Hello and welcome to my TED talk, “Dumps are not backups”. I have only 15 minutes to convince you of this, so let’s get right to it.

Why dumps != backups, in 15 mins

- What are the dumps?
- What's in them?
- How the dumps are like backups
- How the dumps are NOT like backups
- A few words about actual backups!
- What are the dumps? (Redux)

Here's what we're going to cover. Please hold your questions but write them down and I'll check in with y'all at the end.

What are the dumps?

- Datasets available for public download,
- for public mirroring,
- and for upload to www.archive.org

(Hey, *I* liked it!)

The salient points here are that these are available to the public, to anyone at all, to grab a copy. Likewise they are mirrored; we have sites that rsync from us directly, but the more sites want to grab copies from us or from one of the existing mirrors, the better. We would love it.

Um so that logo. Well it was just an idea because we didn’t have one. A community member nixed it. Any proposals? :-D

What is in the dumps?

Type ONE:

Public content from all wiki projects, in sql or xml format

- Some db tables in sql format, can be imported into a new wiki
- Data in xml format, convertable to sql for import into a new wiki, via special scripts (fast), or imported directly via importDump.php (SLOW)
- Useful to researchers, analysts, editors, WMF teams, and others

Type TWO:

Public datasets of other content, in various formats

- Cirrussearch data, Wikidata entities, Commons MediaInfo, global locks, content translation pairs, adds/changes dumps, article category information, etc.
- Not suitable for import into a new wiki, but useful to researchers, analysts, editors and others

We are only interested in Type ONE, as potential backups of the wikis.

Okay, so when we talk about “the dumps” there’s really two piles of datasets we could be talking about. One is the “classic dumps”, i.e. the sql/xml dumps which have been available for download since I’m not sure when but we have datasets in the historical archives all the way back to 2001!

These have all the public content of the wikis, some of it as sql files which are just raw table dumps out of the database, and some of which are XML files because we’re having to skip some private info in those tables, and so we ask MediaWiki to tell us which things the public can see and which not.

The “other dumps” include everything else and they are in all kinds of formats: sql, rdf, json..

The important thing here is that the classic dumps are IMPORTABLE into a new wiki, the other ones not so much.

How the dumps are like backups

- They contain historical revisions of all pages
- They cover all public wikis
- They are copied to hosts not owned by the WMF
- They are copied to hosts outside of the United States
- They can be used to set up **mirrors** of all the projects

So let’s talk about how the dumps are like backups so we can get all of your arguments out of the way :-D

They do have all of the public content, including the full history, which is now billions of revisions across all of the projects!

They are copied to third party mirrors which we do’t own; this is a Good Thing (TM).

And they aren’t all hosted in the United States; recent developments may make us breathe easier about the likelihood of servers being seized, but it’s still nice to have some resilience as far as jurisdictions go.

And one can set up a mirror with these files; people have done it!

How the dumps are not like backups

Missing data!

- User account data
- Deleted articles
- Hidden revisions
- ALL THE MEDIA

We actually believe in privacy.

Of course, that’s not where the discussion ends, or we wouldn’t be having this little chat.

Backups are meant to contain all the data being backed up, in case one needs to, well, restore form the backups :-D The dumps DO NOT.

User data like email addresses, hashed passwords and so on, is all private. Any mirror restored from the dumps would be missing user accounts; users would have to create accounts from scratch with a new name, no edit count, and NO PRIVILEGES.

Some articles and revisions are “deleted” (removed from public view). Occasionally these are “Undeleted”, meaning that the public can see them again. Because deleted content is no longer accessible to the public, it does not get dumped. No restores from that data are possible on such a mirror.

Also NO MEDIA IS DUMPED WHATSOEVER.

Still not like backups

INCONSISTENCY

- Db tables are not consistent with each other
- Db tables are not consistent with article data
- Article data may not be consistent within itself
- Only fix: import XML via importDump.php (SLOW SLOW SLOW)

original

restored from the dumps

Ah but we're not done.

We want backups from which we will restore, to be a consistent snapshot of the data. The dumps are not.

Example: pages in the dump have wikitext putting them into a category that does not exist, because the page content is dumped later than the category table.

Worse example: MediaWiki versions can change between parts of a dump of a single wiki.

