LinkedMusic Queries

Review and Handover

Junjun Cao

alienmusedh@gmail.com

There are corresponding hyperlinks in the following slides, pointing to specific documents or resources.

Query across Different Graphs

- SPARQL-linked data query <- Subject+Predicate+Object’s triples
  - e.g.: Wikidata SPARQL endpoint: https://query.wikidata.org/
- Query across different “graphs”(Federal Query):
- LinkedMusic 2024 Oct. yearly meeting: NLQ2SPARQL
  - The Session: (1) contrast with webpage (2) function beyond webpage
  - The Session + Wikidata
  - The Session + Wikidata +MusicBrainz
- RISM: <-XML for elaborate metadata
  - see https://github.com/DDMAL/linkedmusic-queries/issues/61

the basic

new challenge:
(1) difficult for data reconciliation
(2) difficult for schema-based NLQ2SPARQL

NLQ2SPARQL

- Issue A: What is the “best context” ? (refer to discussion 27)
  - The more concise, the better: try to leverage the pretrained knowledge of LLMs <- reconciliation with Wikidata
  - Query based on different context/prompt: (1) RDF snippets (2) Example pairs of NLQ & SPARQL (3) “schema”
  - ——Literature Review: “for those database without schema” …

prompt types	database type	NO. of shots	technology feature
RDF snippets	oversized, unknown schema	1-shot or few-shots	algorithm & computing-strength-oriented
NLQ&SPARQL	e.g. XML-featured	1-shot	the highest similarity between...
“Schema”	with schema or latent schema	0-shots	knowledge-representation-oriented

NLQ2SPARQL

- Graph Database: schema-free & flexibile & extensible
- What if we can obtain a ready-made “nominal” schema?
  - All the 14 databases must have schemas;
    - after the conversion, will the schema disappear?
  - CSV2RDF/RDB2RDF within Virtuoso: ontology can be as a schema of RDF automatically generated via mapping with schema
    - e.g., Taking CantusDB as an example for RDB2RDF into Virtuoso)
    - Tutorial: https://github.com/DDMAL/linkedmusic-datalake/tree/main/doc/rdb2rdf

NLQ2SPARQL

- Schema:
  - (1) ontology (open world hypothesis)
  - (2) shapes (close world hypothesis)

Which is better, can we rely on only one?

- Ontology: RDFs -> OWL
- (1) RDFs:
  - rdfs:domain - what is the class for the subject of the predicate
  - rdfs:range - what is the class for the object of the predicate
- (2) OWL: ObjectProperty; DataProperty; owl:inverseOf...

