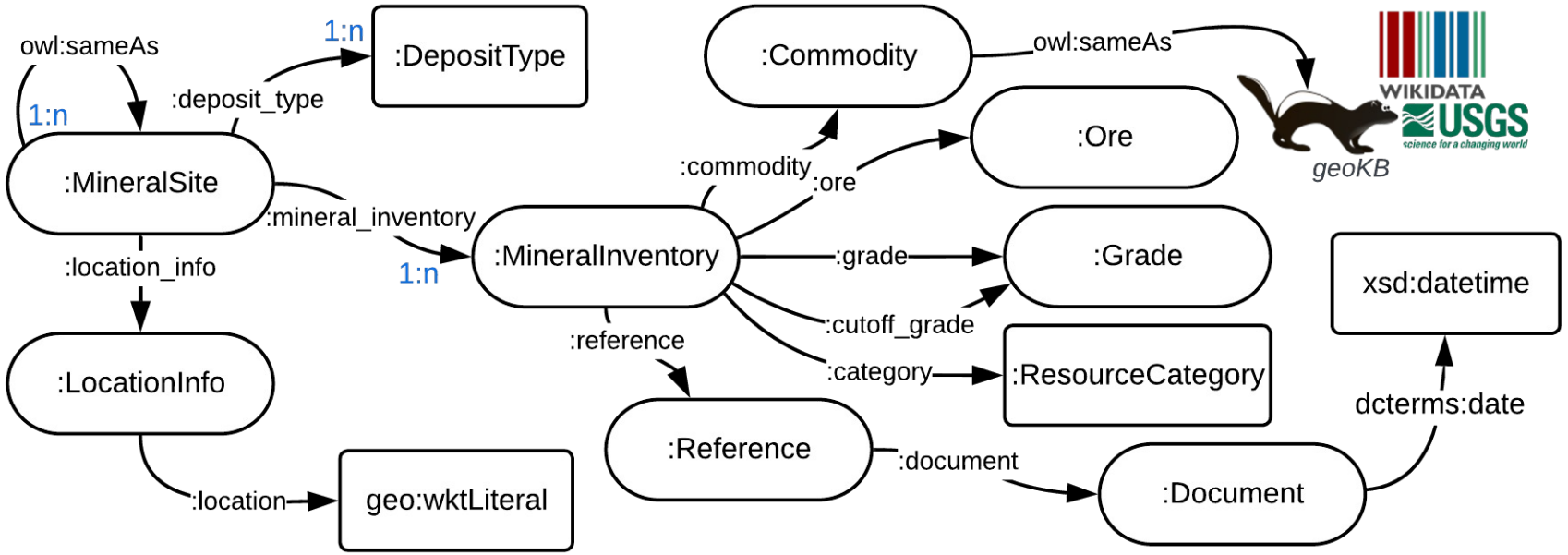
Figure from Green, C. J., Lederer, G. W., Parks, H. L., & Zientek, M. L. (2020). Grade and tonnage model for tungsten skarn deposits—2020 update (No. 2020-5085). US Geological Survey

Agenda

- Intro
- Problem
- Approach ⟸
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work
- Future Directions
- Conclusions

Approach

- Step 1. Semantic Modeling & URI assignment (Data Representation)
  - Transform & materialize the data (construct KG)
    - Generate entities (URIs) based on unique identifiers
    - Provide a useful semantic representation supporting downstream tasks
  - Construct a meaningful semantic model
    - Follows W3C & OGC standards (GeoSPARQL)



GeoKB as a Target KB for EL

GEOKB

Nickel (Q162561)

mineral species in the Iron Group sourced from Mindat and the Geoscience Ontology

Niccolum | nikle | Nikel | نیکل | Nikelo | ニッケル | Nikal | Nichel | Nikiel | Nikl | IMA1966-039 | 니켈 | Niķelis | Nikkeli | Նիկել | Níquel | Nikelj | निकेल | ნიკელი | ನಿಕಲ್ | Никел | Nikèl | Никĕль | Nîkel | Iztāctepoztli | निकेल | നിക്കൽ | نیکل | Niquèl | Никл | ניקל | Nickyl | நிக்கல் | Нікель | Nikeli | Nikil | Nikelis | Νικέλιο | Néckel | Никель | Konukōreko | Niken | Nichele | 自然镍 | Nikkel

