Building Anime Scene Search Engine

Story of building and running a world-wide popular image search service: "WHAT: What Anime Is This?"

Presented by @soruly

URL of my slides

https://github.com/soruly/slides

About Me

- Graduate from **CUHK** (BSc) Computer Science
- Former committee of **Animation and Comic Society**
- Part-time in **Oursky** before graduate
- Created **whatanime.ga**
- Game Developer in **Derivco Hong Kong** now

https://about.me/soruly

The Motivation

Help Identifying The Anime
Know the anime but can't find the scene
Anime Guessing Game

Image reverse search engines:

- Google Image - Results are limited
- TinEye - never works on anime
- iqdb - tailored for doujin artwork, not anime
- SauceNAO - covers iqdb plus game CG

SauceNAO recently expanded its **database coverage** to recent anime

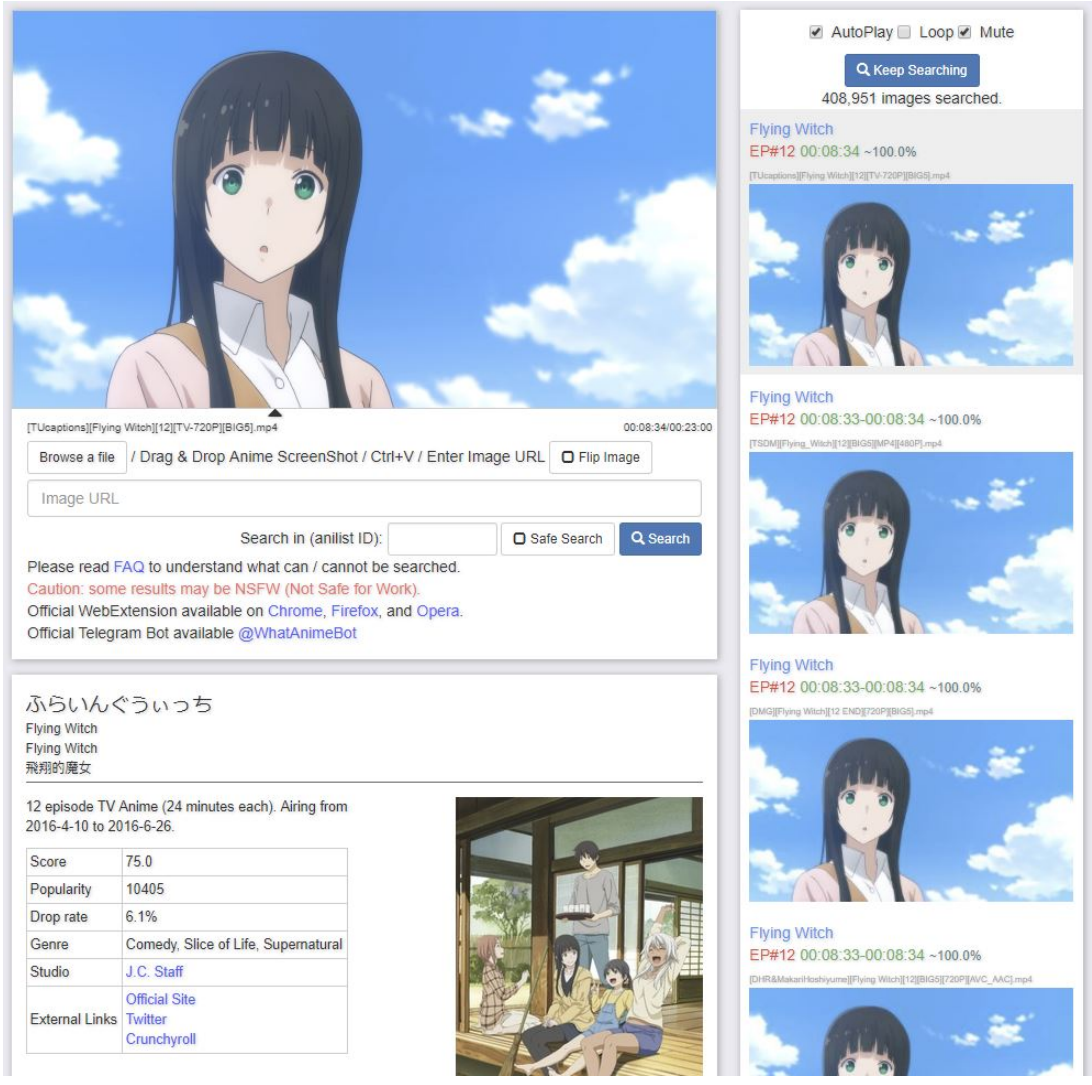
Build_Your_Own_X

CBIR research projects/demos/open source projects [edit]

