Data dumps/FAQ

< Data dumps

Contents

Frequently asked questions about the XML dumps

What happened to the SQL dumps?

In mid-2005 we upgraded the Wikimedia sites to MediaWiki 1.5, which uses a very different database layout than earlier versions. SQL dumps of the 'cur' and 'old' tables are no longer available because those tables no longer exist.

We don't provide direct dumps of the new 'page', 'revision', and 'text' tables either because aggressive changes to the backend storage make this extra difficult: much data is in fact indirection pointing to another database cluster, and deleted pages which we cannot reproduce may still be present in the raw internal database blobs. The XML dump format provides forward and backward compatibility without requiring authors of third-party dump processing or statistics tools to reproduce our every internal hack.

We are looking to provide sql dumps compatible with a recent version of MediaWiki on a regular basis; see the notes on experimental sql dumps. Additionally, there are tools available which you can use to convert the XML files to sql; see the list of tools.

Why don't you use BitTorrent?

We focus our energy on doing things that volunteers can't do as easily. In the case of torrents, some community members have stepped up and produced torrent files; you can see the list or add to it here.

Why do you still use bzip2 and not just 7z compression?

Bzip2 uses a block-oriented compression format, which means that if a job dies partway through we can recover from that reusing most of the data, instead of having to start over. It's also pretty fast to check that a job completed, because we can just look at the last few blocks of the file instead of uncompressing the whole thing. Especially for some of our larger wikis, this is crucial for producing these files on a regular basis without corruption. There was a period of a few years without regular English language Wikipedia runs; we never want to go through that sort of thing again :-)

How big are the en wikipedia dumps uncompressed?

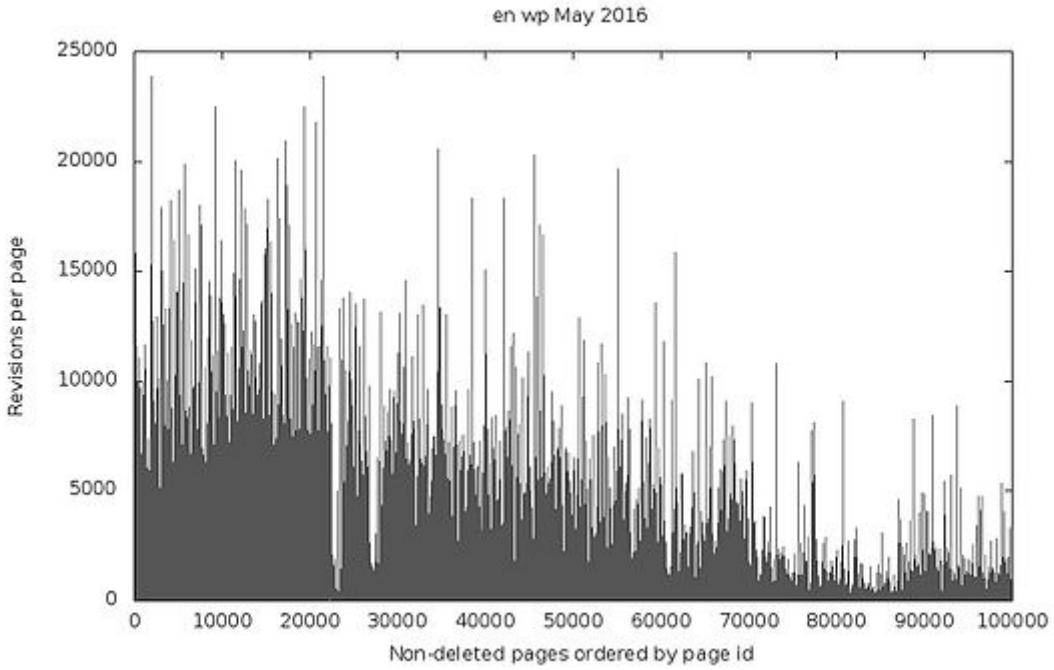
Big. I mean **really** big. As of May 2016, the XML file containing current pages only, no user or talk pages, was 57,080,072,830 bytes uncompressed. The XML file with current pages, including user and talk pages, was 127,884,910,101 bytes uncompressed. The full history dumps, all 202 files of them, took 14,371,294,447,992 bytes. For folks keeping track, that's 10.4 TiB.

You can always get the latest answer to this question by subscribing to the xmldatadumps-l (https://lists.wikimedia.org/mailman/listinfo/xmldatadumps-l) mailing list; once a month, the size of these files, uncompressed, plus the sizes for one other wiki, randomly chosen, are mailed to the list near the end of the month. Example: update for September 2018 (https://lists.wikimedia.org/pipermail/xmldatadumps-l/2018-October/001435.html).

Why do the en wikipedia dumps run in 27 pieces?

The host that job runs on is 32 cores; it turns out that running 32 jobs causes various jobs to fail, primarily those with gzip as part of a pipe. Cutting back to 27 was a nice compromise that lets puppet run over there, lets me fool around in another window looking at consistency or getting byte counts of the files, or whatever else needs to be done.

Why are some of the en wikipedia files split so unevenly?



The older articles have many more revisions.

The 'stub' files (files containing just the page and revision metadata but not the text content) are produced all in one pass, for all three types of stubs. The stub files are then used as input for the text content file production step. All revisions of a given stub file wind up in the corresponding page content file with their text content. Since the generation of the text content and its compression is what takes the lion's share of the time, we try to split up the stubs so that generating the text content files from them takes about the same length of time for each stub. Articles entered early in Wikipedia's history (and having a lower page id, so occuring earlier in the list of stubs) have many more revisions, so we want fewer of those in a single file compared to later revisions. The files with current page content only or current pages plus meta pages will therefore vary wildly in size as well.

Why is the en wp stubs meta history 27 file so much larger than the rest?

See the above question for how we split these files up. In addition, this is the file to which all revisions created since the last dump are written, so it continues to grow over time, whereas the other files should not.

What are all those pxxxx in the en wp history dump filenames?

The history files are dumped in many pieces so that any given piece can be rerun if it fails. In order to know which pages would have to be rerun, we embed the starting and ending page ids into the filename. This means that if you want to work on the revisions belonging to a specific range of pages, you can go to the file(s) that would contain those pages instead of hunting through all of them or through the stubs files to find them.

Which page has the most revisions in the en wp dump?

Strictly speaking this is a data question and not a dumps question but I'll put the answer here anyways. (But check the database reports for more of this sort of thing.) As of August 2016, the page with the most revisions was w:Wikipedia:Administrator intervention against vandalism with 1184048 revisions, and the runner up was w:Wikipedia:Administrators' noticeboard/Incidents with 922774 revisions. In the main name space the page with the most revisions was w:George W. Bush with 46013 revisions, and not too far behind it was the w:List of WWE personnel with 44224 revisions.

Which page has the longest text in the en wp dump?

Well, like the previous question, this is not really a question about the dumps but about the data. But here you go anyways: w:Wikipedia:Upload_log/Archive_1 has text which is 3597391 bytes long. The runner up is w:Wikipedia:Upload_log/Archive_2 with 2610901 bytes. These numbers are from the Feb. 2013 dump and current results may be different. Note that MediaWiki configuration on Wikimedia projects limits revisions to 10 megabytes (this is number of bytes, not the number of characters, which for multi-byte character sets may be significantly less)

If we look at older revisions and not just current page content, we find a revision of w:Talk:Japan which has length 10597086 bytes, hitting the 10MB limit. In the main namespace there's a revision of w:The Holocaust that's 10245346 bytes long.
