Adoption by Book Lovers

GSoC 2020 Proposal for Internet Archive's Open Library

OpenLibrary.org is the world's best-kept library secret: Let's make it easier for book lovers to discover and get started with Open Library.

Tabish Shaikh Indian Institute of Technology, Jammu.

I've been a contributor to the Open Library project for over two years now. I joined the community because I was inspired by Aaron Swartz and his vision for an Open Library. Previously building the Book Sponsorship Program as an IASoC intern, I see Google Summer of Code 2020 as an opportunity to make a significant community impact through my proposal for a Distribution Program, to play a small role in the open-source community by providing people access to information and resources to reach their fullest potentials.

Email: tabish.shaikh91@gmail.com

Website: https://tabshaikh.github.io/portfolio/

Twitter: https://twitter.com/tab_tabshaikh

GitHub: https://github.com/tabshaikh

Location: Pune, India.

Timezone: UTC+5:30 hours

University: Indian Institute of Technology, Jammu.

Major: Computer Science and Engineering (CSE)

Degree level: Final year undergraduate, BTech. (Bachelor of Technology)

Graduation year: August 2020

Why Open Library Matters

Open Library is a non-profit library website run by Internet Archive which aspires to be a catalog of every book. The catalog spans 26 million book editions, 4 million of which are made available online, in digital form, for patrons to read or borrow through the Internet Archive's Controlled Digital Lending (CDL) program.

Open Library meaningfully affects the lives of a wide range of patrons:

1. Student's and Researcher's: They can find more books that they can't find in their bookstore, but can be found nearby or online only.
2. Free Library Creators, Open Source Software Developer's, Archivers - including historians, genealogists, technology adapters (archiving in new media - like the internet): By making Open Library discoverable to them we can improve our catalog, content and find more volunteer's for our community.
3. Book lover's, enthusiasts, trend followers - like events or new books by authors, people looking to read books in other languages.

Services

- Free books!
- Audio reader!
- Discovery! (what else did this author write?)
- Keep track of books! (what page am I on? wishlist? what have I read?)
- Search-inside! Find a quote you're looking for
- Easy barcode scanner!

Problem

As the age-old adage goes, "if a tree falls in the woods does it make a sound"? If you were surprised to learn of any of the services Open Library offers, you're not alone. Book lovers need to discover openlibrary.org in order to benefit from its value. Several indicators give us confidence that Open Library could be serving a much larger audience:

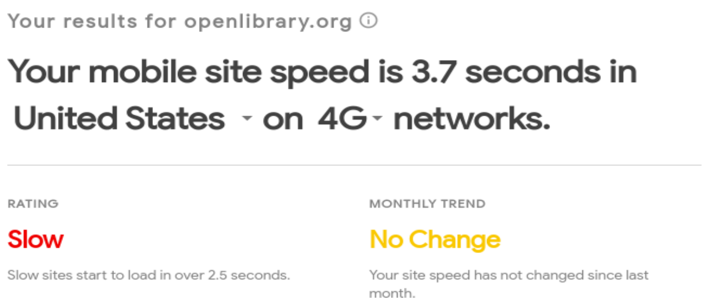
Page Rank. Today, Open Library's international alexa rank is #11,079. Compare this to goodreads.com which is a top #300 website, a website which doesn't even have books to borrow.

Number of Patrons. If Goodreads' members are some indicator of the market size, there are more than 90 million book lovers on the web and Open Library only serves 3% of the book-lovers market (3M patrons).

Book Utilization. We can also surmise that Open Library is underutilized because Of the Internet Archive's 1.2M borrowable books, only 73K (6%) are checked out meaning -

Technical Audits.

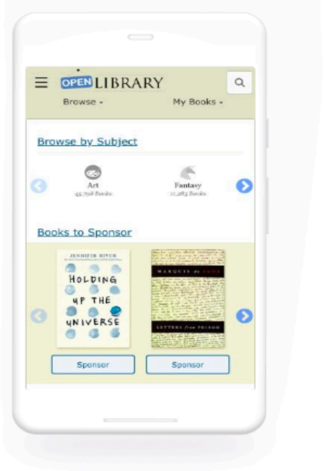
- Many book lovers discover Open Library through Google search, however Google audits reveal we are being penalized because:
    - ○ The load time of **OpenLibrary's homepage** mobile site is **3.7s** in the United States on a 4G network(https://www.thinkwithgoogle.com/feature/testmysite/) which is very slow and according to DoubleClick study by Google, **we may be losing approx. 53% of our mobile site visits.** Also for desktop site our loading time is **2.7s.**



Your results for openlibrary.org ⓘ

Your mobile site speed is 3.7 seconds in United States ▾ on 4G ▾ networks.

