

From: Trevor Bedford [REDACTED]@bedford.io
Subject: Re: data packages made available for download
Date: October 1, 2025 at 9:43 PM
To: GISAID Secretariat [REDACTED]@gisaid.org
Cc: [REDACTED]@gisaid.org



Dear GISAID Secretariat and dear Peter,

Thanks so much for getting in touch. I had completely forgotten about the outstanding email from Michael Songer from January 2023. There was a lot that I dropped in that interim. My apologies for this.

Thank you hugely for providing the backbone for ongoing genomic surveillance of SARS-CoV-2 and for keeping provisioning the flat file at www.epicov.org/epi3/3p/nextflu/export/corona2020_fulldump.json.bz2 for our continued updating of nextstrain.org/ncov/gisaid/. It's hard to believe that it's almost been six years since we started this work together in January 2020 (and still longer for work on flu).

I'd love to keep this partnership going and keep nextstrain.org/ncov/gisaid/ updating. I'd love to know if there's specific things you'd like us to be doing differently. I had noted in an email to TechDev on July 2, 2024 (copied below) about a strategy to lighten the load for provisioning data (and our processing of data), but there's many ways to do things.

But all this said, perhaps it would be best if we could have a phone call? It would be great to talk.

Separately, I can still plan to follow up on Michael Songer's queries. Should I respond directly to him or to you?

Take care,
- Trevor

On Jul 2, 2024, at 2:16 PM, Trevor Bedford [REDACTED]@bedford.io> wrote:

Dear GISAID,

Thank you so much for your continued shepherding of SARS-CoV-2 genomic data. I hugely appreciate the efforts of data curation and promoting data sharing. Thank you as well for continuing to share a flat file of the entire SARS-CoV-2 dataset to:

https://www.epicov.org/epi3/3p/nextflu/export/corona2020_fulldump.json.bz2

This has allowed us at Nextstrain to continue to keep nextstrain.org/ncov up to date over these past 4 years. I do have a request here however (along the same lines as the emails I sent below in fall 2022).

Currently, there's just the 43 Gb single file of everything. We check this file every day to see if there are updates and if there are updates we need to process the file. If instead we had something like the following:

- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2020.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2021.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2022.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2023.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2024.json.xz

each with sequences submitted in 2020, in 2021, etc... we'd only need to check the 2024 file daily. This would save us substantial download and processing overhead compared to the current 43 Gb file and I imagine it would also save you substantial overhead in terms of creating a database dump and in terms of bandwidth. This way you'd only need to update the much smaller [epicov_2024.json.xz](https://www.epicov.org/epi3/3p/nextflu/export/epicov_2024.json.xz) file every ~3 days when updates are made.

Would this sort of "sharded" database dump be something you might consider? Again, I'd hope this would improve things for both organizations. Though of course I'd be open to other suggestions as well. But generally, it would be welcome to have ways to more easily deal with the very large dataset, especially considering that it only grows every day.

Thank you again for everything you continue to do for SARS-CoV-2 genomic surveillance.

Best,
- Trevor

On Oct 1, 2025, at 2:35 AM, GISAID Secretariat [REDACTED]@gisaid.org> wrote:

Dear Mr. Bedford,

During a recent review of data packages made available for download in EpiCoV we noticed that your package 'nextregion' in the 'Genomic epidemiology' section is for the most part no longer being downloaded by users. To be precise, during the past seven months (Mar-Sep 2025) only 79 users have made a total of 88 downloads of your package 'nextregion.'

Because GISAID's resources are limited, we have made the determination to no longer make the 'nextregion' package available for download in EpiCoV.

Consequently, the data provision GISAID has generated for you on a weekly basis for the past five years to facilitate updates of the 'nextregion' package in the 'Genomic epidemiology' section, will no longer be produced.

Of course, GISAID will gladly work with you to keep up-to-date your global subsampled phylogenetic tree of the 'Genomic epidemiology of novel coronavirus (nCoV)' GISAID agreed to support in January 2020 (see your email update 26-Jan-2020).

Additionally, we draw your attention to an outstanding matter. We came across an email you sent on 2. January 2023 to our legal counsel, Mr. Michael Songer (see attached). You indicated at the time that you would follow up on his questions, once you are able following a medical leave. Our counsel, you will recall, had previously provided you with 'detailed clarification on data use under the general DAA vs. web application data agreements' you sought. We would appreciate if you provide us with your answers to our counsel's questions, and we apologize in advance, if we have misplaced your reply.

Best wishes

GISAID Secretariat

email 26-Jan-2020 Bedford
global tree Jan-2020.pdf



eMail reply Bedford 2-Jan-2023
.pdf



