# *soweego*

# Link Wikidata to large catalogs

*Anywhere, 14 September 2021*

*Marco Fossati / Hjfocs*

# Disclaimer

A tool or not a tool?

# a WMF project

Project Grants program

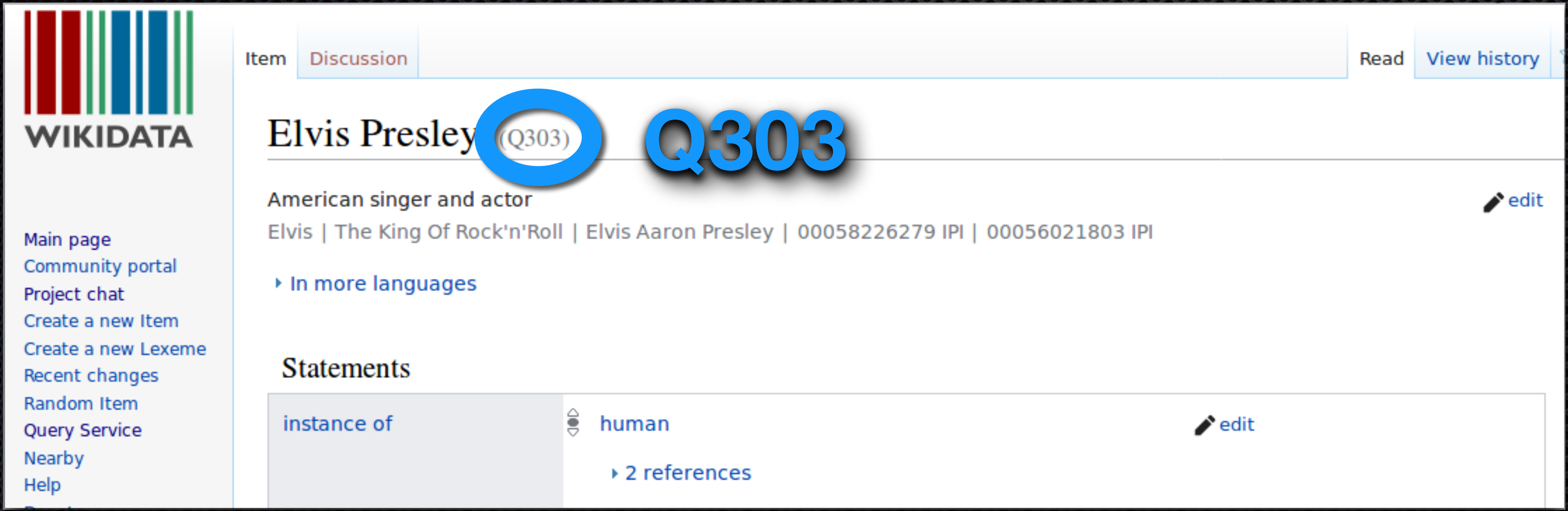
# Example

Q303 → 01809

# Idea

- *Record linkage* AKA *data matching* AKA *entity resolution*
- Input: dataset **pair**
  - source = Wikidata
  - target = external catalog
- Output: **links**, as Wikidata identifier statements

# Example



Elvis Presley (Q303)