Really awful example: the schema could be changed mid-dump. Now we would never do this intentionally… but it could happen!

To keep everything consistent, the only option is to use importDump.php and ignore the tables. Not all data is in the content dumps, but worse than that, importDump requires parsing of every entry… a restoral using this method would take weeks or maybe months.

Nope, not like backups yet

TIMELINESS

- Dumps with full content run once a month
- In the past month there were:
- 5,182,288 edits on Wikipedia
- 15,159,499 edits on WikiData
- 23,711,774 edits on Commons

WMF engineer realizing how much data would be lost in a month

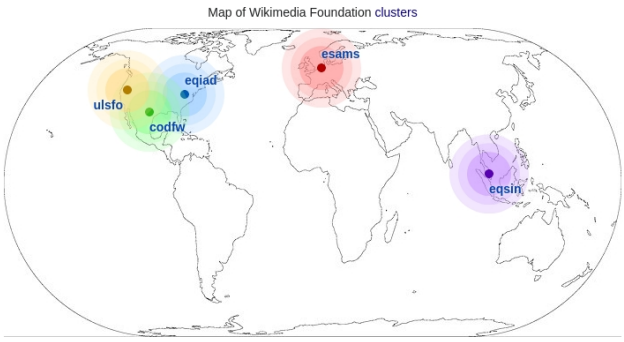
But that's still not all.

Backups from which we want to restore need to be fresh. We don't want to be missing a bunch of data from the last time we ran a backup until now. And there's just no way we're going to have daily full content dumps unless we have a LOT more hardware thrown at the problem, including dedicated database servers. I dare say that's not gonna happen.

If we were unlucky we could lose out on over 40 million edits, just look at the numbers!

You wanted backups but got these

- AVAILABILITY
- Backups should succeed and be available if we lose:
- A host – ok, we can
- A rack – mm probably
- A DC – NOPE. All dumps hardware is in ONE DC ONLY



Aaaaaand we’re still not done. Backups need to be complete and consistent and fresh but also THERE. If something dies and your backups die with it, that’s not very helpful.

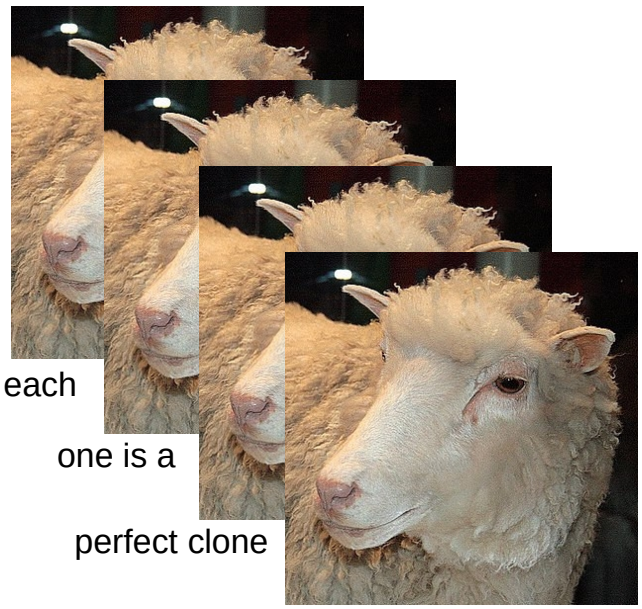
We have resiliency against a snapshot host or a dumpsdata host going away, though it’s not some fancy HA thing going on.

There’s one rack that has both a dumpsdata NFS server for generation and a snapshot host for generation, but we can lose one of each and with some scrambling still get the dumps done.

A row? Not a change. And if we lose Eqiad, it’s all over. There are no dump hosts in any other DC nor plans to have such a cluster.

My kingdom for some backups

- Bacula
- Run on all wiki dbs
- 5 times a week
- Cover public and private data
- In eqiad and codfw
- ONLY in the US, no third party copies
- No media (but there are plans)



Okay, but the good news is that there are real honest-to-goodness backups. There are actual snapshots taken of the databases themselves four times a week, stored in Bacula, with a copy in eqiad and another copy in codfw.

There are so-called logical backups of the databases generated via mydumper, and stashed also in Bacula, on a daily (?) basis.

These contain ALL the data (yay!)

