Preliminary findings of the Librarybase project

James Hare

WikiCite 2017 • Vienna, Austria • 23 May 2017

What is Librarybase?

- All of the citations on Wikipedia
- Structured and queryable
- Recommendations for sources
- **…eventually**

Accomplished to date

- Wikimedia Foundation grant funding
- Consultation with community members
- Data architecture
- Some mass import work
- **Quality control**
- **Code optimization**

The plan

- Storage of **source metadata** in Wikidata or Librarybase-wiki
- Storage of **citation events** in a dedicated database
- A high-performance **identifier cross-reference** service

Quality control

- Identifier cross-referencing means de-duplication and richer data overall
- Filling in the gaps
- DOI normalization: all uppercase!
- Results: **significant drop in constraint violations**

Here are some metrics

- Items with DOIs:
  - Before: **866** violations across **249,946** items
  - After: **225** violations across **605,198** items
- Items with PMIDs:
  - Before: **23,558** violations across **249,951** items
  - After: **6,007** violations across **598,919** items
- Items with PMCIDs:
  - Before: **29,958** violations across **187,102** items
  - After: **11,364** violations across **339,643** items

Why does this matter?

- We want Librarybase to be as close to **real-time** as possible
- This means **automation** and **editing at scale**
- This is particularly **risky**

Best practices for mass import

- Define an informal schema
- Preserve **provenance**
- More: https://meta.wikimedia.org/wiki/Grants:Learning_patterns/Wikidata_mass_imports

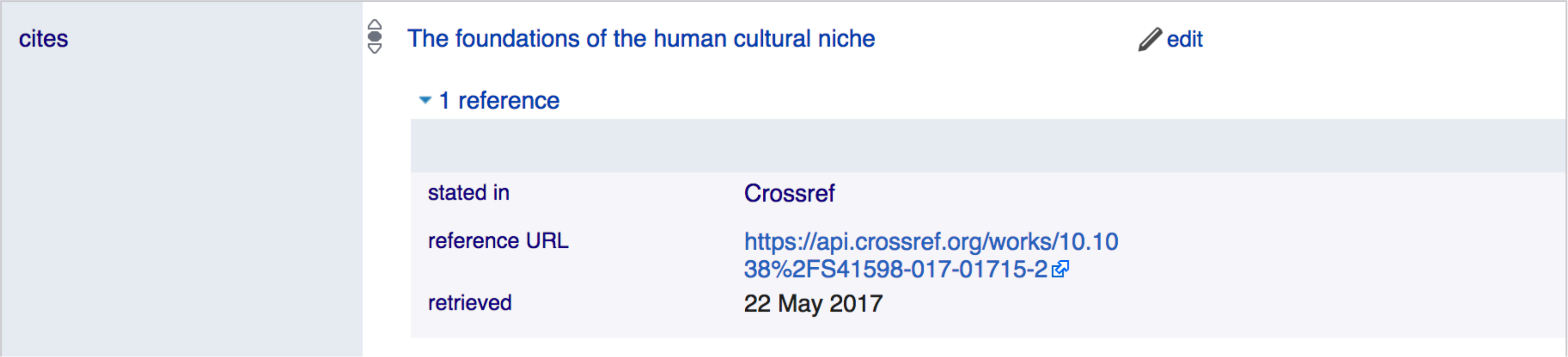
Code optimization

- Taking existing Python scripts to create and edit Wikidata items
- Make it run *fast* to keep up with a growing Wikidata
- Wikidata Integrator
- Redis caching!

Wikidata Citation Graph

- Supply chain of knowledge: Wikipedia cites a source, which cites another source…
- Citation networks help us map information's provenance
- **3.35 million citation connections on Wikidata**

Young children copy cumulative technological design in the absence of action information (Q29994699)



cites: The foundations of the human cultural niche

stated in	Crossref
reference URL	https://api.crossref.org/works/10.1038%2FS41598-017-01715-2
retrieved	22 May 2017

What's next?

- Implementation: citation event and identifier cross-reference databases
- Continuously updating data
- Modeling sources other than journal articles
- Recommending sources for Wikipedia articles