MetaBrainz Summit 2018

State of the Brainz

Rob gave a talk explaining how the Foundation has been doing, talking about finances, staffing, progress etc etc.

Infrastructure

Zas gave a presentation on the infrastructure that we're using these days.

- 1.5 TB of RAM
- 18 TB of storage
- 11B of HTTP 200s in the last year

Ruaok: we focus more on the bad stuff, which is good because it keeps us improving, but our numbers are actually very good.

Hetzner has been really good for us.

Link: https://github.com/zas/hovercraft_presentations

State of the Projects

MusicBrainz

Michael (bitmap) gave a presentation on updates to MusicBrainz over the past year.

- Work has been focused on design overhaul, converting to React, things are a bit slower than we'd like but should probably finish it in 6 months.
- Chhavi did work on the design system, read her blog post.
- Reo worked on an initial version of the genre system, to be released soon.
- Sambhav finished the new SOLR search, applause!
- No schema changes this year.
- We have 2.5 people working on MusicBrainz now, which is good - Rob
- We're working on making VMs easier, first Rob, now Yvan, initially started by jsturgis

ListenBrainz

- Data dumps
- ListenBrainz recommendation engine and how it changes the music industry
- MessyBrainz

CritiqueBrainz

- Need more users and communication
- Comments on reviews
- More features
- Beta site (beta.critiqurbrainz.org) for unreleased/test features/bug fixes

AcousticBrainz

- Rashi has worked a lot on integration with MusicBrainz for here GSoC project
- We've migrated AB to our main infrastructure, which meant 15 hours of downtime

BookBrainz

Presented the new roadmap for Bookbrainz:
https://docs.google.com/document/d/1wStZLaIwjDMYM9yxAKckm8i6qymSjy8MMppNIC5dziw/edit?usp=sharing
Short summary of the next few months:

- Some UI improvements and an FAQ/started guide page for the upcoming Google Code-In
- Two important features we hope to have finished before GCI:
    - Warning when creating an entity if another one with same the name exists in the database
    - Basic entity merging tool
- Minimum Viable Product:
    - Creator Credits
    - More advanced merging
    - Schema review and implementation
    - Replace ElasticSearch with Solr
    - Dockerize and move hosting to Hetzner servers

Schema and vocabulary discussion:
https://docs.google.com/document/d/1jBl8YwBrF0DosjZ49lVC-buel8eAc8Kja3ceVEgLk3Q/edit#heading=h.uvws1ndjpifs
Went quickly over new entity definitions and associated UI vocabulary, and later in more detail in a break-out session. Feedback has been included in the above document

Also had two other one-on-one break-out sessions about Solr search and hosting at Hetzner that were mostly brain dumps

Community

Freso had updates from the community side for our projects.

We have around 150–200 new accounts a day (most of them probably spammers), last year we had 500–1000/day. We updated the captcha for new user registration which is likely what has brought this number down. We still get a lot of spam, but it is still significantly better.

On the forums we get 90–140 new accounts a day. Users are created when they verify their emails. Freso has been looking at the people who get flagged as spam and has been blacklisting their domains, further decreasing the amount of spam accounts carried over to the forums.

SpamBrainz is the Next Big Thing™ that will hopefully help to deal with much of the remaining incoming spam.

Edits and votes per week are pretty much the same as last year.

1100–1400 active editors and users on daily basis, same as last year.

~100 users cast a vote every day.

Very close to 2M MusicBrainz accounts.

Forums have fewer new users, but more activity compared to last year.

Looking into CHAOSS for getting a better picture of the community across platforms.

GCI is about to start up again and so far seems to be coming along nicely with preparations.

SpamBrainz

Leo Verto presenting this:
https://github.com/LeoVerto/spambrainz_ml/blob/master/lodbrok_evaluation.ipynb

- The Lodbrok model has achieved a 99+% percent accuracy at identifying spam editors while not mis-classifying more than 0.2% of real editors
- Some engineering effort still required to get it running on MeB infrastructure
- Work needed on MB side to allow spam ninjas to handle reports
- Future models should be able to also deal with spam entities

The DOREMUS Project -- Loujine

- Doing Reusable Musical Data
- https://github.com/DOREMUS-ANR
- https://loujine.github.io/musicbrainz-summit/2018-doremus.html

MessyBrainz

- First there was discussion on closing open pull requests.
- The project still requires a lot of effort before it gets close to production.
- The algorithms used to create clusters still needs some good way of checking if the algorithms work correctly.
- After the clustering of entities using present MBIDs. We need to do some kind of string matching for further clustering.

Main Agenda

BookBrainz Data Import / Cleanup

Error Pages

Gender

We agreed to split “Not Applicable” out of Other, and to rename the remaining Other into something else if the community can reach a reasonable level of consensus on what that something else should be.

Google Code-In

A Q&A session where **q**uestions about GCI were **a**nswered. I don’t remember exactly what was asked, but most of it was pretty much repeating what is written in other documents. If you have a question about GCI, poke me (Freso) on IRC or elsewhere and I’ll happily answer. :)

Jira Integration

Mini Summit India

MusicBrainz/BookBrainz UI Design

Save CritiqueBrainz

Decision was taken to take steps to increase number of users on CB. This would include more communication (regular community and blog posts), contribution from the community and pacing up development of cross-brainz features.

