

Interview with Diana Duncan - Overview of the BHL Author Names Project

December 2, 2021

About Diana

Diana formerly served as the BHL Cataloging Committee Chair and Technical Services Librarian at the Field Museum. She is now an active BHL and Wikidata volunteer with a deep bench of metadata expertise. She uses Wikidata as a research database and source of URIs for Author name records in the BHL database, with the goal of merging and disambiguating these names to facilitate better search and indexing. She has piloted and refined an end-to-end OpenRefine workflow to enable reconciliation between the two knowledge bases: BHL and Wikidata.

Overview of the BHL Author Names Project

Duplicate author names associated with title records have always been a problem for the BHL. With the addition of article-level metadata, this problem has only grown exponentially. Data providers are not required to submit the author name data with an accompanying URI; this means there is much work to do for the BHL cataloging committee to disambiguate duplicate names in the BHL. No URI, means BHL catalogers must use external systems such as VIAF, LC Name Authority file, Wikidata, google searches etc. to do painstaking research and reconciliation work.

See Diana's: [Open Refine / Wikidata author name reconciliation process](#)

What was the catalyst for the work to reconcile author names in BHL?

- Article creation has created a lot of issues for us: Names in BHL have become a nightmare; ideally we would have one record per unique creator but we are

very far off from this ideal

- Data providers are giving us different variations of names
 - Names attached to title
 - Rules for formatting names are all different across partners resulting in dupes
 - Names attached to articles (an even bigger problem)
 - Pensoft, sometimes has ORCID IDs which is very helpful but not always
 - Biostor is the most problematic because authors don't have any URIs.
- Diacritics create problems too by creating duplicate entries.
- VIAF and Wikidata have duplicate entries as well – so the problem of duplicates is not necessarily resolved by consulting these databases
- VIAF is more comprehensive than wikidata – and better for older content
- Alternative names in wikidata give synonyms to search on

When did you start doing this work?

- Summer of 2020

How often are you doing the work?

- Frequency 1x per month for pulling down spreadsheets to work on
- Once the spreadsheet is cleaned-up volunteers will go in and make sure entity is with the right wikidata / then volunteers go and merge and add wikidata ID

How long does it take to process one 1000 row spreadsheet?

- Typically, a full day or more

Who else is doing this work?

- Diana Shih
- Becky W
- Diane Shaw (BHL doing a lot of wikidata work)

Impact of the work?

- Improved search / metadata
- Users are write in and are noticing the work

Is this type of work being done on other BHL entities?

- As far as I know, no work is not being done for geographies and/or subjects – I don't even think you can use the wikidata Q number for these records in the BHL database at this time.

Brainstorming: Some ideas to help solve the author name problem?

- Document Diana's process, create Open Refine recipe for other staff to use. We need more people on this.
- Load BHL author names data into Wikidata as a project and then use their MixNMatch tool to crowdsource this tsunami of reconciliation work
- Speak to Rod Page about his current workflow and the addition of URIs for author names -- should we use wikidata as a staging area. Deposit citations into wikidata first, disambiguate author names over there, then bring them into Biostor for eventual harvest?

- Talk to Crossref and ask if it's possible to enforce author name URIs – do we have any sway, can we make some type of appeal?
- For author names lacking a URI completely, generate one by adding it to Wikidata
- Allow the Wikidata Q number to be used for other entity types (geographies, articles, subjects etc.)
- Tech development- Allow the cataloging group to edit subject headings too. Need a conversion project to FAST heading or some RDF vocab.
- For partner data can we retroactively z-fetch author authority records that have the LC control number, then at least disambiguate at the title level?
- ACTION: Need more discussion and a policy around preferred name authority database to use for BHL, should the preferred name be pulled from VIAF, LC Authority, Wikidata, or elsewhere?