RATING	MONTHLY TREND
Slow	No Change
Slow sites start to load in over 2.5 seconds.	Your site speed has not changed since last month.



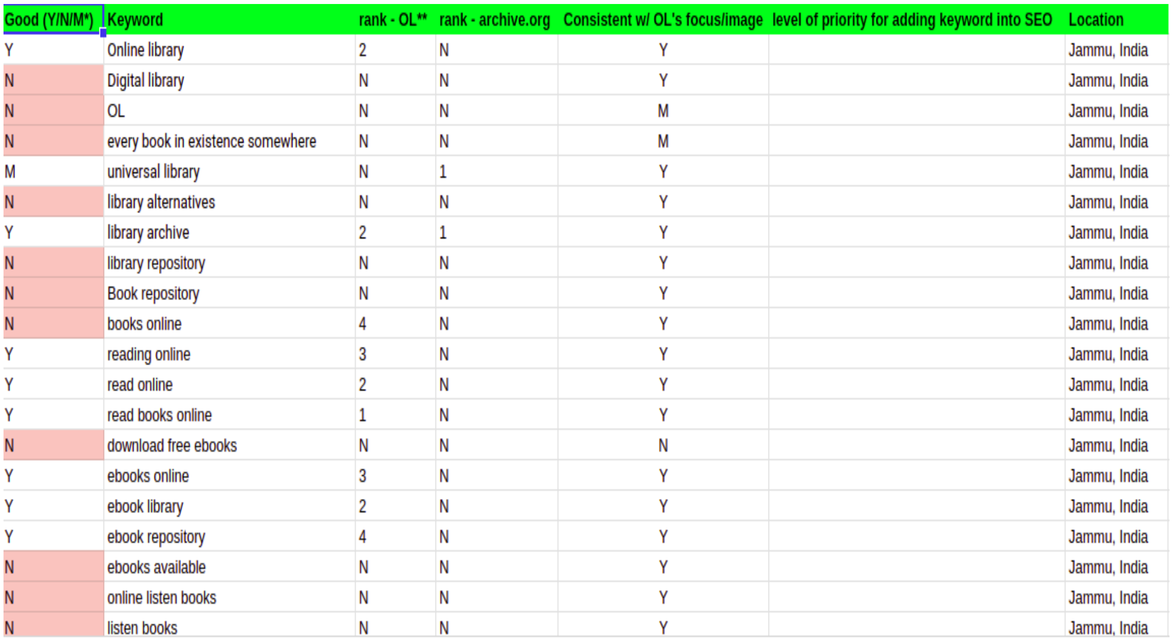
Performance plays a major role in the success of any online venture. Here are some case studies that show how high-performing sites engage and retain users better than low-performing ones. Pinterest increased search engine traffic and sign-ups by 15% when they reduced perceived wait times by 40%. A couple of case studies where low performance had a negative impact on business goals: The BBC found they lost an additional 10% of users for every additional second their site took to load. DoubleClick by Google found 53% of mobile site visits were abandoned if a page took longer than 3 seconds to load.

Performing the **lighthouse performance audit** for Open Library's home page, author page, and edition's page also reveals that for the homepage the performance score was **58/100 for desktop** and **62/100 for mobile**.

The page performance for various other pages is shown below:

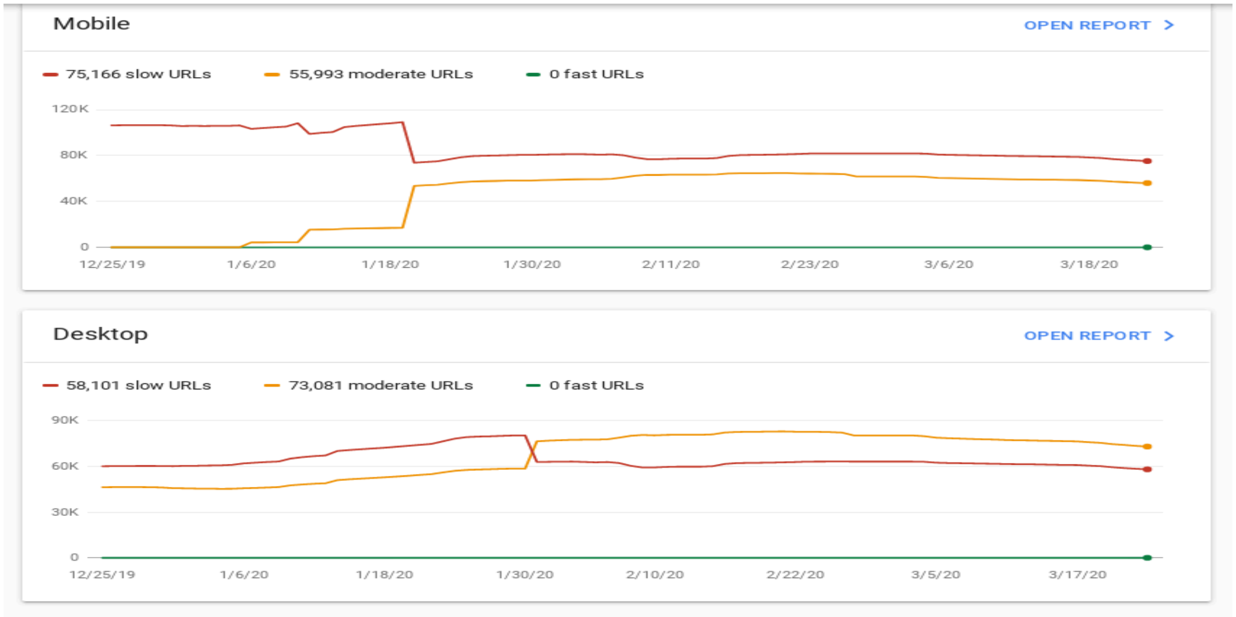
	Desktop (out of 100)	**Mobile (out of 100)**
Home Page	58	62
Author Page	80	87
Edition's Page	52	59

- **Missing meta `description` for keywords:** While performing keyword search audit on various search engines like google.com, ecosia.org, bing, yahoo we saw that for many keywords the search engine did not show up openlibrary.org result which led us to find that the /home page did not have meta description for many important keywords.