These are not copied to third party mirrors, because they have the data (boo!)

And there’s no media backed up. But there are plans, talk to Jaime about that when he says it’s ok :-)

What are the dumps? (Redux)

- Besides data for researchers, analysts, editors, us:
- A guarantee of the Right to Fork
- Insurance in case WMF Turns Evil ™
- Insurance that the contents of the wiki projects is and will always remain free

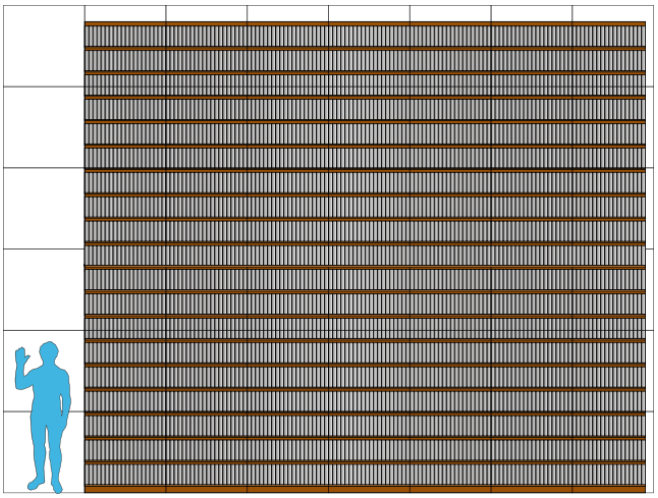
OK so given all that, what are the dumps really?

Besides having all the uses they have (see me for an incomprehensive but wordy list of some of them), they are a guarantee of the right to fork: that anyone at any time can take them freely (free as beer and free as in freedom), and set up a copy.

Not that we ever expect the WMF to turn evil, but like an umbrella against the rain, it’s always good to have that insurance :)

Why aren’t scrapers enough?

- Scrapers are:
- Slow (must run in serial, must collect billions of revisions)
- Not guaranteed (can be blocked at any time)
- Not shared (each user would have to run their own, or publish their files; why not us?)
- Not in the spirit of the GFDL/CC-BY-SA licenses (convenient access to all content, not just bits of it)