Name		External Image Query	Metadata Query	Index Size (Estimate, Millions of Images)	Organization Type	License (Open/Closed)
Windsurf	A general framework where different altern effectiveness and eff	Yes	No		University	Open but not free
Anaktisi	This Web-Solution im accurately retrieving	Yes	No	0.225M	University	Open
FIRE	Open source query b can easily be combin	No	No		University	Open
MIFile	Image similarity sear	No	No	100M	Research Institute	Open
Pastec	C++ LGPL index and	Yes	Yes		Private company	LGPL
BRISC	BRISC is a recursive	Yes	No		University	GPL
digiKam	Extensive photo man detection of duplicate	Yes	Yes	Desktop-based	KDE	GPL
Caliph & Emir	Creation and Retriev	Yes	No	Desktop-based	University	GPL
GNU Image Finding Tool	Query by example im	Yes	No	Desktop-based	GNU	GPL
imgSeek	photo collection man	Yes	No		Individual	GPL
LIRE	Java GPL library for bag of visual words a	Yes	Yes		University	GPL
akiwi	akiwi is a semi-auton	Yes	Yes	15M	University	Closed
ALIPR	Developed by Penn S	Yes	Yes		University	Closed
ISSBP	Similar Image Search	Yes	Yes	free-beta limited to 4k	Private	Closed

https://en.wikipedia.org/wiki/List_of_CBIR_engines

Demo



search results on the right, anime info below is provided by anilist.co

How does it work?

whatanime.ga has nothing to do with:

- AI
- Machine Learning
- blockchain

whatanime.ga is kind of:

- Content-based image retrieval (CBIR) search engine
- computer graphics program
- big data

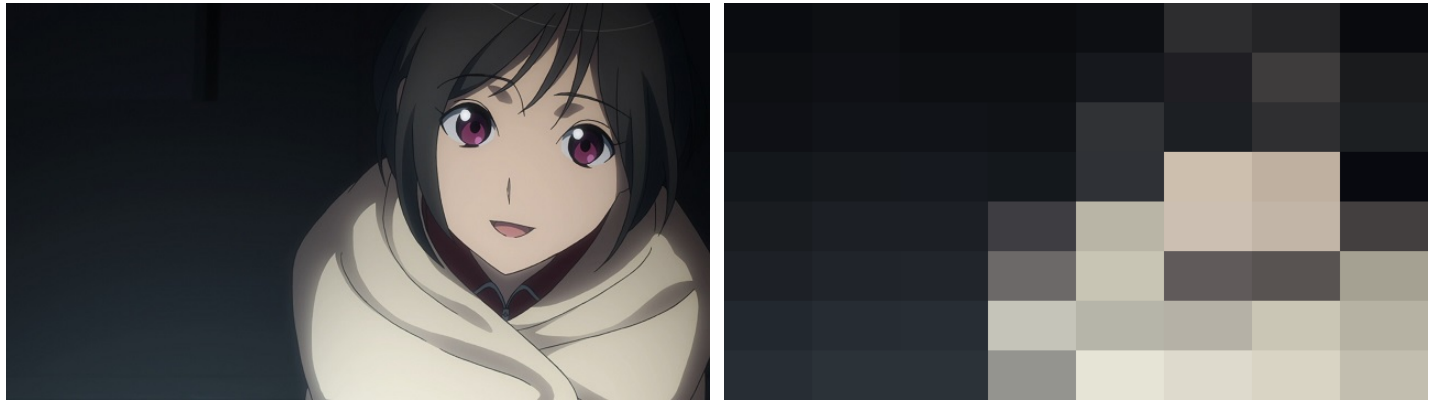
Common image descriptors:

- Color Layout, Edge Histogram, Opponent Histogram, ScalableColor, etc.