Statements

subclass of	01.AA.05 - Copper-cupalite family 2 references
	mineral material 1 reference
same as	http://www.mindat.org/min-2895.html 0 references
	https://w3id.org/gso/mineral/nickel 0 references

has chemical element	nickel 1 reference
member of	Iron Group 1 reference

Approach – cont'd

- Step 2. Entity Linking
  - Link the generated entities to a domain data-rich vocab (i.e., GeoKB)
    - Determine similarity by textual similarity (i.e., Jaccard)
    - Directly within SPARQL

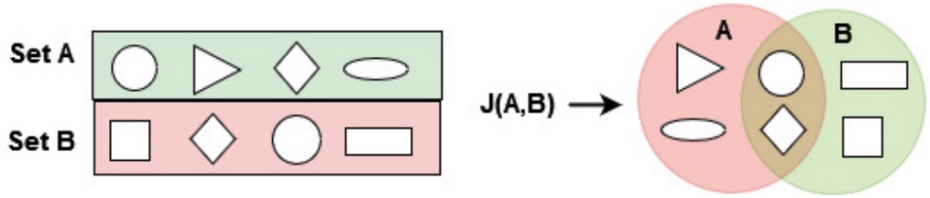
```
"Clay, Fire (Refractory)"
```

		entity.value
0	uri	https://geokb.wikibase.cloud/entity/Q413
1	uri	https://geokb.wikibase.cloud/entity/Q424
2	uri	https://geokb.wikibase.cloud/entity/Q423
3	uri	https://geokb.wikibase.cloud/entity/Q421
4	uri	https://geokb.wikibase.cloud/entity/Q162319
5	uri	https://geokb.wikibase.cloud/entity/Q425
6	uri	https://geokb.wikibase.cloud/entity/Q426
7	uri	https://geokb.wikibase.cloud/entity/Q428
8	uri	https://geokb.wikibase.cloud/entity/Q427
9	uri	https://geokb.wikibase.cloud/entity/Q429

	entityLabel.xml:lang	entityLabel.type	entityLabel.value
0	en	literal	high alumina clay aluminum
1	en	literal	bloating material clay
2	en	literal	brick clay
3	en	literal	clay
4	en	literal	Clay
5	en	literal	bentonite clay
6	en	literal	chlorite clay
7	en	literal	fire (refractory) clay
8	en	literal	fullers earth clay
9	en	literal	glauconite clay

```
1  SELECT ?entity ?entityLabel WHERE {
2    ?entity rdfs:label ?entityLabel.
3    ?entity gkbt:P1 gkbi:Q406. # instance of mineral commodity
4    FILTER(CONTAINS(LCASE(?entityLabel), "nickel")) }
```