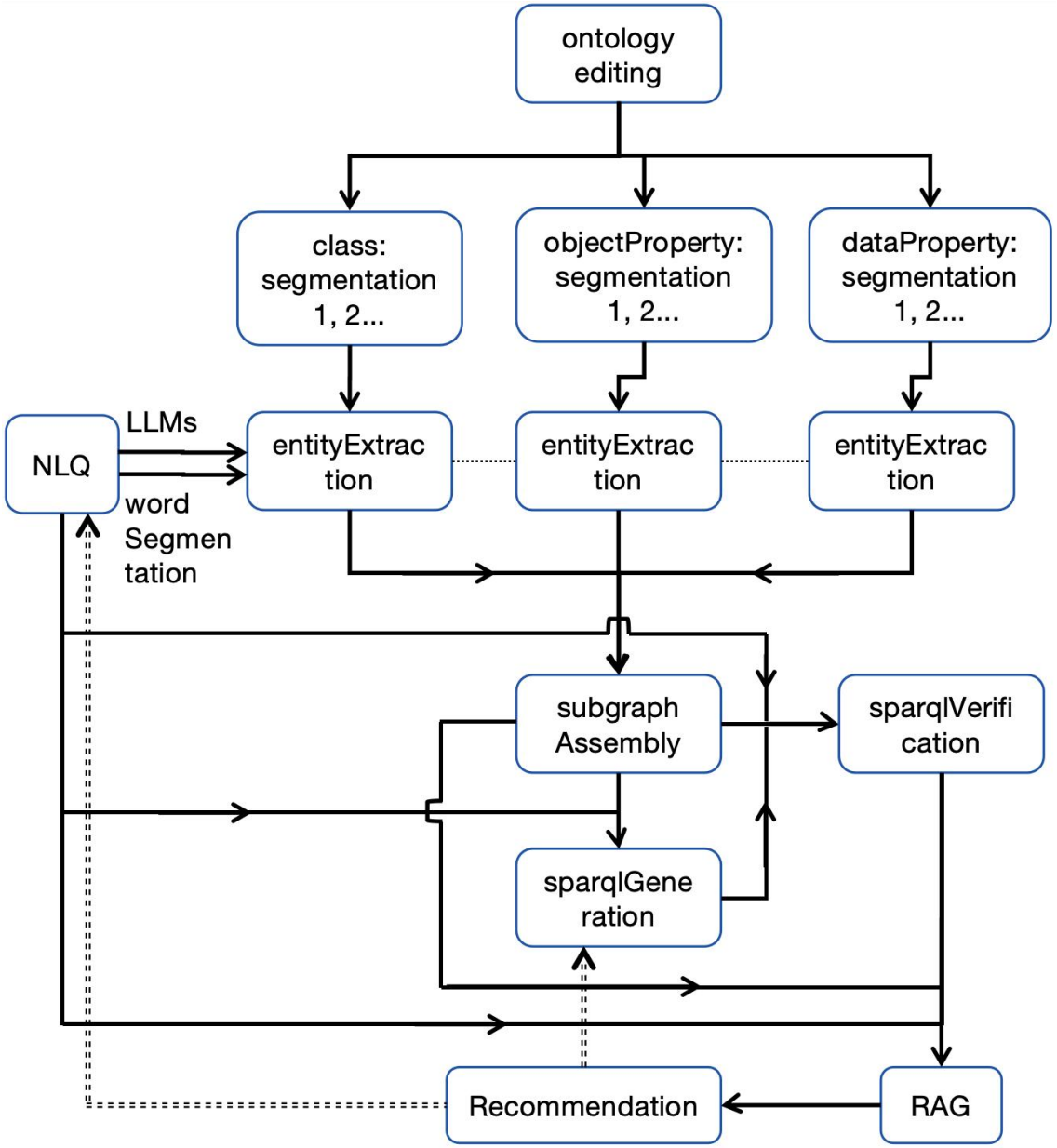
OP: When the property value is another node
DP: When the property value is
purely data such as literals, strings, integer, dates, boolean...

NLQ2SPARQL

- Issue-B: LLMs can not concertrate when the ontology is too large
  - (why)
- Experiment and solution:
- E.g., Chinese Traditional Music Culture Knowledge Base (CTM or ESEA)
  - which has a complicated existing ontology
    - which can also support “knowledge reasoning” which may relieve LLMs
  - which has visualization facilities
  - whose vocabularies would like to be shared with LinkedMusic eg:instru...

Python Script Workflow

- **Script**:
- 1_entitiesExtractionFromNLQ_basedOnOntology.py
- 2_subGraphAssemblyFromOntology_3_SPARQLgeneration.py
- **Workflow:**
- 1. Specific Ontology Editing
- --Protege
- (1) Clarification and Enrichment of ...
- (2) Semantic reinforcement. Eg:inverse
  - use ontology to replace shapes
- (3) Simplification



Python Script Workflow

- 2. Ontology Segmentation： into 3 parts
- 3. Entity Extraction from the Ontology Segments

  LLMs are prompted to extract (isolated) entities from an NLQ by mapping them with all segments of the ontology. Hereby, the “entities” also include class, property, or instance.

  ——key point! It decides whether this approach is robust enough
- 4. (Ontology) Subgraph Re-Assembly
  - Ontology is also a special graph! The nodes represent classes, and edges properties
  - Ambiguity of NLQ->Over-generalization->base for recommendation

Python Script Workflow

- 5. SPARQL Generation and Verification Based on Subgraph
  - (1)claude 4 (2) reflection: ontology for validation instead of shapes
- 6. Retrieval Augmented Generation(RAG) and Recommendation
  - Illustration on the retrieval result in respect to 3 contexts...
  - 2 scenarios of the rusults:
  - (1) If the result is too large or complicated, e.g.: issues 60
  - (2) If the result is too small or even empty:

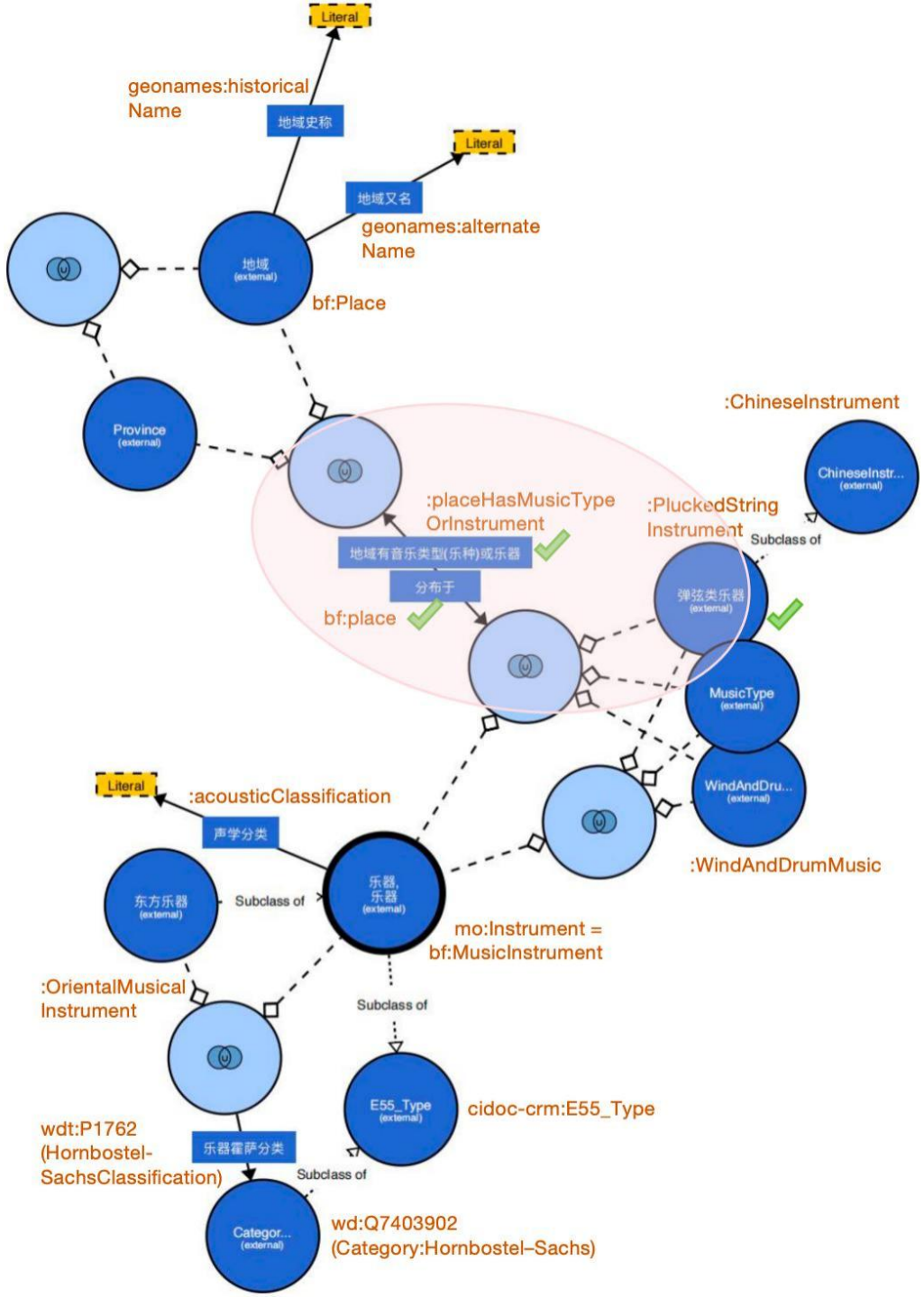
    the retrieval scope is broadened by relaxing query conditions/constraints in the SPARQL query, and other possible query patterns can be recommended

Python Script Workflow

- 6. Retrieval Augmented Generation(RAG) and **Recommendation**
  - 2 scenarios of the rusults:
  - (2) If the result is small or empty -> Recommendation Based on:
  - A. Relaxing SPARQL Constraint
  - B. the neighborhood within the Ontology Subgraph

The aforementioned can be found in the paper: *ESEA (East-and-Southeast-Asian) Traditional Music Knowledge Base and its Ontology-subgraph-driven NLQ2SPARQL Intelligent Question-Answering System Research*