whatanime.ga only uses Color Layout due to Hardware limitations

Brief idea of Color Layout

- one of the MPEG-7 standard (I didn't invent this)
- Raw Image -> Partition to 8x8 blocks -> take average color of each block -> convert color space to YCbCr -> DCT transform (quantize) -> Zigzag scanning



- extracted image feature (fingerprint):

```
FQYLBAQRFgoYFBANEBIQDw0QCw0PDxAeEhEQDhAfDQ8PEA8=
```

https://en.wikipedia.org/wiki/Color_layout_descriptor

Color Layout for each frame

Raw Video -> Extract all frames by **ffmpeg** -> Extract image features by **LIRE** -> Deduplicate hashes -> Append timestamp -> Load into **solr** (database)

```
<?xml version='1.0' encoding='UTF-8'?>
<add>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0</field>
    <field name="cl_hi">FQYCEBAQEBAQEBAQEBAQEBAQEBAgEBAQEBAgEBAQEBA=</field>
    <field name="cl_ha">89f 2c3 cc6 a57 540 926 32d 9aa e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    af1 2b f81 dd 1bf cc5 c2c 530 78b c6c cb8 d2f d69 dbb a43 ed2 205 eb7 e36 b0c 8aa 3ad a14 661 b0e 89
  </doc>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0.0833333</field>
    <field name="cl_hi">FQYCDxAQEBAQDxAPEBAQEBAQEA8QEBAgEBAQEBAgEBAQEBA=</field>
    <field name="cl_ha">89f 2c3 cc6 a57 540 926 32d 9ea e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    af1 2b f81 dd 1bf cc5 c2c 730 78b c6c cb8 d2f d69 dbb a43 ed2 205 eb7 e36 b0c 8aa 3ad a14 661 b0e 88
  </doc>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0.166667</field>
    <field name="cl_hi">FQYDDxAQEBAQDxAPEBAPEBAQEA8QEA8fEA8QEBAfEBAPDxA=</field>
    <field name="cl_ha">89f 2c3 cce a57 540 926 32d 9ea e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    2b f81 2dd 1bf cc5 c0c 730 78b c6c cb8 d2f d69 dbb a43 ed2 205 e37 e36 b0c 8aa 3ad a14 661 b0f 88b 5
  </doc>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0.25</field>
    <field name="cl_hi">FQYEDxEREBERDxAPEBAPEBAQEA8QEA8fEA8QEBAeEA8PDxA=</field>
    <field name="cl_ha">89f 2c3 cce a57 540 926 32d 9aa e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    2b b81 2dd 1bf cc5 c0c 730 78b c6c cb8 d2f d69 dbb a4b ed2 205 eb6 e36 b0c 8aa 3ad a14 661 b0f 88b 5
  </doc>
```

`image similarity = similarity of two binary strings`

Comparing image features, at scale

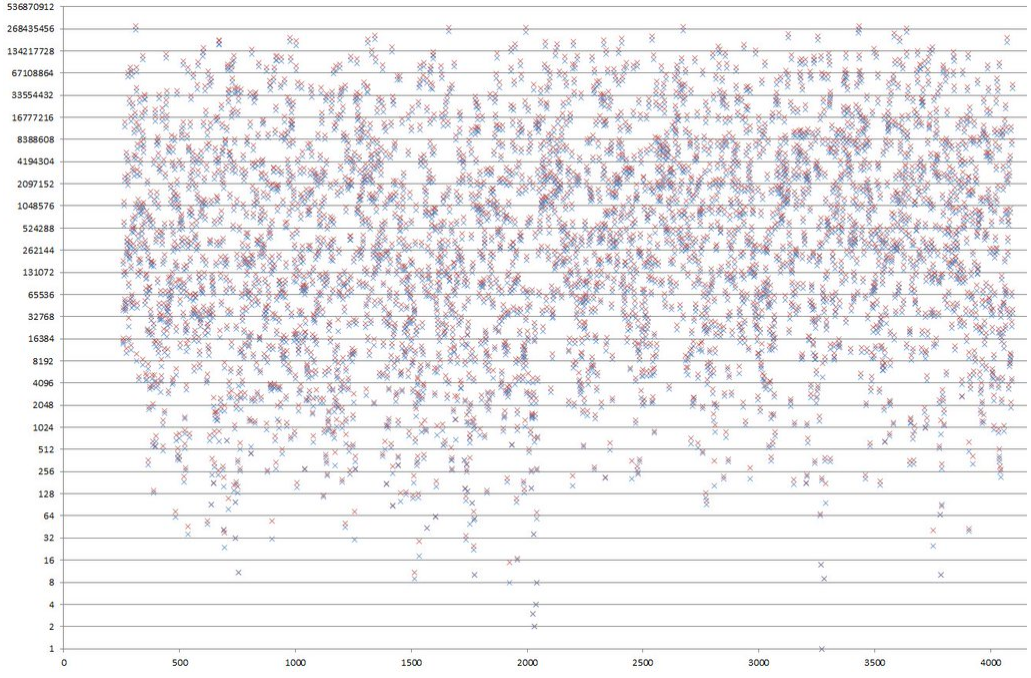
- 30000+ hours of video (~2,600,000,000 frames)
- Deduplicate frames with a running window
- There are still ~804,000,000 images to compare
- Reduce search area by Locality Sensitive Hashing
- Comparing ~800 million strings -> compare ~1 million strings