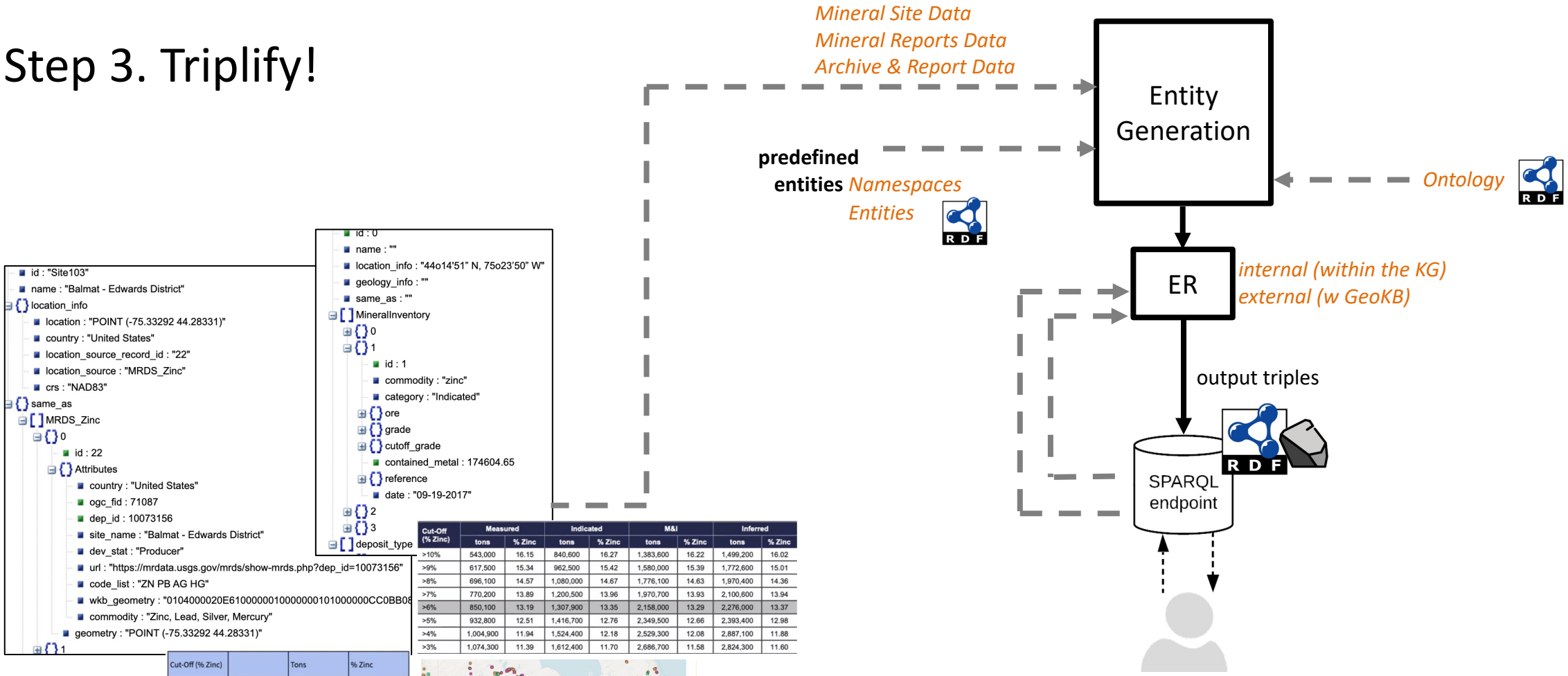
Jaccard



USC Viterbi School of Engineering

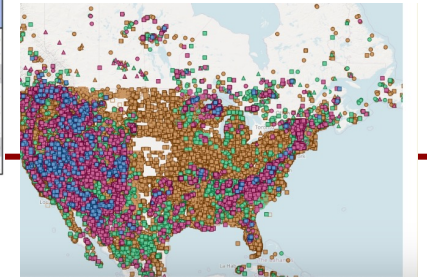
Approach - cont'd

- Step 3. Triplify!



Cut-Off (% Zinc)	Measured		Indicated		M&I		Inferred	
	tons	% Zinc	tons	% Zinc	tons	% Zinc	tons	% Zinc
>10%	543,000	16.15	840,600	16.27	1,383,600	16.22	1,499,200	16.02
>9%	617,500	15.34	962,500	15.42	1,580,000	15.39	1,772,600	15.01
>8%	696,100	14.57	1,080,000	14.67	1,776,100	14.63	1,970,400	14.36
>7%	770,200	13.89	1,200,500	13.96	1,970,700	13.93	2,100,600	13.94
>6%	850,100	13.19	1,307,900	13.35	2,158,000	13.29	2,276,000	13.37
>5%	932,800	12.51	1,416,700	12.76	2,349,500	12.66	2,393,400	12.98
>4%	1,004,900	11.94	1,524,400	12.18	2,529,300	12.08	2,887,100	11.88
>3%	1,074,300	11.39	1,612,400	11.70	2,686,700	11.58	2,824,300	11.60

Cut-Off (% Zinc)		Tons	% Zinc
>10%	Measured	543,000	16.15
	Indicated	840,600	16.27
	M&I	1,383,600	16.22
	Inferred	1,499,200	16.02
>9%	Measured	617,500	15.34
	Indicated	962,500	15.42
	M&I	1,580,000	15.39
	Inferred	1,772,600	15.01



Agenda

- Intro
- Problem
- Approach
- Evaluation & Discussion ⟸
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work
- Future Directions
- Conclusions

Evaluation & Discussion

- Data completeness (SHACL)
- Entity linking
- Query performance

Characteristic	Count
Total Triples	2,397,708
Distinct Classes	16
Instances (Non-literals)	226,267
Geospatial Instances	2,884
Blank Nodes	1,518,981

Data: 2.4m triples // 135 commodities // focus on 2 critical mineral: nickel, zinc