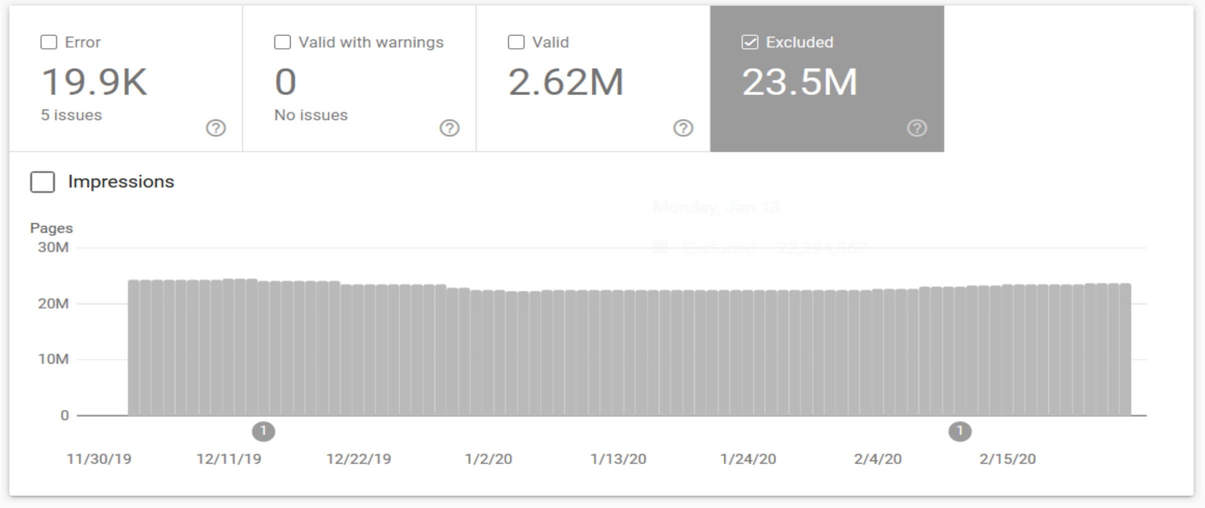


Good (Y/N/M*)	Keyword	rank - OL**	rank - archive.org	Consistent w/ OL's focus/image	level of priority for adding keyword into SEO	Location
Y	Online library	2	N	Y		Jammu, India
N	Digital library	N	N	Y		Jammu, India
N	OL	N	N	M		Jammu, India
N	every book in existence somewhere	N	N	M		Jammu, India
M	universal library	N	1	Y		Jammu, India
N	library alternatives	N	N	Y		Jammu, India
Y	library archive	2	1	Y		Jammu, India
N	library repository	N	N	Y		Jammu, India
N	Book repository	N	N	Y		Jammu, India
N	books online	4	N	Y		Jammu, India
Y	reading online	3	N	Y		Jammu, India
Y	read online	2	N	Y		Jammu, India
Y	read books online	1	N	Y		Jammu, India
N	download free ebooks	N	N	N		Jammu, India
Y	ebooks online	3	N	Y		Jammu, India
Y	ebook library	2	N	Y		Jammu, India
Y	ebook repository	4	N	Y		Jammu, India
N	ebooks available	N	N	Y		Jammu, India
N	online listen books	N	N	Y		Jammu, India
N	listen books	N	N	Y		Jammu, India

- **Google Search Console** also shows that we have **0** fast loading pages.



- **Google not indexing 23.5M OpenLibrary.org Sitemap URLs:**



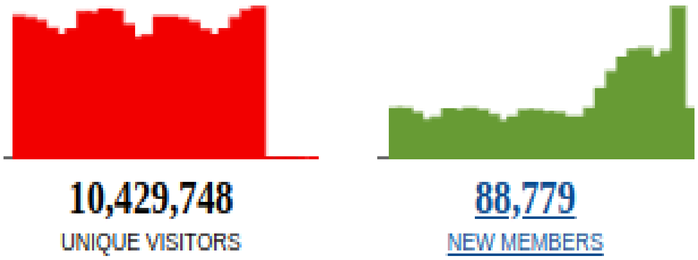
☐ Error
19.9K
5 issues

☐ Valid with warnings
0
No issues

☐ Valid
2.62M

☑ Excluded
23.5M

- **Up to 20% of patrons who discover Open Library drop off at registration because the form is unreliable:** We have also seen a shocking 1/5 of users (i.e. 500 a day) are hitting issues when creating an account (either email or username taken, or captcha broken), therefore the signup process should be fixed because of which users may be dropping off. It is also seen in /stats where the number of unique IPs is 10M but the number of new members is 88K (in a period of 28 days). (Link)



Adoption by Book Lover's

By improving OpenLibrary.org's distribution, performance, and book utilization, there's evidence to believe Open Library can join other top 500 websites like Goodreads in becoming a first-stop for book lovers.

What does adoption look like?

Adoption means Open Library may be discovered by more book lover's who utilize its services like book lending. Adoption occurs when patrons discover OpenLibrary.org, they register for an account, and they engage with its services, like the reading log and lending library. Therefore, measuring growth in account registrations, improvements in site performance, and referred traffic are good proxies for determining if adoption is happening.

Target

In 2018, Open Library welcomed 500k new patrons to the service. In 2019, this number grew to 800k (+60%) new patrons to the service. This year, because our efforts will focus on distribution, we believe we can increase this rate by an additional 30% and welcome ~1.5M patrons in 2020.

Impact

For Open Library, success would mean reaching our target of 1.5M new patrons, Improving page performance rating and load times. With increased numbers of new patrons we would see an increase in the number of sponsorships, borrowing, read-logging, and increased number of list creations.

Hypothesis

In terms of site performance, the DoubleClick study by Google shows that 53% of patrons drop off if page load is more than 3 seconds. For similar reasons, there are SEO penalties to having slow-loading web pages. Because Open Library has millions of slow-loading pages, we believe that focusing on performance and mobile, and desktop) and achieving a score which may have a scalar impact on improving its rank in search engine results and increase retention through the lending and registration funnels. In order to take full advantage of these improvements, a big opportunity is to ensure that all of Open Library's pages are indexed correctly in major search engines (currently audits reveal major gaps). Also, by pairing SEO with improvements to our signup process (real-time validation of email and username to reduce failed registrations attempts), we believe we can reduce drop-off of registrations by 5% and increase the number of patrons successfully registering for accounts. Finally, by making bets on social features, like sharing reading logs, we believe we can increase the number of patrons that may discover OpenLibrary.org.