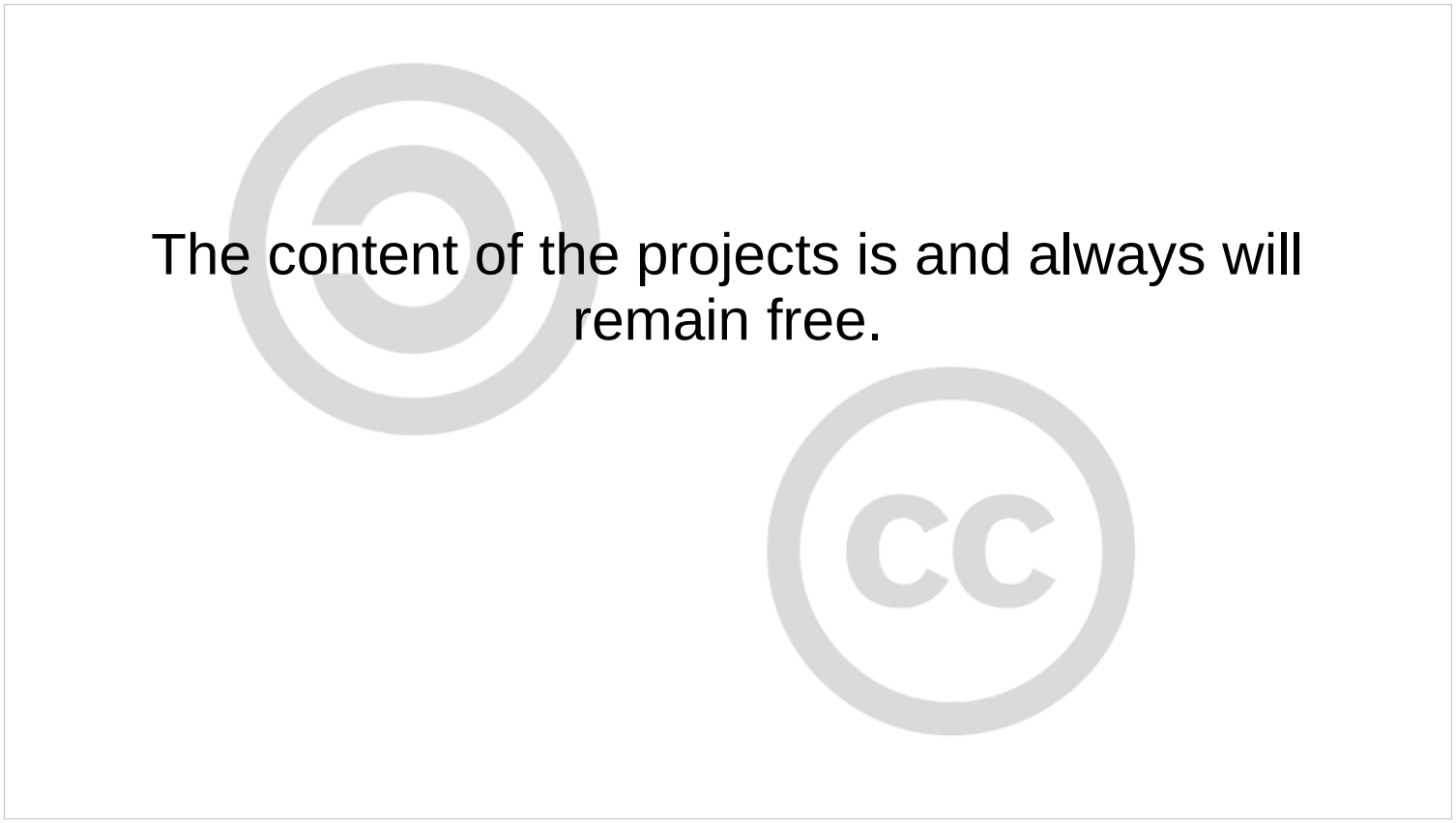
Size of English Wikipedia, August 2010

But can’t anyone get a copy already, just by running a scraper? In theory. But that’s not convenient, having to get revision information one at a time and then cobble together the collection. It might be within the letter of our open content licenses