Method	MRR	Hits@1	Hits@3	Hits@5
String search, then Jaro	0.557	0.459	0.659	0.659
String search, then Jaccard	0.648	0.637	0.659	0.659
Instance search, then Jaro	0.801	0.689	0.926	0.956
Instance search, then Jaccard (proposed)	**0.940**	**0.904**	**0.978**	**0.978**

```
1 ?ms :location_info/:location ?loc_wkt .
2 FILTER(geof:distance(?loc_wkt, "POINT(-118.57 47.56)"^^geo:wktLiteral, unit:mile) < 500)
```

Query Constraint Type	Avg	Min	Max
Textual	450	369	649
Temporal/Numeric	438	388	607
Spatial	708	501	811

Agenda

- Intro
- Problem
- Approach
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL ⟸
- Related Work
- Future Directions
- Conclusions

Spatio-Temporal Analysis via GeoSPARQL

Grade	Assays					
	Cu (%)	Ni (%)	S (%)	Au (g/t)	Pt (g/t)	Pd (g/t)
Concentrate	7.16-10.1	1.66-2.20	18.4-21.5	0.65-1.28	1.17-1.59	5.76-6.71

2nd aggregation: **total** grade & tonnage computation

```
SELECT
  ?ms ?ms_name ?deposit_name ?loc_wkt ?total_tonnage_measured ?total_tonnage_indicated ?total_tonnage_inferred ?total_contained_measured
      ?total_contained_indicated ?total_contained_inferred
      (?total_tonnage_measured + ?total_tonnage_indicated + ?total_tonnage_inferred AS ?total_tonnage)
      (?total_contained_measured + ?total_contained_indicated + ?total_contained_inferred AS ?total_contained_metal)
      (IF(?total_tonnage > 0, ?total_contained_metal / ?total_tonnage, 0) AS ?total_grade)
WHERE {
  {
    SELECT ?ms ?ms_name ?deposit_name ?country ?loc_wkt
           (SUM(?tonnage_measured) AS ?total_tonnage_measured)
           (SUM(?tonnage_indicated) AS ?total_tonnage_indicated)
           (SUM(?tonnage_inferred) AS ?total_tonnage_inferred)
           (SUM(?contained_measured) AS ?total_contained_measured)
           (SUM(?contained_indicated) AS ?total_contained_indicated)
           (SUM(?contained_inferred) AS ?total_contained_inferred)
    WHERE {
        ?ms :deposit_type [ rdfs:label ?deposit_name ] .
        ?ms :mineral_inventory ?mi .
        OPTIONAL { ?ms rdfs:label|:name ?ms_name . }
        ?ms :location_info/:location ?loc_wkt .
        ?mi :category ?mi_cat .
        ?mi :date [ :date_value ?date ] .
        FILTER(geof:sfWithin(?loc_wkt, "POLYGON(...)"))
        FILTER(geof:distance(?loc_wkt, "POINT(-118.57 47.56)"^^geo:wktLiteral, unit:mile) < 500)
        FILTER(?date >= "2000"^^xsd:gYear && ?date <= "2010"^^xsd:gYear) .
        ?mi :ore [ :ore_value ?ore_val_raw; :ore_unit ?ore_unit] .
        ?mi :grade [ :grade_value ?grade_val; :grade_unit ?grade_unit] .
        BIND(IF(bound(?ore_val_raw), ?ore_val_raw, 0) AS ?ore_val_pre)
        BIND(IF(?ore_unit = <http://data.nasa.gov/qudt/owl/unit#MetricTon>, ?ore_val_pre / 1e6, ?ore_val_pre)) AS ?ore_val)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "measured"), ?ore_val, 0) AS ?tonnage_measured)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "indicated"), ?ore_val, 0) AS ?tonnage_indicated)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "inferred"), ?ore_val, 0) AS ?tonnage_inferred)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "measured") && ?grade_val > 0, ?ore_val * ?grade_val, 0) AS ?contained_measured)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "indicated") && ?grade_val > 0, ?ore_val * ?grade_val, 0) AS ?contained_indicated)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "inferred") && ?grade_val > 0, ?ore_val * ?grade_val, 0) AS ?contained_inferred)
    }
    GROUP BY ?ms ?ms_name ?deposit_name ?loc_wkt }
```

```
qudt:hasDimension: "",
qudt:abbreviation": "g tonne-1",
..
qudt:hasPart: [
{
    qudt:hasDimension: "M",
    qudt:quantityKind: "http://data.nasa.gov/qudt/owl/unit#Gram",
    qudt:conversionMultiplier: 0.001,
    qudt:conversionOffset: 0.0,
    qudt:symbol: "g"
  },
  {
    ccut:exponent: "-1",
    qudt:hasDimension: "M",
    qudt:quantityKind: "http://data.nasa.gov/qudt/owl/unit#MetricTon",
    qudt:conversionMultiplier: 1000.0,
    qudt:conversionOffset: 0.0,
    qudt:symbol: "t"
  }
]
```