Single Sign On

Yvan is looking into single sign on for both Jira and the wiki, which are the two most important bits. We want to eventually make MusicBrainz users into MetaBrainz users, but not a short-term priority.

Web Service URL Structure

We talked about trying to standardize the structure of our ws/api URLs between projects. We'll most likely follow the MB structure since it's the most used and hence well known.

Project Specific Agenda

AcousticBrainz

A post-summit meeting happened on Monday. See the details below the main summit section.

BookBrainz

Importing

Schema

We looked into the schema described on https://docs.google.com/document/d/1jBl8YwBrF0DosjZ49lVC-buel8eAc8Kja3ceVEgLk3Q/edit#heading=h.uvws1ndjpifs and made changes as needed (main difference is probably no artist credit for works, mirroring MusicBrainz). Also talked quite a bit about UI ideas to make adding a new edition with all the data less painful.

ListenBrainz

Recommendations

- Iliekcomputers, ruaok and zas will focus on the engineering work to get the Spark cluster running
- A jupyter notebook server will be set up to allow others to build and test recommendation systems
- Third parties (e.g. researches) may be given access to the cluster to run their own models
- Build an MVP with collaborative filtering and a minimal web interface
- Spread the word about the first open recommendation engine and try to get more people on board

MessyBrainz

PRs / Code Review

We still have a lot of open PRs On which Kartikeya is currently working to get those merged. Once merged we can start working on a string matching algorithm for further clustering of entities and association of MBIDs.

MusicBrainz

Areas and Wikidata

We agreed to look into either taking over (if possible) or rewriting the Wikidata import bot for areas, and making it officially supported, probably with some GCI help. Since we now have areas sorted by usage on search results, we don't need to worry about some tiny village in Guatemala called Paris showing up above the capital of France, so we can add every village if we want. At first we'll only worry about adding current areas, but eventually we might want to copy the Wikidata structure for historical areas as well (since we already have more and more historical areas by virtue of official subdivisions changing since we implemented the system, we need to eventually deal with them properly).

Data Quality

We agreed that Chhavi will design two sets of possible data quality icons, and we'll try adding them to the site during the React rewrite of the appropriate pages. We'll then decide on one of the two (with the help of the community) and push it to the main site.

Event Art Archive

We had a request from the Internet Archive to decide on a structure for the Event Art Archive so they can start setting it up.

"Poster" will stay the front type, and there won't be any concept of a "back" image like in the CAA. We didn't finalize any other types yet, but we'll come up with a list soon and run them by the community.

A few other misc. things we discussed:
- It'd be good to keep the same thumbnail sizes as the CAA.

- We probably won't have an image type for photographs of the actual event to start (it was suggested those belong on Wikimedia Commons or elsewhere), but this might change in the future if many people request it.

- We have an "eventartarchive.org" domain parked and talked about whether it's a good idea to bifurcate the project from coverartarchive.org in that way, instead of just extending the CAA API to support coverartarchive.org/event/. We decided we'd prefer to keep everything under a central domain (so, have coverartarchive.org/event/ be the canonical API) but perhaps have eventartarchive.org redirect.

Mockups

New Edit System

We quickly discussed a few possibilities for this (long-term) project. We decided to think about it during the next year and have it as a main topic for the next summit.

React shift

All the Template Toolkit frontend of MusicBrainz is being switched to React. Work will continue on that (estimated time to completion ~6 months), with the hope of finishing long enough before the next schema change so that we can actually have a proper schema change this time (and finish the unimplemented schema changes from 2016). We will concentrate on Artist, Release and Release Group pages at first, since they see the most hits, and see if that frees bandwidth to e.g. let Bing and Google index us faster.

Virtual Machine and New Search

Other Breakout Sessions

Postgres-BDR

PythonBrainz

Copied from
https://docs.google.com/document/d/114YuClRO3soZWQOHmqDGH3vsjWIc0DEQ0HOLbjkulZA/

1) Port AB from python 2.x to python 3.x as all other projects that use BU use python 3.x and we have to write BU code to work for both python 2.x and 3.x.
2) Different versions of python libraries are used in different projects for libraries like flask, SQL-alchemy etc. We should use same versions if possible.t
3) Data dump module should be created in BU as some of code is similar for LB, AB, MsB, CB.

4) Don’t mock queries and put data into SQL lite database so that we can actually test queries.
5) Serialization is done manually we may use sqlalchemy_dst to do the same. Making serialization easier and quicker for MusicBrainz DB module (example).
6) Documentation should be done (we prefer read the docs for that) for BU modules. Maybe just for custom flask, redis and not for musicbrainz_db module in GCI. Inline documentation that is also rendered in readthedocs would be ideal.
7) Different ways of validation, serialization in mbdata and BU musicbrainz_db module. We should use one for those in both.
8) MB data has a sort of incomplete API.
9) CB should use BU’s musicbrainz_db module.
10) Docs on how to HACK BU (How to build, test BU).
11) Use Pipenv for more projects (OTHER-334)
12) MeB should adopt mbdata

AcousticBrainz post-summit meeting