but it’s certainly not within the spirit of it. The idea of an open source license isn’t that with a lot of effort and time you can eventually get a copy of all of the content, but that it’s easy and intended you to get a copy.

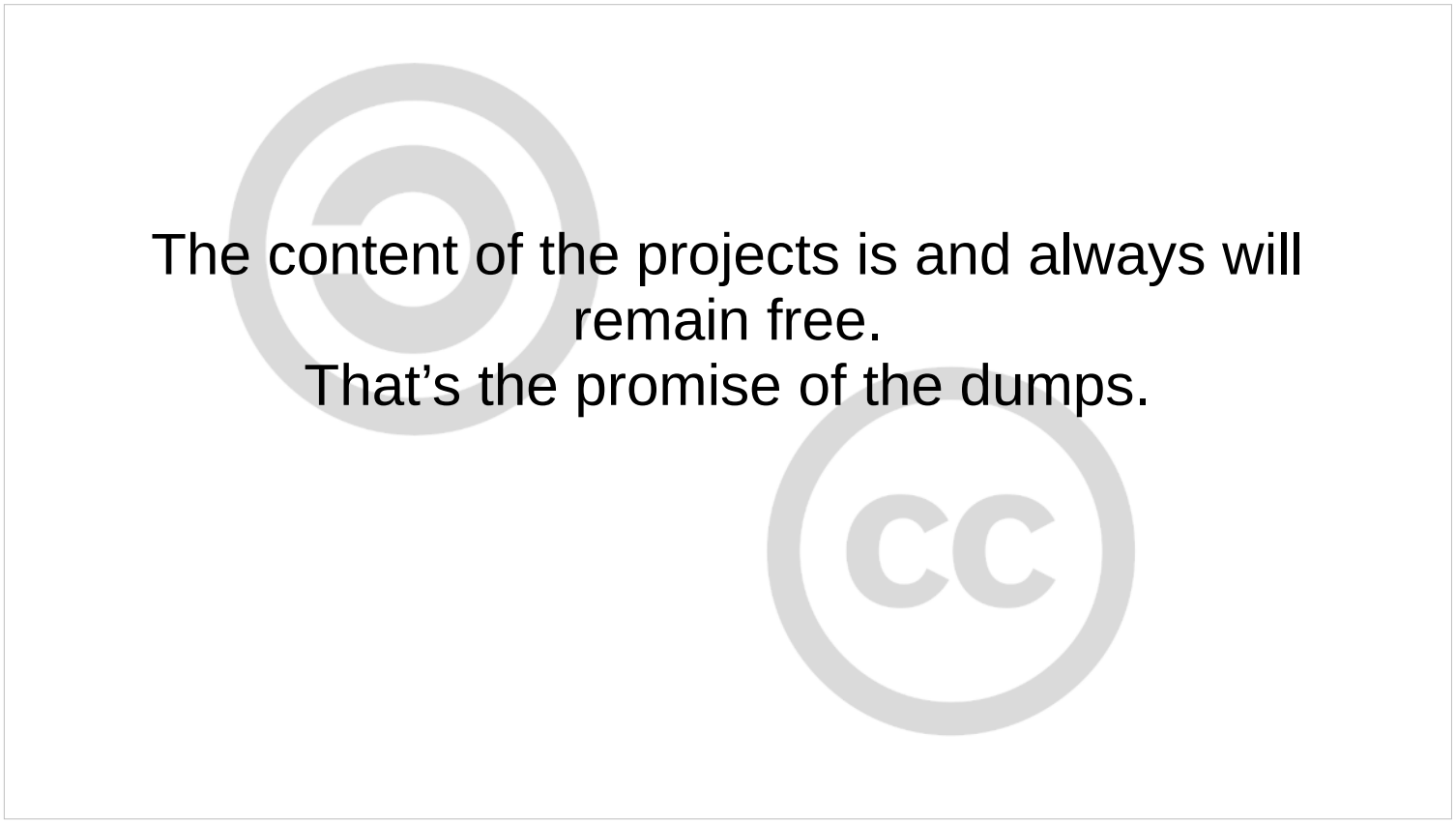
If you wonder how long a scraper might take to get it all:

Commons has 521 million revisions. Let’s say we can get 50 revisions per request to the MW api, and 10 of those a second, because we’re running in serial as the guidelines say. Then, we only have to run for 1024000 seconds. That’s 11 days nonstop. Then there’s … all the other wikis.

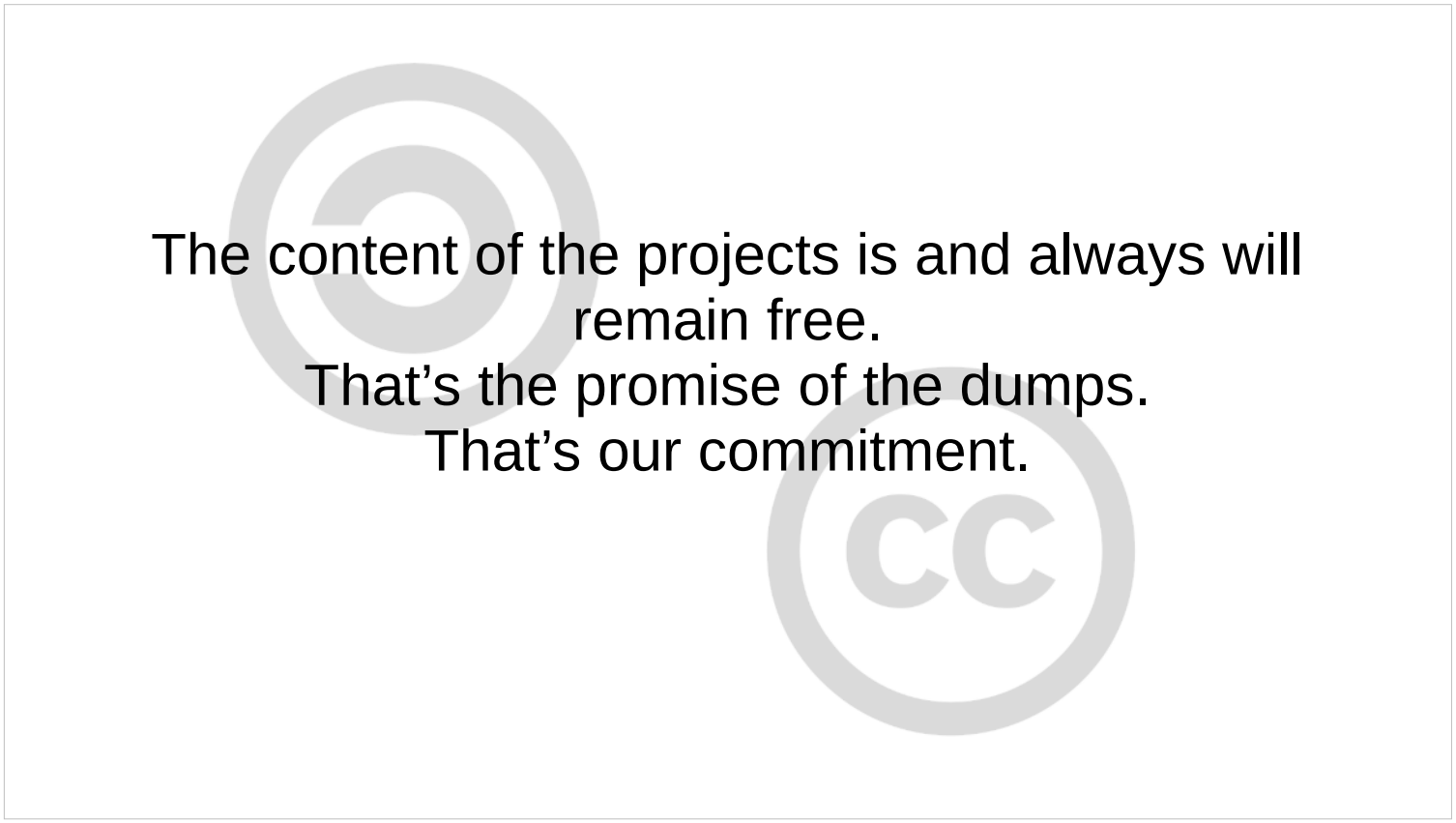
The pic: Using volumes 25 cm high and 5 cm thick (some 400 pages), each page having two columns, each columns having 80 rows, and each row having 50 characters, ≈ 6 MB per volume. As English Wikipedia has around 15887 MB of text (August 2010) ≈ 2647 volumes (2660 in illustration).



No...


The content of the projects is and always will remain free.
That’s the promise of the dumps.

...comments...



...necesssary. This is what it’s all about for me, and always has been. Cheers!

Thanks!

Questions, comments, gripes? You know where to find me:

- irc: apergos on freenode
- element: apergos
- email: ariel@wikimedia.org
- phabricator/gerrit: ArielGlenn
- on the wikis: User:ArielGlenn

Image credits

All images are copied from or derived from:

- https://commons.wikimedia.org/wiki/File:Magnetic_tape_rolls.JPG
- https://commons.wikimedia.org/wiki/File:Czech-2013-Prague-Astronomical_clock_face.jpg
- https://commons.wikimedia.org/wiki/File:Dumps_logo_black_and_white.svg
- https://commons.wikimedia.org/wiki/File:V%C4%9Btru%C5%A1e,_zrcadlov%C3%A9_bludi%C5%A1t%C4%9B.jpg
- https://upload.wikimedia.org/wikipedia/commons/thumb/0/0a/Mark_Zuckerberg_F8_2019_Keynote_%2847774200621%29.jpg/640px-Mark_Zuckerberg_F8_2019_Keynote_%2847774200621%29.jpg
- https://commons.wikimedia.org/wiki/File:Colored_Angora_Goat.jpg
- https://commons.wikimedia.org/wiki/File:20170307Ovis_aries3.jpg
- https://commons.wikimedia.org/wiki/File:Edvard-Munch-The-Scream.jpg
- https://wikitech.wikimedia.org/wiki/Template:ClusterMap
- https://commons.wikimedia.org/wiki/File:Dolly_face_closeup.jpg
- https://commons.wikimedia.org/wiki/File:Darth_vader_hot_air_balloon_1.jpg
- https://commons.wikimedia.org/wiki/File:Size_of_English_Wikipedia_in_August_2010.svg
- https://commons.wikimedia.org/wiki/File:Copyleft.svg
- https://commons.wikimedia.org/wiki/File:Cc.logo.circle.svg