1st aggregation: tonnage computation

Agenda

- Intro
- Problem
- Approach
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work ⟸
- Future Directions
- Conclusions

Related Work

- General geo KBs (Zhu 2017, Brodaric 2020)
  - Mostly encompasses conceptual knowledge & data
  - Does not address: quantitative data integration

- GeoKGs related to mineral data (Qun 2023)
  - Tailored for geochemical data
  - Does not address: quantitative data integration

- Information extraction for geo KGs (Wang 2018)
  - Focus is on the data extraction
  - Does not address: data integration & entity linking

Agenda

- Intro
- Problem
- Approach
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work
- Future Directions ⟸
- Conclusions

Future Directions

- Advanced data modeling
  - More modalities (remote sensing)
  - More data (e.g., rapidly changing geographies)
  - Uncertainty & probabilistic modeling
- Enhanced embedding techniques for ER & EL
  - Expand integration of textual data
  - Utilize subword information & deep learning attention mechanisms
- KG expansion
  - Extend KG linkage to additional KBs & LOD
  - Apply & integrate with additional domains like archaeology & environmental sciences
- Dynamic semantic modeling
  - Create more sophisticated & evolving semantic models for accurate representation across multiple domains

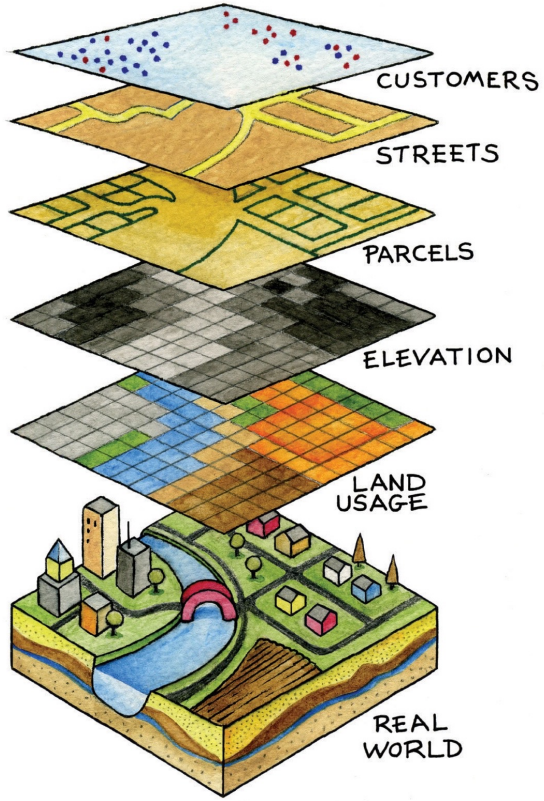


figure from *Essentials of Geographic Information Systems, Ch 7, Saylor Academy, 2012*

Agenda

- Intro
- Problem
- Approach
- Evaluation & Discussion
- Spatio-Temporal Analysis via GeoSPARQL
- Related Work
- Future Directions
- Conclusions ⟸

Conclusions

- Presented a method for the construction of a spatio-temporal KG from geo-referenced spatial entities in archive report

- Contributions
  - pipeline for building a KG from extracted quantitative, spatial & semantic information from historical mining data archives
    - automatically, incrementally, follows LD & SW principles, linked to web
  - method to identify & retrieve instances of a given type from a publicly available KG
    - specifically, entity matching commodities with GeoKB
  - spatio-temporal queries to automatically generate grade-tonnage models
  - publicly available resulting KG in the form of linked data covering two critical minerals: Zinc & Nickel
    - queryable RDF via a (Geo)SPARQL endpoint

https://minmod.isi.edu/

https://github.com/DARPA-CRITICALMAAS/ta2-minmod-data/