Bridging Music Metadata:

Creating a Multilingual Instrument Vocabulary with Wikidata

Kevin Kishimoto, Stanford University

IAML 2025

Outline

1 Issue: What are we trying to solve?

2 Approach: Why Wikidata?

3 Project: Evaluation, prototype, next steps?

Stanford | University Libraries

Issue: What are we trying to solve?

- Link library-created vocabularies with the semantic web.
  - Library of Congress Medium of Performance Thesaurus (LCMPT).
  - IFLA UNIMARC Medium of Performance vocabulary.
  - Link these two vocabularies to one another (and to other vocabularies).

- Create structured data for names of musical instruments in different languages.

- Promote wider adoption of library vocabularies and possibly new tools to use them.

Issue: Current state of instrument terms

- Library of Congress Medium of Performance Thesaurus (LCMPT).

- IFLA UNIMARC Medium of Performance vocabulary.

- Wikidata.

Stanford | University Libraries

Approach: Why Wikidata?

"Wikidata is a free and open knowledge base that can be read and edited by both humans and machines." — wikidata.org

- Structured data based on RDF model (i.e. Linked Data).
- Low barrier to access and relatively easy to understand.
- Global community with many tools and scripts.
- SPARQL endpoint and API access.

- Wikidata data model is flexible and extensible which would accommodate refinement or expansion of the project over time.

Stanford | University Libraries

Approach: Wikidata lexicographical data

- Wikidata **ITEM**: represents a concept.
- Wikidata **LEXEME**: represents a word / phrase in a language.

- Use Items / Lexemes and relationships between them to build a structured multilingual vocabulary.
- Since concepts and words are separate data types, Wikidata lexicographical data model can more accurately represent the names of musical instruments based on language and context.

- Since Wikidata open knowledge base is continually being edited and revised, generalizations / ambiguities may be updated or corrected.

Same concept // Different languages

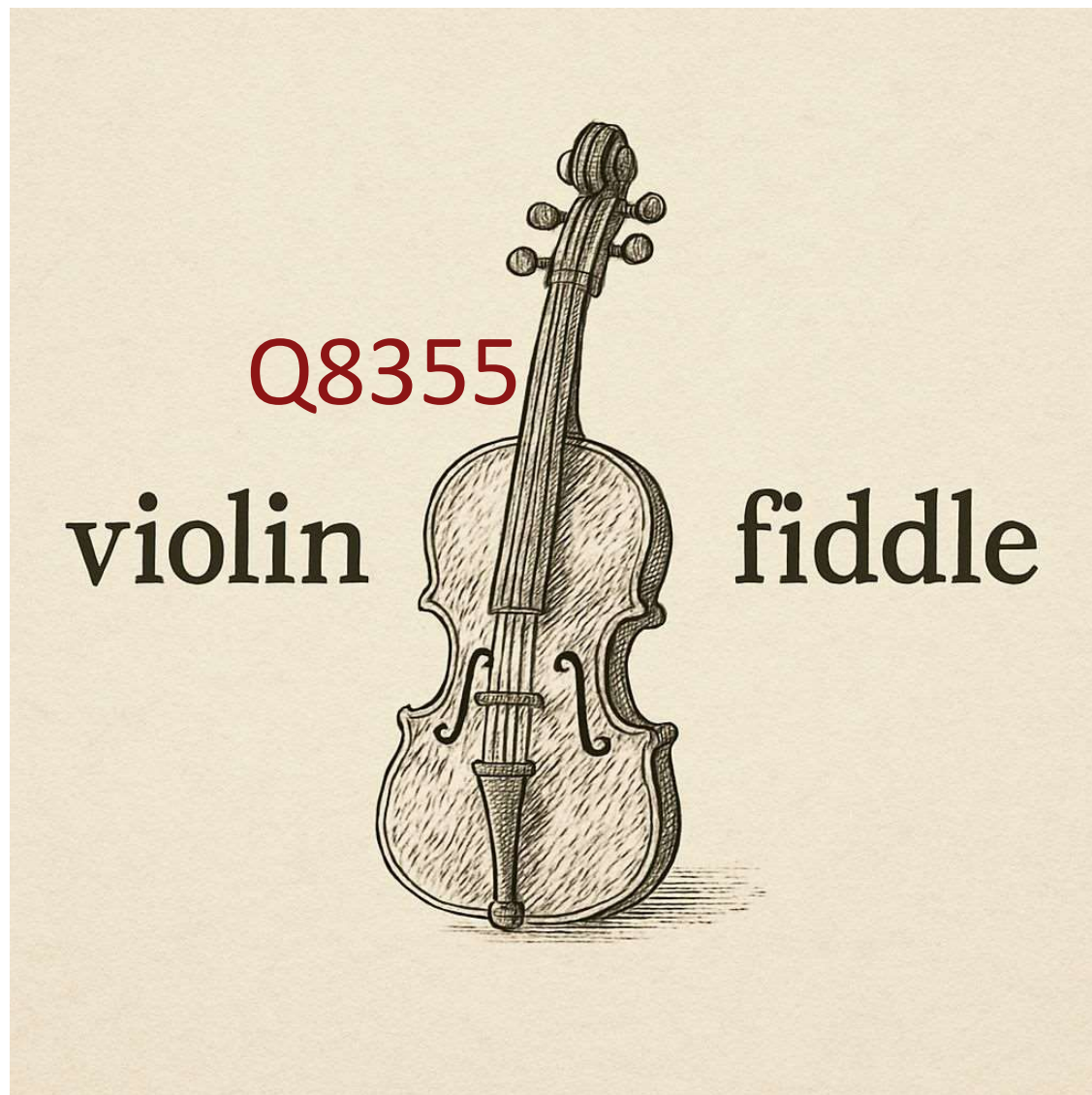A single concept (Item) may be known by different terms in different languages (Lexeme).