Q303

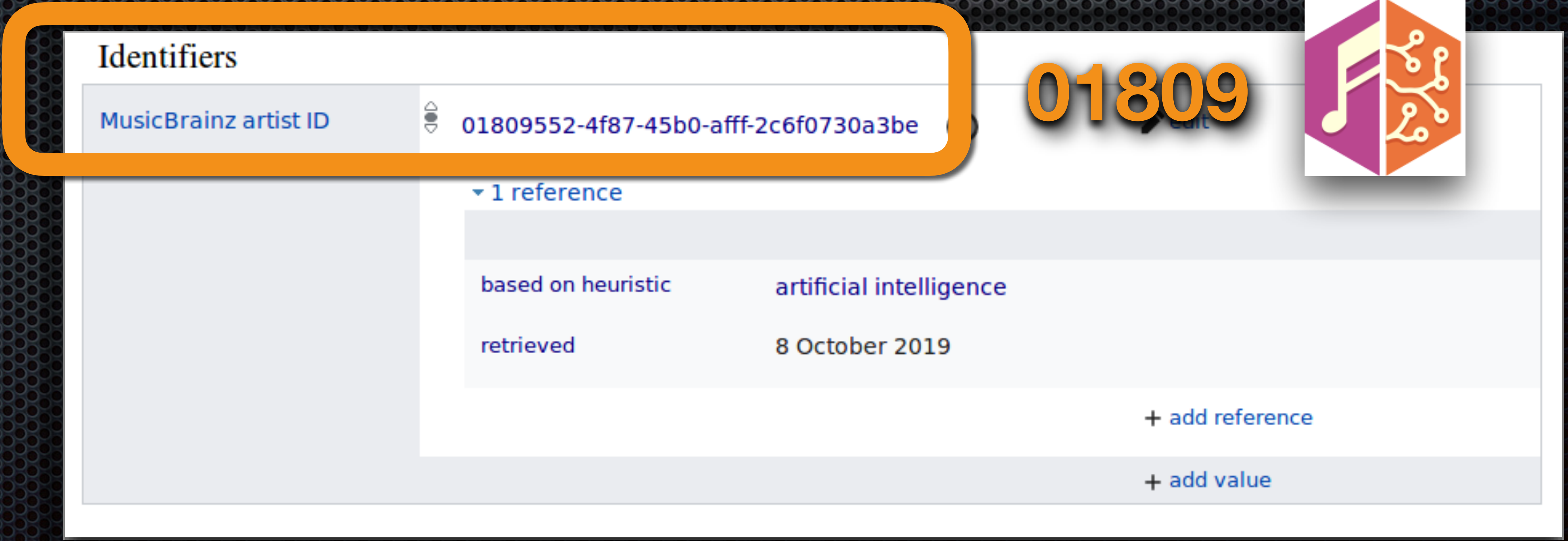
American singer and actor

Elvis | The King Of Rock'n'Roll | Elvis Aaron Presley | 00058226279 IPI | 00056021803 IPI

In more languages

Statements

| | |
|---|---|
| instance of | human<br>2 references |



Identifiers

| | |
|---|---|
| MusicBrainz artist ID | 01809552-4f87-45b0-afff-2c6f0730a3be |

01809

1 reference

| | |
|---|---|
| based on heuristic | artificial intelligence |
| retrieved | 8 October 2019 |

# Why

- Increase Wikidata **quality** & **trust**
  - **quality** = identifiers enable feedback loops
  - **trust** = references to external reliable sources

# How

- Record linkage workflow
- Supervised machine learning

# What

- **People**, a big Wikidata slice
- **Works**, made by people

# What's hot?

**Validator**: sync Wikidata to target catalogs

# In a nutshell

- **Run** validation criteria
  1. ID existence
  2. URLs
  3. domain-specific data

- **Apply** actions
  - enrich items
  - rank statements
  - submit to target communities