Project Plan:

Project Board: https://github.com/internetarchive/openlibrary/projects/27

(Note: a overall summary is provided here with links to issues of each component which describes in detail about each component)

1. **Improving website** - get more pages indexed in google
   a. **Page Performance (Desktop and Mobile) #3227**
      i. Properly size images #3223
      ii. Serve static assets with an efficient cache policy
      iii. Avoid an excessive DOM size
      iv. Twitter preview blocked for IA hosted images #1841
      v. Preload key requests #3221
      vi. Serve images in next-gen formats #3222
      vii. Preconnect to required origins #3224
      viii. Avoid flash of invisible text #3226
      ix. Ensure text remains visible during webfont load
      x. Minimize main-thread work
      xi. Reduce JavaScript execution time
      xii. Avoid chaining critical requests
      xiii. Keep request counts low and transfer sizes small
   b. **Markup**
      i. Preview error's for Facebook and Reddit #3212

   ii. Schema.org tags for authors #3231
   iii. Missing meta `description` for keywords #3234
  c. **Content**
   i. Add Table of Contents text from IA to book pages #3237
2. **Outbound / Distribution: Promoting website**:
   a. Twitter borrow-link bot #3255
   b. Improving Signup Process #3256

Considerations

We want to minimize the amount of manual, repeat work we have to do (i.e. engineering, manually updating lists, manually tweeting, manually submitting to google)

We want to maximize the number of patrons we reach who are going to meaningfully engage with Open Library:

- Borrow
- Star Rating
- Lists, Reading Log (sharing, using)
- Searching
- Sponsoring a book

Inbound Distribution Ideas:

- A /resources page on Open Library which links to other useful services we love like... Librivox, https://www.readinga-z.com/curriculum-correlations/us-state-standards/
- If I "Want to Read" a book, can OpenLibrary show me the Reading Logs of other patrons who have "Want to Read" this book?
- Social Carousels: Show a carousel on Open Library (homepage) of top books twitter mentioned each day or week.
- Monthly Newsletter
- Book challenges
- https://www.reddit.com/r/bookshelf/ 58k members (books on display -- may just be good for "tweeting" pictures). Import through OCR?

GSoC Program Logistics

Mentor(s)

Michael E. Karpeles (@mekarpeles)

Meeting Timings: As the Open Library community meeting takes place on Tuesday, I would like Friday/Saturday for a meeting as, if I am stuck somewhere it would help me clear my doubts easily without wasting a lot of time. ½ an hour to one hour of meeting time would suffice.

Components and estimated time:

Sr No.	Components	Estimated Time(For completion)
1.	Preview error's for Facebook and Reddit #3212	**Approx. 4 days:** 1. 1 day - reviewing schema markup and locating the error 2. 1 day for fixing and making a pr 3. 1-2 days for testing on dev and merging into master.
2.	Twitter preview blocked for IA hosted images #1841	**Approx. 1-2 days:** 1. 1 day - Changing robots.txt on any or all servers where the image is hosted. 2. 1 day Making pr, testing and debugging on dev and merging.
3.	Google not indexing 25M OpenLibrary.org Sitemap URLs #3096	1. Approx. 3-4 days for finding out all the possible reasons why our sitemap is not being indexed using google search console. 2. Approx. 2-3 days for updating sitemap and making pr. 3. According to google documentation testing could take upto 25 days for updated sitemap to be indexed
4.	Schema.org tags for authors #3231	**2 days:** 1. 1 day- review schema.org documentation and make the