- violin (en) = L13249

- violon (fr) = L10281

- Geige (de) = L402499

- violino (it) = L60357

violon

Q8355

violin   Geige

violino

Same concept // Different words

A single concept (Item) may be known by different terms in the same language (Lexeme).

- violin (en) = L13249
- fiddle (en) = L320581

Q8355

violin   fiddle
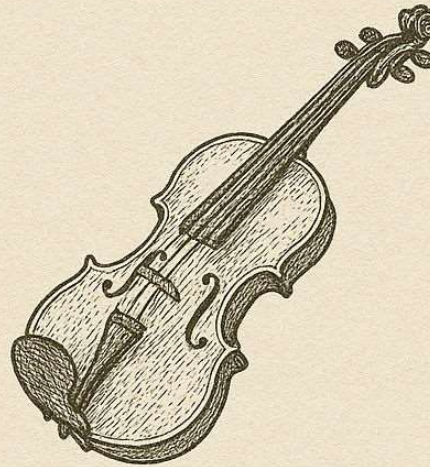
Same word // Different concepts

A single word (Lexeme) may be used to describe different concepts (Items) in a single language.

Q810451

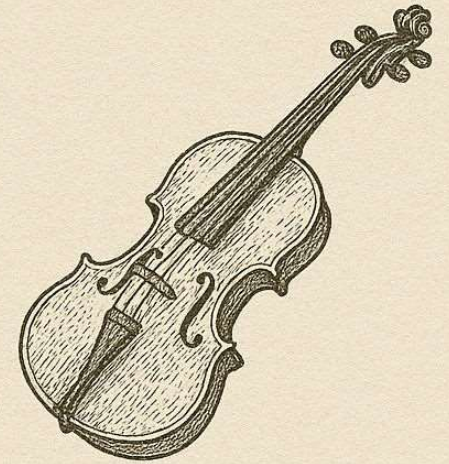Q719678

"bass"

L13251

Q46185

Q80019

Same spelling // Different languages

The same string in different languages (same spelling) may be used to describe different concepts.
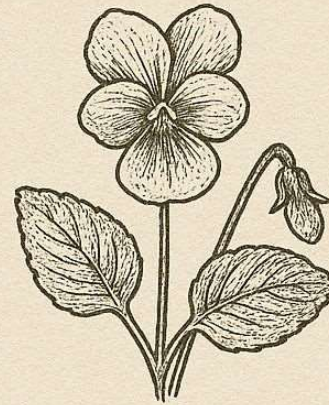
- viola (en) = string instrument
- alto (fr) = string instrument
- alto (en) = singer
- Viola (de) = flower

viola (en)

alto (fr)

Viola (de)

alto (en)

Wikidata Lexeme: Data Model

- Wikidata Lexeme ID (begins with "L")
- Lemma + language (standard form / dictionary form)
- Lexical category (part of speech / word class)

Lexeme	Discussion

(L14142) **guitar** ← lemma
en

Language English
Lexical category noun

Wikidata Lexeme: Identifiers

- Identifiers from online dictionaries / lexical reference sources.
- Identifier types are established as Wikidata properties.
  - New Oxford American Dictionary ID (P12690)
  - Duden lexeme ID (P8376)
  - Larousse Online French Dictionary ID (P11118)
  - etc.

Identifiers

Australian Oxford Dictionary ID	⇕ 0023525 •••
	▾ 0 references

Wikidata Lexeme: Senses

- Senses provide different meanings of a Lexeme.
- Single Lexeme may have multiple senses.
- Gloss: definition of the sense using natural language.
- "Item for this sense" statement links Lexeme to Item.

Senses

L14142-S1	English	gloss →	fretted string instrument
	Spanish		guitarra, instrumento musical de cuerda pulsada que tradicionalmente consta de seis cuerdas
	Basque		gitarra, hari instrumentu edo musika-tresna

Statements about L14142-S1

item for this sense	guitar ← Q6607

Wikidata Lexeme: Forms

- Different forms of a Lexeme used in speech or writing.

Project: Evaluation?

- Parse two MOP vocabularies into format for comparison.
  - LCMPT
  - IFLA UNIMARC Medium of Performance.

- Query Wikidata.
  - Items that have existing links to MOP vocabularies.
  - Items are linked to Lexemes (via senses).

- Devise minimum viable data model for creating usable lexicographical data for musical instruments.

- Discuss model and methods with Wikidata experts.

Stanford | University Libraries

Project: Prototype?

- Scope project.
  - Which types of instrumental names can be accurately modeled in Wikidata?
  - Which types of instrumental names require more subject expertise?
  - Which types of instrumental names cannot be modeled?

- Begin with 'simplest' cases.
  - Focus on Lexemes in native language.
  - Instrument names with entries in language dictionaries and are not ambiguous.
  - Instruments are described in reference sources.
  - Wikidata Items exist for the concepts and may have links to vocabularies.

- Create first-attempt spreadsheet with data to add / update.

Project: Next steps?

- Get assistance in writing scripts for batch updates to Wikidata.

- Present project to IAML. Ask for help.
    - Request for feedback on data model and process.
    - Expand subject and language expertise.
    - Scale up.

- Evaluate new Wikidata dataset for musical instruments.

- Expand to include additional instrument vocabularies in project.

- Share findings and data with external groups / users.
    - LinkedMusic Project: Universal Musical Instrument Lexicon (UMIL)
    - LD4, Wikidata Wikiproject Music, library associations, etc.

Questions?

kishimoto@stanford.edu

Thank You