--DDMAL/linkedmusic-queries

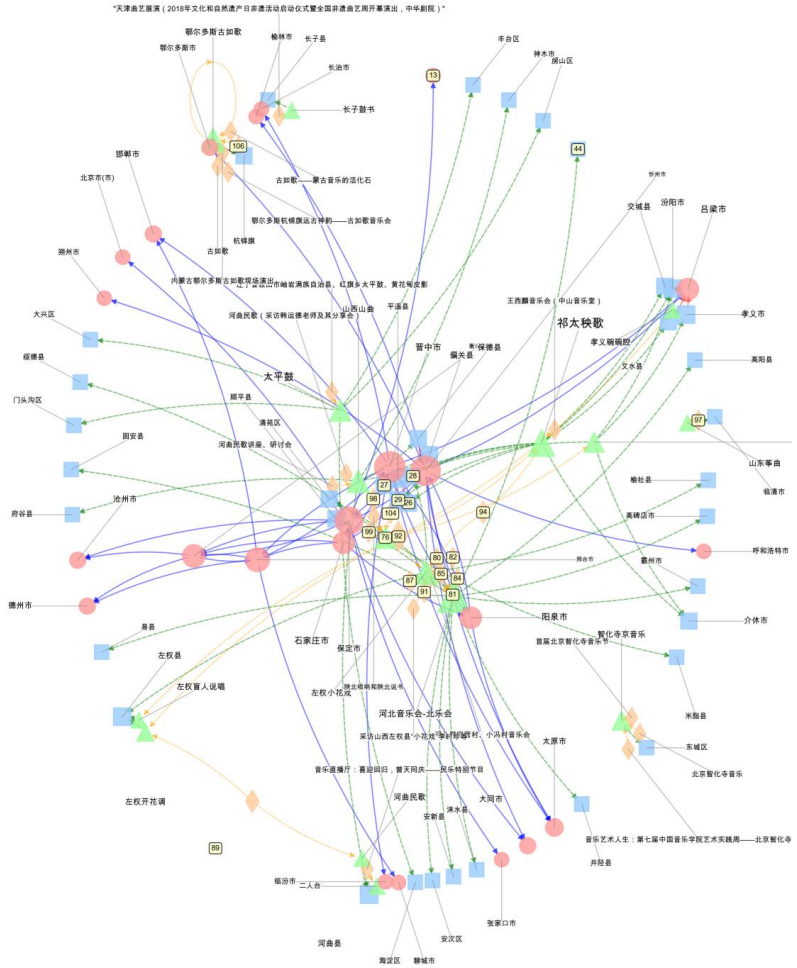


Python Script Workflow

- NLQ: Where is the "dongbula" (dombra, 东不拉) distributed, and what other plucked string instruments are distributed in the same regions/places? -> Light Red Area
- The Recommended pairs of NLQ & SPARQLs are based on the neighbourhood of the subgraph
- Reflection：
- (1) The Ambiguity of NLQ necessiates recommendation
- (2) Limitation of the approach: SPO-triples-question

Advance

- 7* examples (storage)/ vector databases
  - e.g.: (1) nested structure such as that in RISM (2) special functions queries e.g. (3) granularized questions
- 8* NLQ2SPARQL2NetworkAnalysis
  - (1) Obtain data for heterogeneous network analysis: generating “typed edge list” data
  - (2) Visualization



Advance

- Other NLQ examples beyond the scope of previous … **Modular workflow & Agent Development,** 扣子 (e.g., coze developing platform)