		changes for author's mentioned in the issue. 2. 1 day - Making pr, testing and debugging on dev and merging.
5.	/home missing meta \`description\` for keywords #3234	**Approx. 2-3 days:** 1. 1-2 days - Deciding all the important keywords with @Brittany and updating the meta tags on the home page accordingly. 2. 1 day - Making pr, testing and debugging on dev and merging.
6.	Preload key requests #3221	**Approx. 2-3 days:** 1. 1 day - Identifying the elements on the homepage, author and edition page where link=preload needs to be added and making the appropriate changes. 2. 1 day - Making pr, testing and debugging on dev and merging.
7.	Properly size images #3223	**Approx. 3-4 days:** 1. 3 days - for researching, implementation about the strategy to implement from the ones mentioned in the proposal. 2. 1 day - Making pr, testing and debugging on dev and merging.
8.	Serve images in next-gen formats #3222	**Approx. 3-4 days:** 1. 2-3 days - for deciding(together with the whole community) the format to convert the images and implementing it. 2. 1 day - Making pr, testing and debugging on dev and merging.
9.	Preconnect to required origins #3224	**Approx. 2-3 days:** 1. 1-2 day - Identifying the elements on the homepage, author and edition page where preconnect origins need to be added and making the appropriate changes. 2. 1 day - Making pr, testing and

		debugging on dev and merging.
10.	Avoid flash of invisible text #3226	**Approx. 4-5 days:** 1. 3 days for implementing the proposed solution on the homepage, author and edition page’s. 2. 2-3 days for testing and making it compatible for all browsers.
11.	Various other Issues Related to Performance seen during lighthouse audit #3227	There are 7 points whose priorities and issues have not been created: 1. Create issues for the remaining points. 2. For 7 points it would take approx. a month to fix those.
12.	Add Table of Contents text from IA to book pages #3237	1. Figure out a way to get the text data from the book reader. 2. Convert the raw text data into the content format used by openlibrary.
13.	Improving Signup Process #3256	1. There are 3 epic issues related to this task: a. #2053 - update backend API endpoints b. #2055 - Add real-time field validation for email, username, password to Account Registration Form front-end c. Design and implement a new signup process using google, github auth.
14.	Allow Patrons to Enable "Public" Reading Log at Account Creation #2058	1. This task entails modifying the user registration page (front-end) + https://github.com/internetarchive/openlibrary/blob/master/openlibrary/plugins/upstream/account.py#L203 (back-end) to support enabling this setting from the account creation form.

Required Deliverables

Phase I (Week 1 - Week 4): Improving Page Performance + Fixing Schema Issues

Sr. 1. Preview error's for Facebook and Reddit #3212

Sr. 2. Twitter preview blocked for IA hosted images #1841

Sr. 3. Google not indexing 25M OpenLibrary.org Sitemap URLs #3096

Sr. 4. Schema.org tags for authors #3231

Sr. 5. /home missing meta `description` for keywords #3234

Sr. 6. Preload key requests #3221

Phase II (Week 5 - Week 9): Improving Page Performance + Signup Process

Sr. 13. Improving Signup Process #3256

Sr. 7. Properly size images #3223

Sr. 8. Serve images in next-gen formats #3222

Sr. 9. Avoid flash of invisible text #3226

Sr. 10. Issues Related to Performance seen during lighthouse audit #3227

Phase III (Week 9 - Week 11): Improving Signup + Adding Table of Contents

Sr. 13. Improving Signup Process #3256

Sr. 12. Add Table of Contents text from IA to book pages #3237

Phase IV (Week 11 - Week 14): Adding Table of Contents and Wrap Up

Sr. 12. Add Table of Contents text from IA to book pages #3237

Schedule:

June 1	Coding officially begins!

Week 1 + Week 2 (June 1 - June 14)

Sr. 1. Preview error's for Facebook and Reddit #3212

Sr. 2. Twitter preview blocked for IA hosted images #1841

Sr. 3. Google not indexing 25M OpenLibrary.org Sitemap URLs #3096

Week 3 + Week 4 (June 14 - June 28)

Sr. 3. Google not indexing 25M OpenLibrary.org Sitemap URLs #3096

Sr. 4. Schema.org tags for authors #3231

Sr. 5. /home missing meta `description` for keywords #3234

Week 5 + Week 6 (June 28 - July 12)

June 29 18:00 UTC	Mentors and students can begin submitting Phase 1 evaluations
July 3 18:00 UTC	Phase 1 Evaluation deadline

Sr. 6. Preload key requests #3221

Week 7 + Week 8 (July 12 - July 26)

Sr. 13. Improving Signup Process #3256

Sr. 7. Properly size images #3223

Sr. 8. Serve images in next-gen formats #3222

Week 9 + Week 10 (July 26 - August 9)