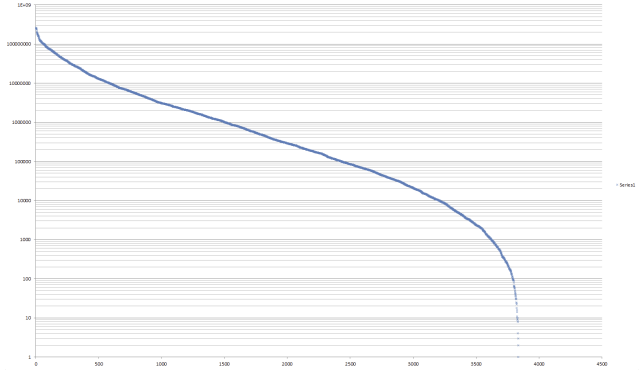
```
<?xml version='1.0' encoding='UTF-8'?>
<add>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0</field>
    <field name="cl_hi">FQYCEBAQEBAQEBAQEBAQEBAQEBAgEBAQEBAgEBAQEBA=</field>
    <field name="cl_ha">89f 2c3 cc6 a57 540 926 32d 9aa e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    af1 2b f81 dd 1bf cc5 c2c 530 78b c6c cb8 d2f d69 dbb a43 ed2 205 eb7 e36 b0c 8aa 3ad a14 661 b0e 89
  </doc>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0.0833333</field>
    <field name="cl_hi">FQYCDxAQEBAQDxAPEBAQEBAQEA8QEBAgEBAQEBAgEBAQEBA=</field>
    <field name="cl_ha">89f 2c3 cc6 a57 540 926 32d 9ea e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    af1 2b f81 dd 1bf cc5 c2c 730 78b c6c cb8 d2f d69 dbb a43 ed2 205 eb7 e36 b0c 8aa 3ad a14 661 b0e 88
  </doc>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0.166667</field>
    <field name="cl_hi">FQYDDxAQEBAQDxAPEBAPEBAQEA8QEA8fEA8QEBAfEBAPDxA=</field>
    <field name="cl_ha">89f 2c3 cce a57 540 926 32d 9ea e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    2b f81 2dd 1bf cc5 c0c 730 78b c6c cb8 d2f d69 dbb a43 ed2 205 e37 e36 b0c 8aa 3ad a14 661 b0f 88b 5
  </doc>
  <doc>
    <field name="id">2014-07/ALDNOAH.ZERO/[KTXP][Aldnoah.Zero][03][BIG5][720p].mp4?t=0.25</field>
    <field name="cl_hi">FQYEDxEREBERDxAPEBAPEBAQEA8QEA8fEA8QEBAeEA8PDxA=</field>
    <field name="cl_ha">89f 2c3 cce a57 540 926 32d 9aa e82 c37 c6b 55f a71 fe6 38f 58 e8 5fe b18 ff7 f2
    2b b81 2dd 1bf cc5 c0c 730 78b c6c cb8 d2f d69 dbb a4b ed2 205 eb6 e36 b0c 8aa 3ad a14 661 b0f 88b 5
  </doc>
```

still not fast enough

Comparing image features, at scale

- Choose 1 out of 100 hash terms for search, starting from the least populated one. (image: cluster ID vs population)
- dermotte (author of **LIRE**) accepted this idea and implemented this as IDF into liresolr (see more in semanticmetadata.net)





