FOR 2020 & beyond: See our Pipeline of Future Ideas, and share your suggestions!

CC Search: 2019 Roadmap

Roadmap is subject to change. The CC Search team will adjust the roadmap at the end of each quarter, after reviewing its OKRs and the pipeline of ideas contributed by staff and community. Our GitHub issues and milestones (CC Search Active Sprint, CC Search Backlog) contain technical details on what we're currently working on. If you have specific thoughts on the roadmap, shoot us a note at support-search@creativecommons.org.

Q1: February-March

CC Search 1.0
- Complete QA sprint to make CC search beta the default at search.cc.org, including:
    - Improve the UI of CC Search to make it look more like the CC home page UI
    - Browse by provider
    - Search by creator
    - HTTPS proxy for images so we don't have 404s
    - Test whether Lists feature is desirable as it stands and decide to continue or kill it

Front End
- Fix, document & ship Attribution Frame MVP
- Ship project documentation & contribution guidelines for Front End
    - Improve test coverage
    - Create contribution guidelines
    - Document project's high-level structure and workflow
    - Miscellaneous tasks
- Develop features for tracking use of images and provide metrics that help know our audience
- Optimize front end loading times

Catalog
- Catalog 325 million works in the commons
    - Identify new platforms by analyzing the top contributors in the common crawl data set or by using existing sources (e.g. the State of the Commons reports).
    - Integrate new platforms and increase the total number of content providers by 80%. The breakdown is as follows:
        - 60% from the Common Crawl corpus
        - 20% from open APIs
- Develop the ETL (extract, transform, load) infrastructure to integrate data streams from APIs and Common Crawl
- Integrate Clarifai into the data pipeline and keep the catalog up-to-date

- Optimize the database to improve execution time of complex operations by at least 25%
- Ongoing catalog work
  - Flickr and other API providers
  - Clarifai tagging
  - Common Crawl providers

API
- Ship project documentation & contribution guidelines for API
- Ship strategy for opening up the CC Search API to third-party developers
  - Figure out whether we're giving Internet Archive everything
  - Make sure to talk to Ryan to figure out what data we do and do not want to share
- Publish API documentation publicly
- Develop and ship API authentication for identifying consumers
- Provide attribution for a CC licensed work server-side via the API
- Improve search result accuracy
- Define and meet internal [SLAs](#) for the CC Catalog API
- Fully automate and document all day-to-day operations tasks

Usability
- Determine metrics to track
- Plan first monthly usability test
- Plan out technical work needed to support the Lean Startup ideas of tracking actionable metrics and split testing

Everything else
- Ship product vision & user research strategy
- Update provider review process to align with new strategy and to start following in Q2
- Devise game plan with Director of Open Education on how to integrate open texts in Q3
- Help Development team develop 3-5 pitches to current / new funders for CC Search
- Ship internal memo that articulates strategy about what providers and content we include in the catalog
- Write one post updating the community on progress via CC blog and Medium
- Google Summer of Code preparation
- Promotion of community efforts:
  - 3 blog posts
  - 2 conference session submissions or meetups

Q2: April-June

CC Search Front End
- Ship CC search 1.0 as default search
- Redesign home page, including front page filters and toolbar

- Improve one-click attribution options
- Implement and test reuse metrics through attribution tools and user feedback mechanisms
- Make source (provider) of image more prominent on image page and in attribution options
- Change "License types" category filter
- Better education (or access to) on CC and CC licenses
- Ship a version of the CC watermark with some reuse metric in place
- A way to search for all works from a specific provider on CC Search
- Improve relevance of search results and guidance on how to narrow
- Make CC search accessible (including infinite scroll)
- Make sure UX on mobile works w/redesign, design and performance

CC Catalog
- Review open text provider APIs
- API/Common Crawl carryover from Q1 & ongoing work
- Identify and review 30 collections for inclusion in CC Search (via Common Crawl and other means), including exploring:
    - Cleveland library
    - Wikimedia Commons / Wikidata
    - Behance
- Explore options for tagging images, including AI development research to build our own and Google Vision
- Automate or otherwise streamline data review process
- Develop a resource for providers with desirable content to integrate into CC Search
- CC Catalog Workflow Documentation – Airflow
- Define an efficient relational schema for adding open texts to catalog

CC Catalog API
- Browse by provider
- API endpoint for images that are similar to a given image.
- Improve relevance of phrases
- improve search result relevance / Prioritize usability issues like search quality
- Get team to review and sign SLA once it's finalized
- Implement suggestions in SLA required to hit 99.99% availability
- More dynamic frontpage? API-driven in some way? (need to change curation set manually right now, so can we do this more automatically?)