July 27 18:00 UTC	Mentors and students can begin submitting Phase 2 evaluations
July 31 18:00 UTC	Phase 2 Evaluation deadline

Sr. 9. Avoid flash of invisible text #3226

Sr. 10. Issues Related to Performance seen during lighthouse audit #3227

Week 11 + Week 12 (August 9 - August 23)

Sr. 13. Improving Signup Process #3256

Sr. 12. Add Table of Contents text from IA to book pages #3237

Week 13 + Week 14 (August 23 - August 31)

Sr. 12. Add Table of Contents text from IA to book pages #3237

August 24 - 31 18:00 UTC	Final week: Students submit their final work product and their final mentor evaluation
August 31 - September 7 18:00 UTC	Mentors submit final student evaluations
September 8	Final results of Google Summer of Code 2020 announced

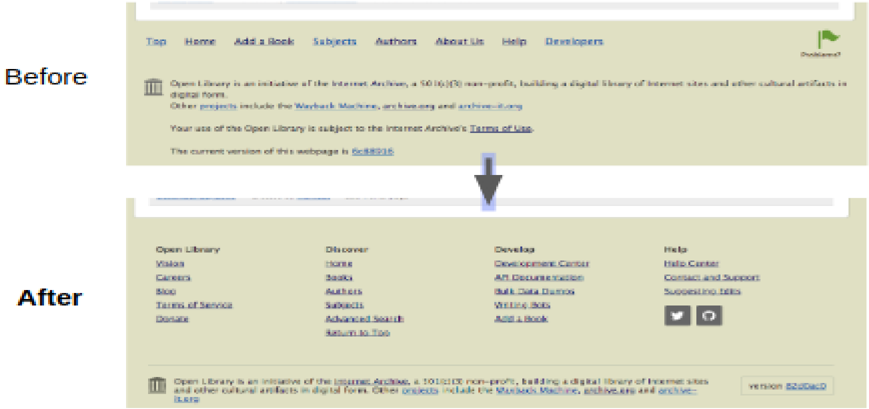
Why me?

Contributions to this repository

I joined the Open Library community in May, 2018 and the first task I worked upon was redesigning our website footer.

Issue number: https://github.com/internetarchive/openlibrary/issues/908

https://github.com/internetarchive/openlibrary/issues/642



Top Home Add a Book Subjects Authors About Us Help Developers

Problems?

Open Library is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form.
Other projects include the Wayback Machine, archive.org and archive-it.org

Your use of the Open Library is subject to the Internet Archive's Terms of Use.

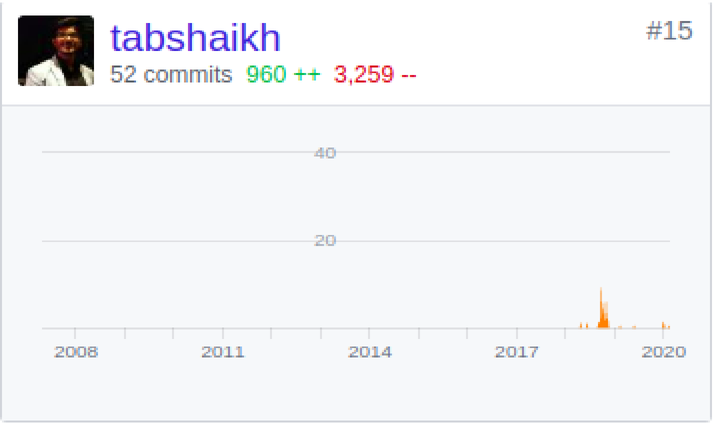
The current version of this webpage is 6c88916

Open Library	Discover	Develop	Help
Vision	Home	Development Center	Help Center
Careers	Books	API Documentation	Contact and Support
Blog	Authors	Bulk Data Dumps	Suggesting Edits
Terms of Service	Subjects	Writing Bots	
Donate	Advanced Search	Add a Book	
	Return to Top		

Open Library is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include the Wayback Machine, archive.org and archive-it.org

