



Batch Data Download and Programmatic Access

Downloading large datasets by FTP, API and other programmatic access to services

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Technological advances enabled the generation of increasingly large biological datasets. Effective access to these data often requires batch downloading or script-assisted access. As one of the largest public provider for biological data, NCBI has seen dramatic increases in data it archives and curates, all of which are available through file transfer protocol (FTP) and/or programmatic access.

A key mission of the NCBI is to provide access to archived biomedical information. The FTP site is available to download large molecular dataset. A faster alternative to FTP, powered by Aspera software, is also available.

The **FTP site** (<ftp://ftp.ncbi.nlm.nih.gov/>) contains downloadable content consisting of all of our publicly-available data. You can access the site by a web browser, a command line interface, or an FTP client. From site with blocked FTP traffic, you can use HTTPS protocol (<https://ftp.ncbi.nlm.nih.gov/>). You can also use the **Aspera-specific site** (<https://www.ncbi.nlm.nih.gov/public>) for faster download of large volume of data. The contents of this site mirror that of the FTP site. This requires the installation of the Aspera Connect browser plugin. For SRA sequence reads, we recommend using utilities from sratoolkit instead of FTP.

NCBI provides several mechanisms and data-streams for people interested in creating scripts or applications to access databases or utilize online analytical tools.

Entrez Programming Utilities (EUtils) API (<https://www.ncbi.nlm.nih.gov/books/NBK25501/>) is a set of eight server-side programs that provide a stable interface into the Entrez search and retrieval system at NCBI. These utilities use a fixed URL syntax to translates a standard set of input parameters into the values necessary for server-side software components to search for and retrieve the requested data. Three tools or packages are available to assist novice scripters in the use of the E-Utilities API:

- Ebot <https://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi>
- EDirect <https://www.ncbi.nlm.nih.gov/books/NBK179288/>
- EUtils Video Tutorial <https://www.youtube.com/watch?v=BCG-M5k-gvE>

BLAST API provides a programmatic method for you to submit BLAST search through an HTTP-based command-option interface (commonly known as **QBLAST**).

- QBLAST URL API User's Guide <https://ncbi.github.io/blast-cloud/dev/api.html>
- BLAST Cloud package <https://go.usa.gov/xmGNQ>
- BLAST Help Document <https://www.ncbi.nlm.nih.gov/books/NBK1762/>

PubChem APIs offers you several programmatic interfaces to access PubChem data and services: the E-Utilities, the XML-based Power User Gateway (PUG SOAP), and an HTTP-based RESTful version of PUG (PUG REST). They are described in the follow help document:

- Programmatic Access <https://pubchemdocs.ncbi.nlm.nih.gov/programmatic-access>

Please note that programmatic access must follow strict guidelines to ensure that NCBI resources and specialized services remain available to the general public at large.

- EUtils API Key requirement: <https://go.usa.gov/xmGNp>

Additional documents on data specifications are available online:

- NCBI Data Model <https://go.usa.gov/xmG5R>
- ASN.1 Data Specification https://www.ncbi.nlm.nih.gov/data_specs/asn/
- XML DTDs https://www.ncbi.nlm.nih.gov/data_specs/dtd/

File formats

To save space, most of the files are compressed with “.tar.gz” extension, which requires decompression utilities (such as tar, gzip, WinZip, StuffIt, or others) to expand and extract. Extracted files could be in human unreadable binary formats that will require reading or processing by specialized software, such as binary ASN.1 and pre-formatted BLAST databases. Or they could be huge text file in various formats (.xml, vcf, tabular, csv, bcp, gff, bed, etc) meant for data importation into local database or further processing by specialized tools.