Usability, metrics, and user research
- Ship new user survey and ccsearch@ feedback page
- Set up tracking reuse doc or dashboard for testing different metrics / dashboard for CC Search team to easily track how our metrics are doing
- Anonymize licensebuttons.net logs and share
- Review licensebuttons.net data

- Run two monthly usability tests, incorporating learnings from each into sprints

Developer community
- Support Google Summer of Code by reviewing proposals, answering questions from students, reviewing external code contributions promptly, and mentoring accepted students
- CC Search: set up develop branch and automatically deploy latest develop to a server
- CC Search Front End contribution guidelines and documentation

Planning & Strategy
- Plan integration of open texts for Q3, including provider engagement, with Director of Open Education
- Execute Comms plan for shipping CC search 1.0
- Global Summit sessions on CC search
- Global Summit session on user research
- Global Summit session on engineering
- Seek a partner for prototyping a CC Catalog API partnership
- Every product team member write one post updating the community on progress WRT their respective work via CC blog (+ Medium if applicable)

Q3: July-September

CC Search UI
- Fix navigational issues:
  - ~~footer and infinite scroll are not compatible; reevaluate infinite scroll as design choice~~
  - users being taken back to the top of the search results every time they click the back button from a single search result
- Add new filters:
  - clip art, vector graphics, photographs
  - image dimensions (size)
  - NSFW (mature) works
  - geo-coordinates where image was taken
- Button to report images so that we can identify and remove content that users license incorrectly (e.g. when the user does not have the right to license that content) or tag mature content appropriately
- General bug fixes and improvements

CC Search backend (CC Catalog API)
- Search algorithm improvements
  - ~~Use descriptions of works in the search algorithm~~
  - Use popularity data gathered from providers like Behance and Flickr to contribute towards ranking of images
  - Experiment with stemming and lemmatization changes in Elasticsearch

- - Experiment with weighing sources differently
  - Release stable 1.0 version of the CC Catalog API
  - ~~Collection of anonymous usage data to identify high quality images and identify any search algorithm blindspots~~
  - ~~Plan out a Push API that organizations can use to automatically upload their collections to the CC Catalog~~
  - *Stretch goal:* Cache thumbnails rather than retrieving them from source's servers

CC Catalog
- Use APIs for some image sources than we previously used CommonCrawl for (e.g. Behance)
- Integrate 10 image sources using CommonCrawl data and 5 image sources via API
- Identify 30 new image sources to consider adding in Q4
- Plan out adding open textbooks as a searchable content type in Q4
- ~~Integrate images from Wikimedia Commons~~
- ~~Infrastructural improvements to test cases and ETL workflow~~
- ~~Plan out adding OCR and object recognition data for all images~~

Usability
- ~~Set up A/B testing infrastructure~~ and run experiments comparing UI designs as well as efficacy of different search algorithms
- Seek out contract designer for general usability improvements for current and future features
- Conduct usability tests every month
- Set up a dashboard to track usage metrics internally

Developer Community
- ~~Google Summer of Code mentorship for CC Search browser plugin, CC Catalog data visualizations, new CC license chooser~~
- Guide for external contributors to work on CC Catalog and Search tasks they are interested in building but that are not yet on our roadmap

Q4: October-December

CC Search UI
- Make open textbooks searchable

CC Search backend (CC Catalog API)
- Implement Push API

CC Catalog
- Index open textbooks from various sources
- CC Catalog Documentation & Contributors guide (Github) that allows users to discover new providers and write scripts to help us get that data into the Catalog

- Implement adding OCR and object recognition data for all images

2020 Priorities

- Add audio to CC Search
- Add 3D designs to CC Search