June 2016: ~1k of daily users, search time varies from 1-40sec

Image search, at scale, with speed

- Cache search results in redis
- Reduce search accuracy
- Disable swap
- Replace SATA SSD with NVMe SSD

June 2017: ~2k of daily users, search time varies from 1-30sec

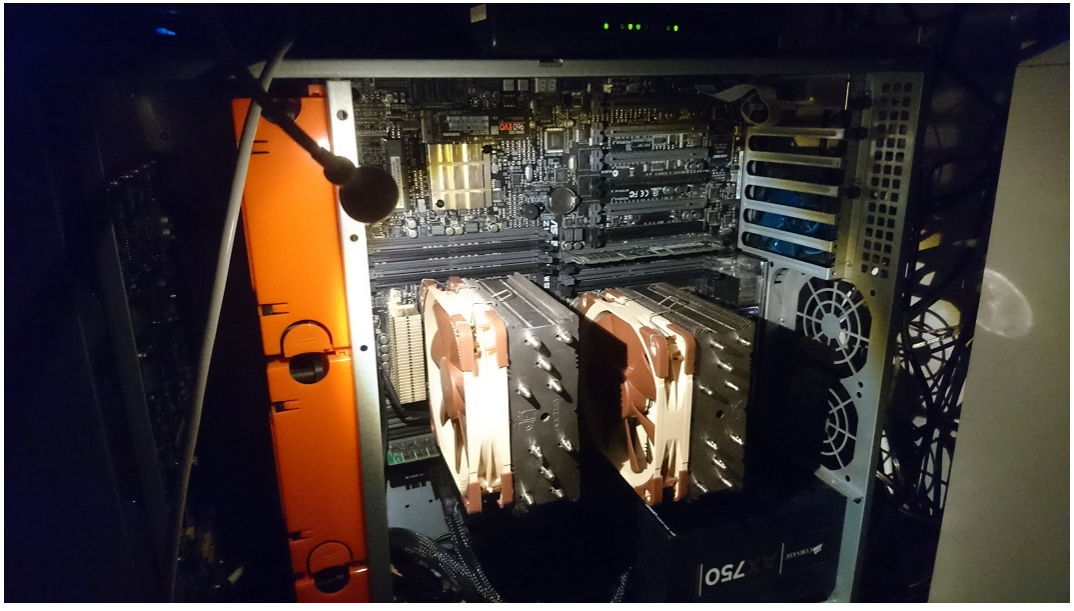
Data keeps growing, Traffic keeps rising