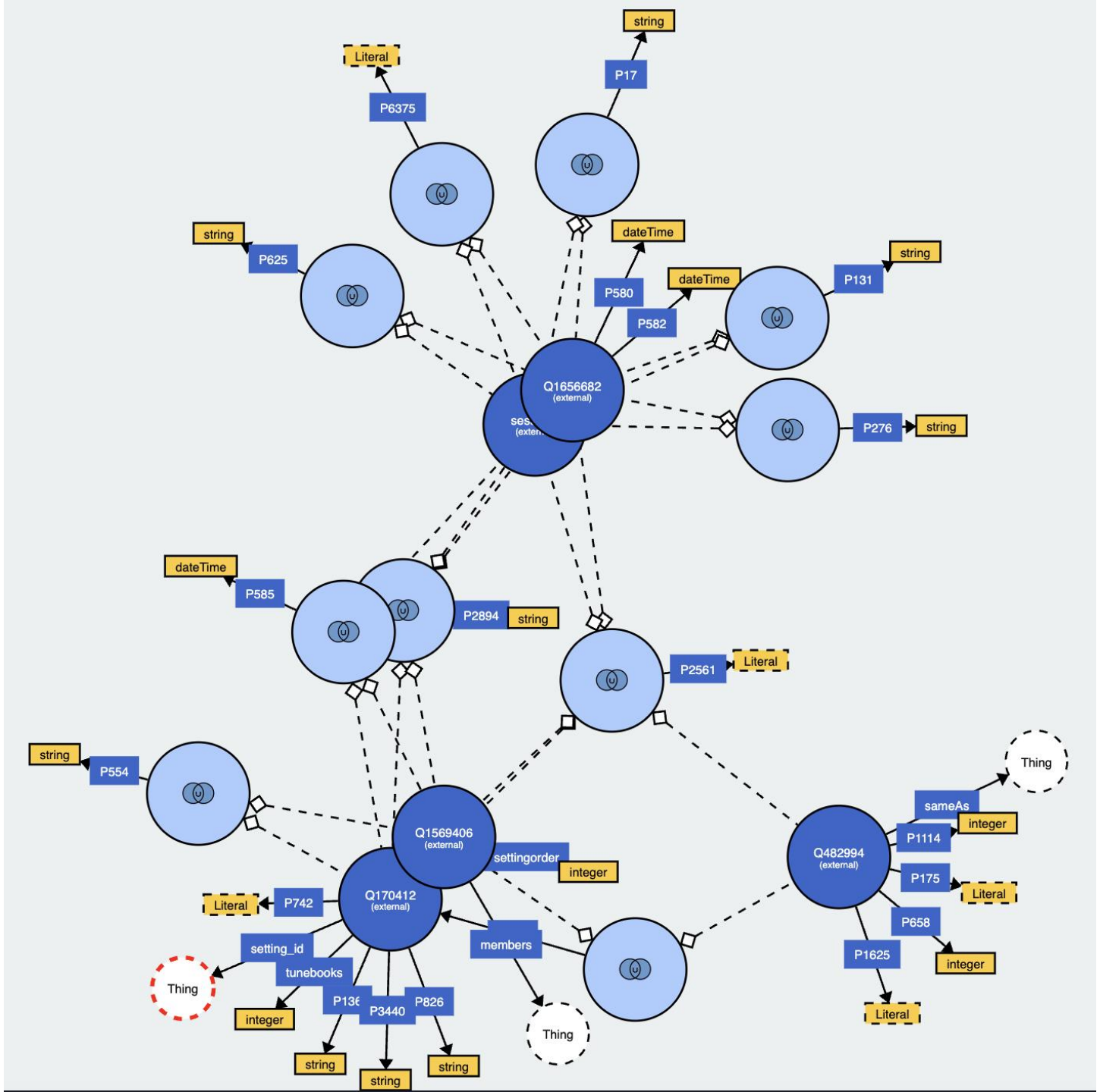
  --to faciliate live interaction between users and LLMs

- Retrospection on:
  - 2 orientations for NLQ2SPARQL
- Issue-C
- How to obtain ontology from existing RDF data graph?
- **VOID** (Vocabulary of Interlinked Datasets):
  - https://github.com/DDMAL/void-generator

It's almost impossible to always generate the corect or expected SPARQL <- ambiguity of NLQ itself

Advance

- Future Work:
- Issue-C: How to extract ontologies from... graphs in data lake?
- (1) For CSV2RDF, e.g. TheSession, MusicBrainz...
  - VOID generator
  - e.g.: Generate the ontology for TheSession
- (2) For RDB2RDF, e.g. CantusDB, SimmsaDB...
  - The internal process of RDB2RDF in Open Link Virtuoso
  - e.g.: https://github.com/DDMAL/linkedmusic-datalake/tree/main/doc/rdb2rdf



Advance

- -> TheOntologyFor TheSession
- unionOf classes
  - light blue circles with “U”
- owl:objectProperty
  - between 2 nodes
- owl:dataProperty
  - yellow squares
- owl:Thing

Addition/Reflection on Data Reconciliation and RDF Conversion

- Different scenarios and ways of None-graph-DBs to RDF Bulk Loader
  - https://github.com/DDMAL/linkedmusic-datalake/tree/main/doc/CSV2RDFInVirtuoso
- A summary of data reconciliations:
  - Guidelines or suggestions for data reconciliation (updated from time to time; collecting advice from everyone)
- Logs & Archives for reconciliations: (1)Properties in ...mapping.json file (2) Archived Excels
- The co-existence of unreconciled data and reconciled data
- To what extent will the reconciliation be conducted? -> for queries
  - preparation for ontology; balance between ObjectProperty and DataProperty
- Blank nodes and Named Graph

Thank you!

Junjun Cao
alienmusedh@gmail.com