# Example

Q303 → 01809

# #1 Existence

**Q303** > **01809**, *but* **01809** is *no more* in MusicBrainz

WIKIDATA

**Deprecate** statement

# #2 URLs

Q303 has 7 URLs - 01809 has 8 URLs - 3 overlap

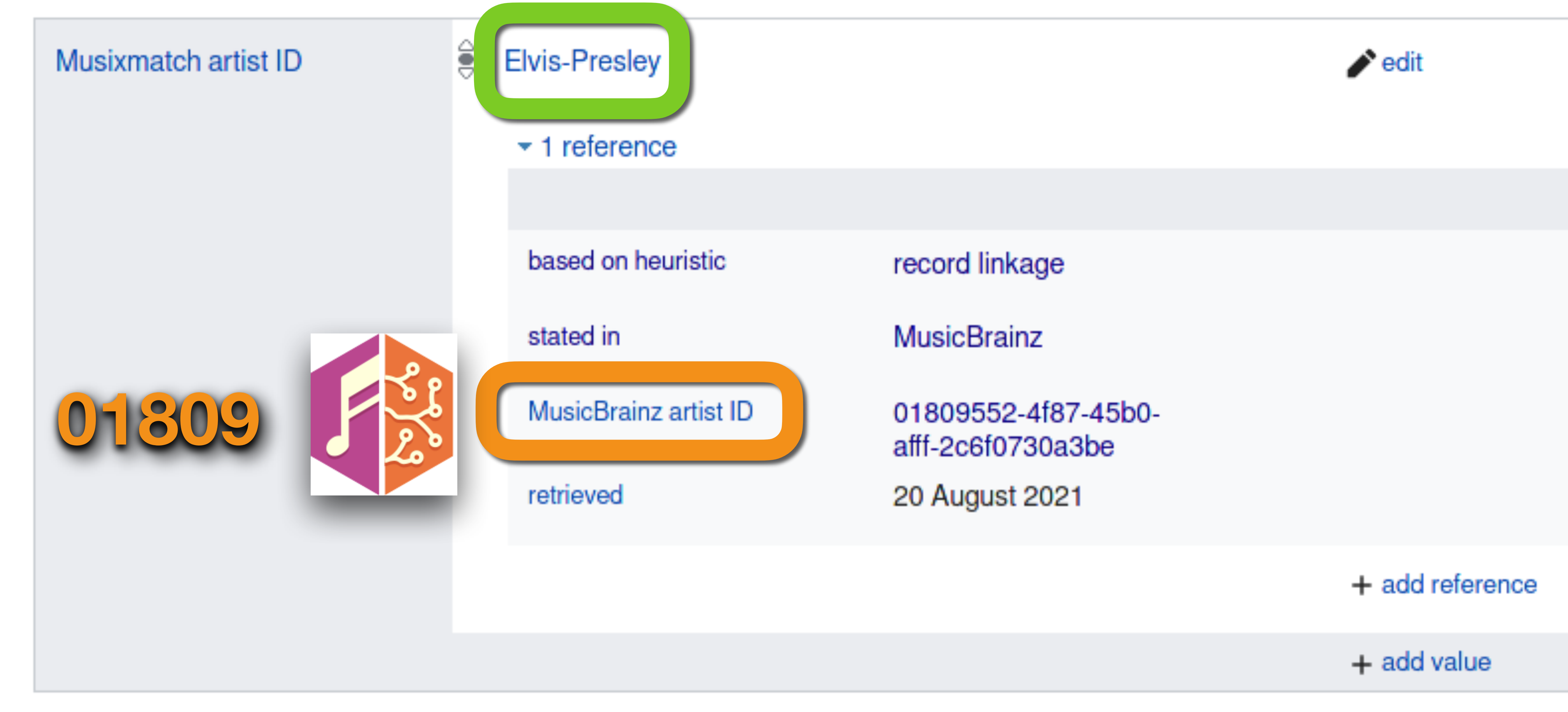
WIKIDATA

**add 5** URLs to **Q303**
(8 - 3)

MusicBrainz

**submit 4** URLs
(7 - 3)

# URLs are dangerous: let's add **IDs** first



Musixmatch artist ID

Elvis-Presley

edit

1 reference

| | |
|---|---|
| based on heuristic | record linkage |
| stated in | MusicBrainz |
| MusicBrainz artist ID | 01809552-4f87-45b0-afff-2c6f0730a3be |
| retrieved | 20 August 2021 |

+ add reference

+ add value

01809

# #3 Domain-specific data

**Q303** was born on **1935** in **Tupelo**
**01809** was born on **1934** in **Memphis**

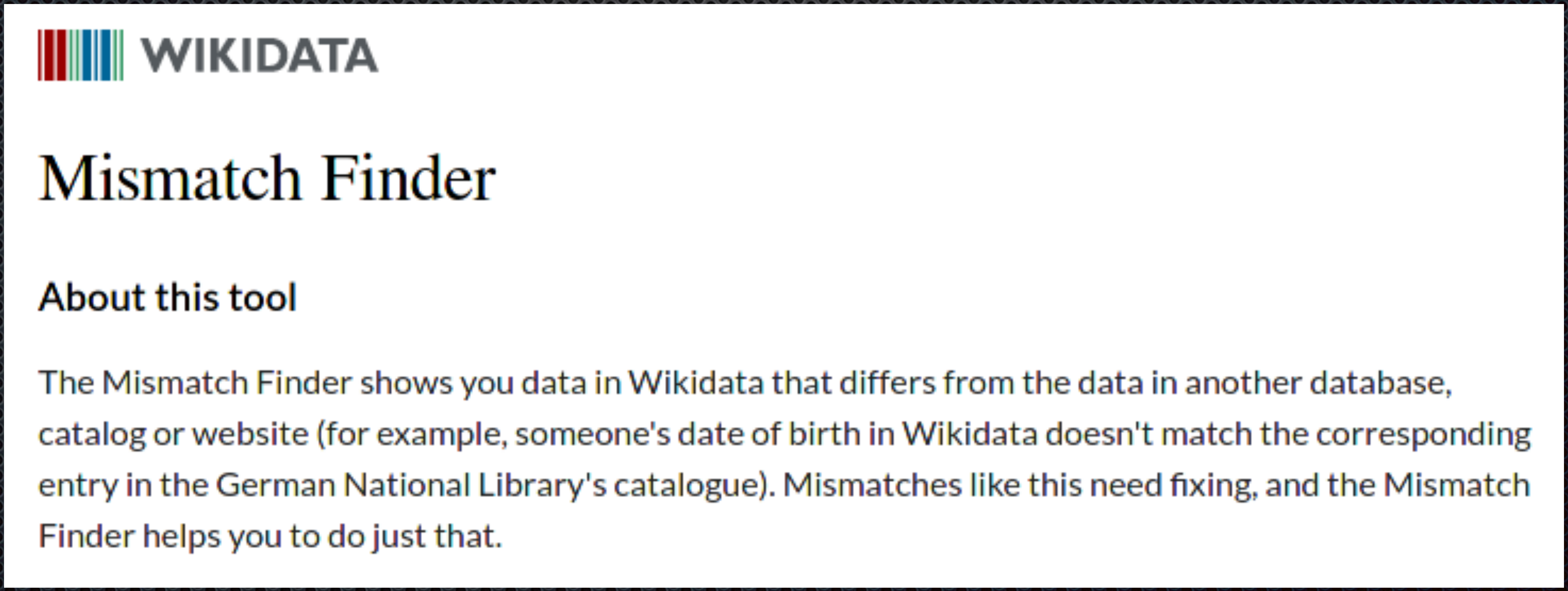
WIKIDATA

**add 2** statements to
**Q303**

MusicBrainz

**submit 2** values

# They can be controversial: let's feed the **mismatch finder** first



WIKIDATA

## Mismatch Finder

### About this tool

The Mismatch Finder shows you data in Wikidata that differs from the data in another database, catalog or website (for example, someone's date of birth in Wikidata doesn't match the corresponding entry in the German National Library's catalogue). Mismatches like this need fixing, and the Mismatch Finder helps you to do just that.

python powered

# Thanks!

https://soweego.readthedocs.io/

*Anywhere, 14 September 2021*

*Marco Fossati / Hjfocs*