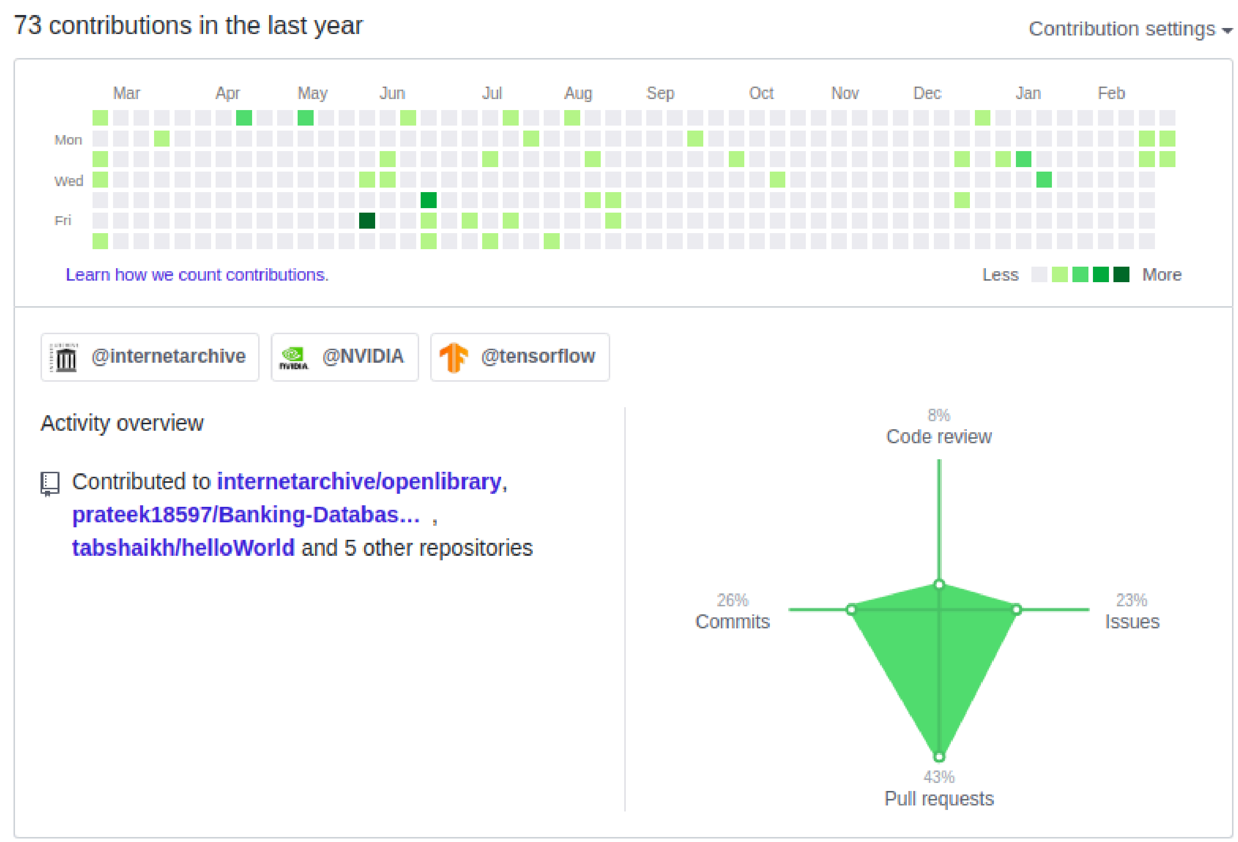
version 82d0ac0

Then went onto resigning the mobile login experience, making numerous front-end fixes and implementing style rules with Jon Robson, helping with internationalization and Hacktoberfest coordination. Helped in developing the Book Sponsorship Program as an Internet Archive Intern which raised over $55k. I have also contributed to documentation such as the Contributor's FAQ, changes in readme after removing vagrant and tens of minor fixes.


tabshaikh #15
52 commits 960 ++ 3,259 --

Other Projects:

Contributed to organizations such as Tensorflow (migrations of tf.contrib loss functions to tf.addons), OSSN (contributed towards fixing the sidebar and improving the website's security to the repository ossn/fixme which connects new contributors to issues of different organizations according to the technology, labels, difficulty, etc)



73 contributions in the last year

Contribution settings ▾

Learn how we count contributions.

Less ⬜ 🟩 🟩 🟩 🟩 More

@internetarchive @NVIDIA @tensorflow

Activity overview

Contributed to internetarchive/openlibrary, prateek18597/Banking-Databas... , tabshaikh/helloWorld and 5 other repositories