~~"We need to build a wall"~~

Minimum search time becomes 10-30sec, server keeps overloading

More Cores, More RAM

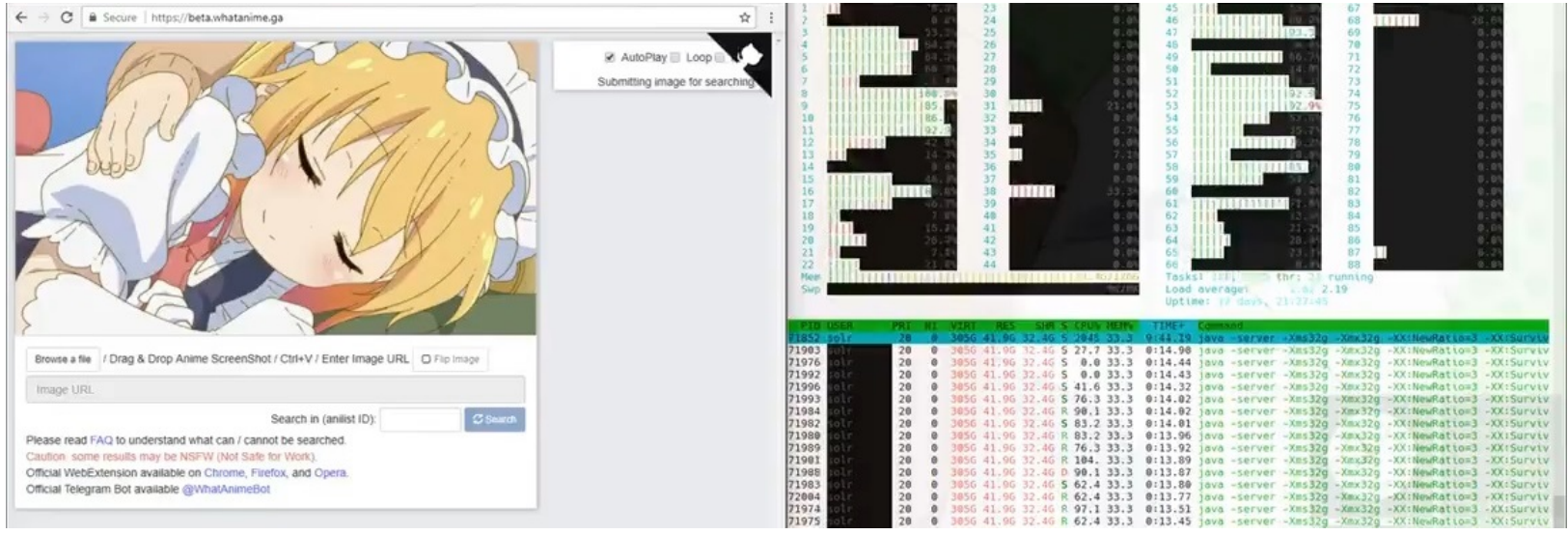
- Old server: just a quad-core Desktop PC with 32GB RAM
- New server: 2 x E5-2696v4 (44 Core 88 Threads), 256GB RAM



Dec 2017: Slightly better, but server still keeps overloading

Squeezing All CPU Powers

- **liresolr** is single thread... and **solrcloud** does not work well with plugin schemas
- Split index into 32 smaller databases (solr cores)
- Balance cores by loading hashes into least populated cores
- Search all databases in parallel, and combine results
- All database (solr cores) are hosted in one server



April 2018: See how it utilize all cores

More RAM

- Database size (index) is 150GB now
- Use vmtouch to keep the database in RAM

https://twitter.com/soruly/status/1030122636725051392

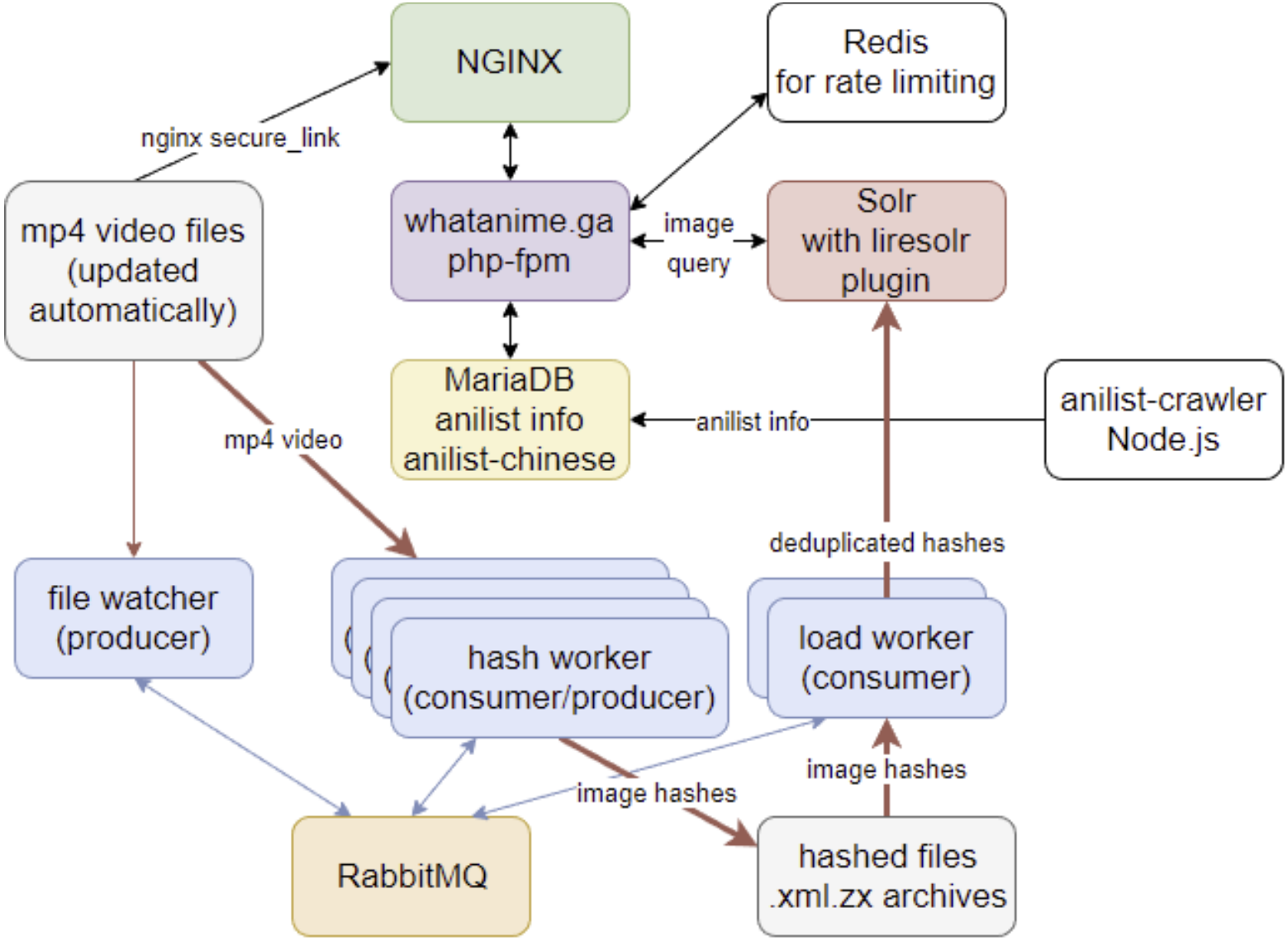
Aug 2018: Search time consistantly 0-2sec

