Interview with Andra Waagmeester - Perspectives on Wikimedia Projects from a Bioinformatician

March 23, 2022

About Andra - Q19845625

Andra is a bioinformatician, semantic "webby," and works on Genewiki professionally. Genewiki is a project that began its work by making articles in Wikipedia to expose and preserve gene data created during research funding cycles for the public. (the nature of funding cycles means this data was getting lost)

It is now moving data into Wikidata because while useful in Wikipedia researchers wanted to expose Genewiki knowledge in other languages. (2012)

He has been involved with Wikidata since 2014; he takes the info from public databases → Wikidata.

Andra uses the taxonomic world as a playground to explore and invent.

Can we talk about your work getting GBIF images into Wikimedia Commons and Wikipedia articles using your app Tarsier?

It started at Wikimania 2018 where a group of Wikimedians and Andra decided they wanted to create an app that would link to iNaturalist images. To date, 44,000 thousand images from iNaturalist have made their way into Wikipedia articles.

Andra wanted to do this with GBIF images too. One complication is that GBIF does not annotate its image licenses separately from its metadata license.

During this work, he has found interesting things such as:

- Frogs are difficult to find openly licensed images/dragonflies are very easy to find exposing the particular values of each community and levels of “openness”

As you may know, user Fae already uploaded BHL’s Flickr image collection in 2015 and those Flickr images are on Wiki Commons. Any thoughts on getting more of those images into Wikidata items and Wikipedia articles?

Follow-up: send Andra the link to machine tags in Flickr and blog post about info. We will discuss more soon; possibly on a next Wikipedia Weekly: Biodiversity edition.

Opinions about the future fate of Wikispecies?

I really don’t have enough time to work on Wikispecies; if you are interested in getting BHL images into Wikispecies then you have to add 300+ languages pages manually. I think there is value in the community and the work but my interest is data reuse which is Wikidata.

Wikibase Federation: can we go into a bit more detail on options for federation. Do you recommend staying with the Wikidata ontology for the purposes of federated querying?

Wikidata - natively it is a relational database that stores JSON blobs; it is very similar to Wikipedia built on MediaWiki. Stores 1 article as 1 single record – Wikidata does the same thing but instead of storing text it is storing a JSON blob.

In 2015 they deployed Blazegraph as an additional search layer which means there is a copy of the data from wikibase to the triplestore and there is lag.

Note: Andra’s semantic peers don’t love wikibase / wikidata because it is not actually a triple store and has its own namespace that doesn’t conform to W3C RDF specifications.

However, the unique selling point of wikibase is the ability to edit a single statement →

you don't need to have a PhD in computer science to add a statement. In order to do a single statement in an RDF database you would need to do an insert using APIs etc.

"What Wikidata brings to the table is that it takes away (some of) the necessity to set up and maintain infrastructure. Which allows users to focus primarily on content with the need to deal with ICT issues. Maintaining Wikibase requires that users have to consider maintenance issues, however since Wikibase uses the same Infrastructure as Wikidata, less effort is needed to overcome learning curves.

People do know how to use wikidata, so it is more intuitive to use the same interface on other datasets." Essentially, it allows anyone to interact with the semantic web/

Wikibase limitations:

- Because BIBFRAME is an RDF ontology it is not so easy to deploy on Wikibase right now
- It can't handle millions of statements (e.g. BHL); there will be a workshop in March about increasing the intake of Wikibase. Andra to send an invite.

Should BHL join the Wikibase Stakeholders Group (WBSG)?

Wikibase Stakeholders Group is about institutions discussing requirements and needing similar functionality. They pool financial resources to hire developers to build out features. Additionally, it's about surfacing institutional needs to WMDE.

BHL should be on the WBSG. It's for institutes, libraries, museums etc. (it's a hangout place, they meet 1x per month) → they have a central fund everyone chips for development costs. Membership fees currently = $0 but may change later.

JJ to email Andra's colleagues about BHL joining in.

Top recommendation to help BHL open-up its data to the Wikimedia ecosystem?

Look into bots.

Gene wiki bots: http://jenkins.sulab.org; Andra is happy to work with us on creating a BHL bots; work on 1 bot on a specific dataset with the caveat of he will not maintain it; needs to pass it of to BHL technical team for data curation/maintenance stuff.

Most effective way to get all BHL entities into wikidata and interlinked ?

We have:

- BHL bibliography ID (P4327) (books/journals)
- BHL creator ID (P4081) (authors)
- BHL Page ID (P687) (pages) - > he uses page IDs
- BHL part ID (P6535) (articles)
- BHL name ID (P8724) (taxons) ← let's work on this one.

Not in Wikidata:

- BHL Item ID → not represented
- BHL contributor ID → not represented
- BHL collection ID → not represented

JJ sent Andra all of our datasets. He is going to look at them more.

What excites you most about Wikidata / Wikibase?

He is semantic webby by training; quickly realized that setting up semantic infrastructure is very expensive.

Wikibase is essentially a kickstarter for any LOD venture and helps get folks on the linked data bandwagon.

Also check out: https://triplydb.com - free triplestore then starts at $700 per month for paid tier

Similar to the BnF project, Luxembourg is using wikibase for their ILS but using CIDOC and has built a wikibase that has mapping rules. They have a large technical team working on this for:

- CIDOC mapping
- Wikibase doesn't have strong security layers so built an additional SSL layer.
- Here is a recent talk about it: https://www.youtube.com/watch?v=MDjyiYrOWJQ