Dumps are not Backups

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)

What are the dumps?

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

(Hey, *I* liked it!)

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.

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

How the dumps are not like backups

Missing data!

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

We actually believe in privacy.

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

Nope, not like backups yet

TIMELINESS

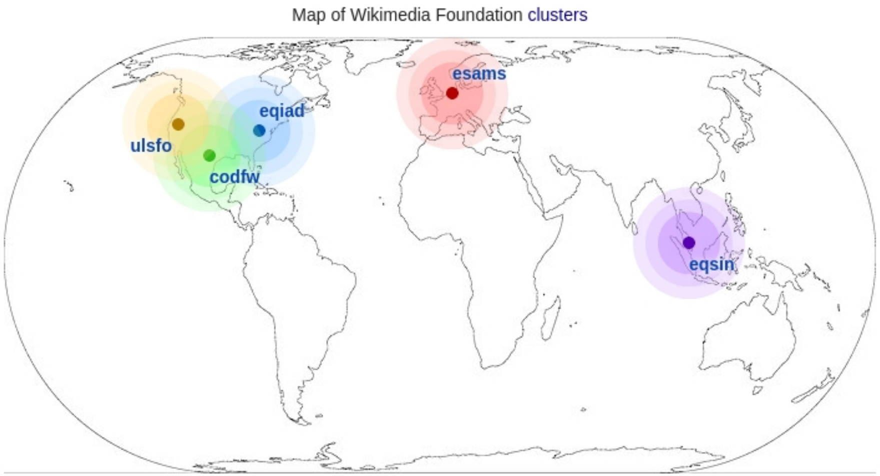
- Dumps with full content run once a month
- In the past month there were:
- 5,182,288 edits on Wikipedia
- 15,159,499 edits on WikiData
- 23,711,774 edits on Commons

© WahooArt.com – Edvard Munch – The Scream

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

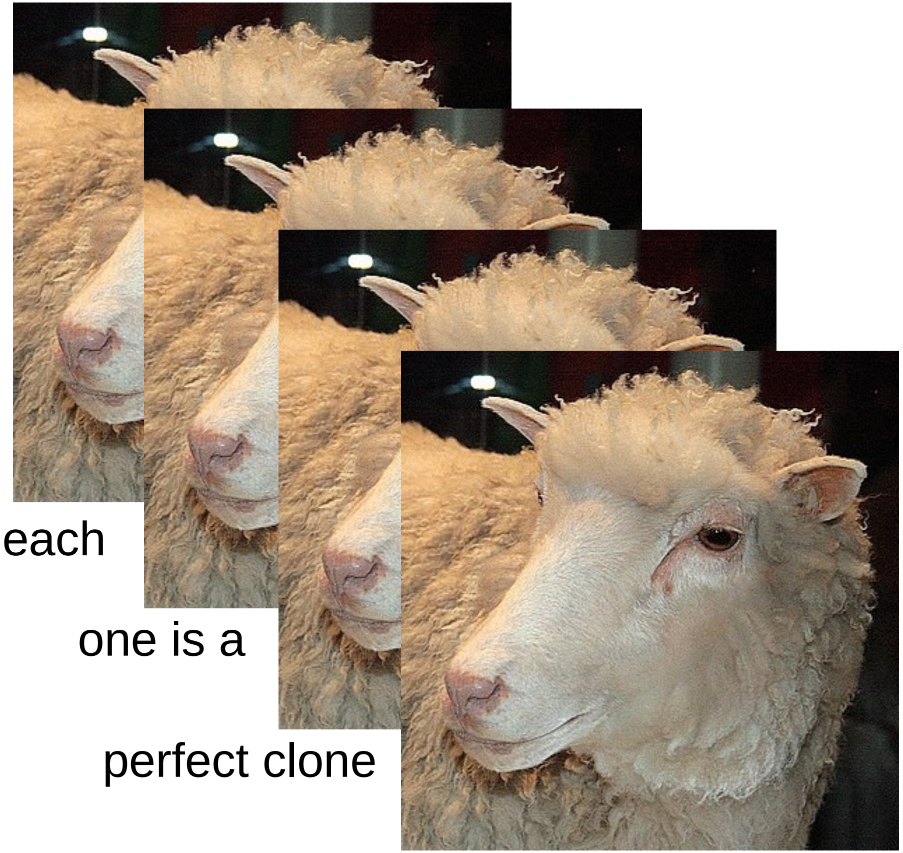
You wanted backups but got these

- 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



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)



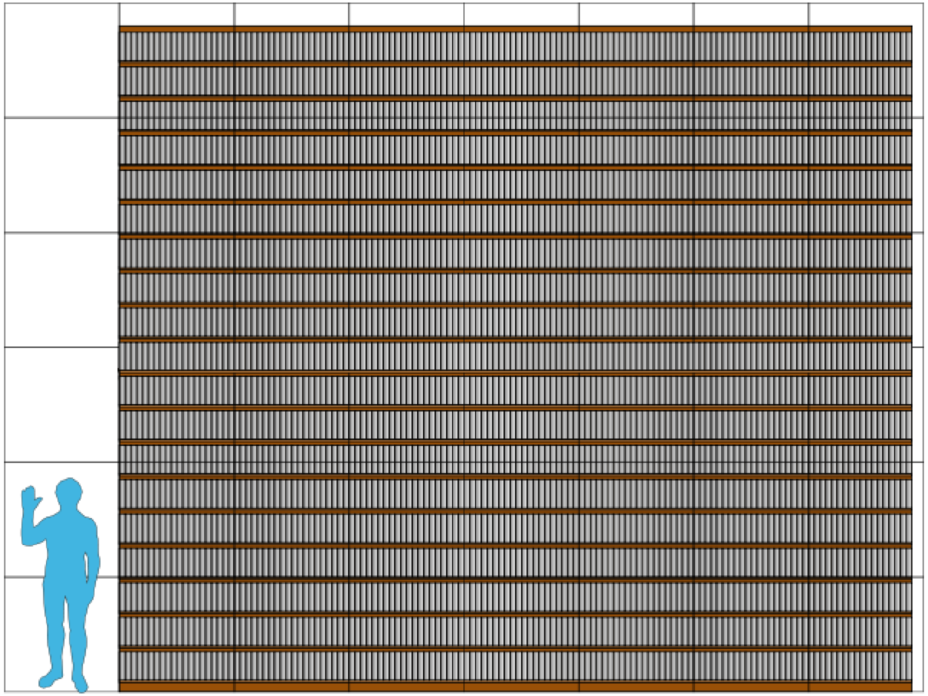
each one is a perfect clone

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

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

The content of the projects is and always will remain free.

The content of the projects is and always will remain free.

That’s the promise of the dumps.

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

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