Auto black border detection and crop

- Using findContours from OpenCV to crop black borders

similarity: with black border 89.4%, without border 96.3%

System Overview



All parts of whatanime.ga are open source!

https://github.com/soruly/whatanime.ga
https://github.com/soruly/whatanime.ga-WebExtension
https://github.com/soruly/whatanime.ga-telegram-bot
https://github.com/soruly/anilist-crawler
https://github.com/soruly/anilist-chinese
https://github.com/soruly/liresolr
https://github.com/soruly/sola

API Docs:

https://soruly.github.io/whatanime.ga

「你不需要很厲害才能開始，但不開始就沒辦法很厲害」

It's time to build your own Anime/Video Scene Search Engine!

Future Plans

whatanime.ga will not:

- cover comics / artworks
- allow search by timecode
- increase duration of preview
- add ads to websites

whatanime.ga will:

- Increase database coverage (fill in missing anime and maybe crawling from youtube)
- Reduce duplicates in database
- Support multiple image descriptors like FCTH (Fuzzy Color and Texture Histogram)
- Rebuild web front-end for language and mobile support
- move to a new domain (considering trace.moe)

Get Involved!

If you love whatanime.ga , share it!

Report bugs on GitHub, Telegram or Discord

Support soruly on Patron

- https://www.patreon.com/soruly

Support soruly via PayPal

- https://www.paypal.me/soruly

Join official pages / channels:

- Discord Channel
- Telegram Channel
- Facebook Page
- Google+

Credit

- Dr. Mathias Lux for LIRE Project and liresolr
- Josh for providing anilist.co info via Anilist API
- bちゃん, Desmond, FangzhouL, Snadzies, WelkinWill, yuriks, and 16 other Patrons
- ccd0 for integrating whatanime.ga into 4chan-x
- Xamayon for integrating whatanime.ga into saucenao.com
- egoist for docute that makes API docs
- bestshow for reporting an XSS issue regarding CVE-2017-6390
- fans that help me to answer questions on discord
- whoever shared, complained and made suggestions
- whoever bring anime to this world ❤

Thank you!

https://about.me/soruly

https://twitter.com/soruly

https://telegram.me/soruly

https://www.instagram.com/soruly

See more awesome ACG projects:

https://github.com/soruly/awesome-acg