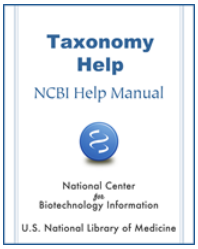
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Schoch C. NCBI Taxonomy. 2011 Apr 7 [Updated 2020 Feb 11]. In: Taxonomy Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



NCBI Taxonomy

Conrad Schoch[1]

Created: April 7, 2011; Updated: February 11, 2020.

Overview

The National Center for Biotechnology Information (NCBI) Taxonomy includes organism names and classifications for every sequence in the nucleotide and protein sequence databases of the International Nucleotide Sequence Database Collaboration (INSDC). It provides a framework for clustering elements within other domains of NCBI web pages, for internal linking between domains of the Entrez system and for linking out to taxon-specific external resources on the web and relevant publications. It is also the standard nomenclature and classification repository for the International Nucleotide Sequence Database Collaboration (INSDC) that comprises of GenBank, the European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ). It consists of a curated set of names and classifications for all of the source organisms represented in the INSDC.

The NCBI taxonomy database contains a list of names that are determined to be nomenclaturally correct or valid (as defined according to the different codes of nomenclature), classified in an approximately phylogenetic hierarchy (depending on the level of knowledge regarding phylogenetic relationships of a given group) as well as a number of names that exist outside the jurisdiction of the codes. That is, it focuses on nomenclature and systematics, rather than documenting the description of taxa. It requires a curator to select a current name out of all possible names for a specific taxon. This is done by a specialist at the NCBI Taxonomy database, relying on published taxonomic data and external expert opinion. This single, current name are indicated on all INSDC records.

A brief history

An early version of the NCBI Taxonomy in 1991 was included in the first version of Entrez, the search and retrieval system for NCBI's linked databases. This was the first system to link nucleotides and proteins from numerous disparate sources with different taxonomic classification systems. In 1996 the first version of NCBI Taxonomy Web Browser was presented to the public and the INSDC decided on its annual meeting to utilize the NCBI Taxonomy as the only source for taxonomic classification to maintain consistency among databases. This included the decision that all issues regarding nomenclature and classification would be resolved prior to the public release of any sequence data. INSDC partners now send any requests for new names (taxonomy "consults") to NCBI Taxonomy curators before final data release. Consequently, the NCBI taxonomy pages only display taxa that are linked to public sequence entries.

Author Affiliation: 1 NCBI; Email: schoch2@ncbi.nlm.nih.gov.

Codes of Nomenclature

NCBI Taxonomy distinguishes between **formal** and **informal** names. Formal names are declared based on rules laid down in four relevant codes of nomenclature (although other codes do exist). These are the International Code of Nomenclature for algae, fungi, and plants (ICNafp), the International Code of Nomenclature of Prokaryotes (ICNP) and the International Code of Zoological Nomenclature (ICZN). The viruses are governed by the International Code of Virus Classification and Nomenclature (ICVCN, also referred to as the ICTV Code). Informal names follow internal rules that are dictated by practical considerations outside of the Codes. For example, names lacking species epithets are commonly applied to GenBank records.

Treatment of Type Material in NCBI Taxonomy

A major change in database curation over the last decade was to document and track data related to the standard references for taxonomy: type strains and specimens. The word 'type' forms part of many compound terms used by taxonomists. In all the above-mentioned codes one or more types is designated as an objective standard to fix the scientific name of the species or infraspecies (i. e. subspecies, variety or forma). The type of a species or infraspecies (taxon) is designated when the new taxon is described and illustrates the trait(s) that distinguish the new taxon from all other comparable taxa. Viruses rely on lists of names approved by the International Committee on Taxonomy of Viruses (ICTV) and their usage of exemplars, focused on sequence accessions do not apply.

Under the ICNafp and ICNP, types are never living, and usually consist of dried or pickled specimens or physiologically inactivated cultures. Prokaryotic types (Bacteria and Archaea) are always living cultures (type strains). The most common types encountered, as well as certain non-nomenclatural terms, are listed below. Types may be indicated by collector name and number plus institution code where the specimen is deposited, and/or institution accession or barcode number. We list the abbreviations of the nomenclatural codes that use them in parentheses (see above).

- **Type material** (not designated in any code). This is an internal term to NCBI Taxonomy, and if not further designated, refers to specimens or cultures where the type of type is unknown or unspecified.
- **Holotype** (ICNafp, ICNP). is the most important. There is only one holotype, usually a single specimen, and it is the 'name-bearer' of the described taxon. It serves as the standard to which all subsequent examples of the described taxon are compared. The equivalent term we use for prokaryotes is type strain (ICNP) and there can be multiple co-identical type strains, cultured from a single source.
- **Neotype** (ICNafp, ICNP, ICZN). If the holotype is lost or destroyed, a neotype specimen is designated from a collection considered to be representative of the original holotype. There is only one neotype. The equivalent term we use for prokaryotes is neotype strain (ICNP) with multiple neotype strains possible when cultured from a single source.
- **Isotype** (ICNafp). Under the ICNafp one or more duplicate specimens from the holotype collection can be deposited in other institutions. Usually the collection number is the same as the holotype, but the institution code has to be different. Iso- can be appended to other kinds of types to indicate duplicates, e.g., isosyntype, etc. Isotype is not a formally accepted term in the ICZN.

The next set of types are considered of lower nomenclatural importance, because they are derived from specimens that act to inform and expand the concept of the preceding set of types and will generally be different genetically, although still potentially closely related.

- **Paratype** (ICNafp, ICZN). One or more additional specimens chosen to further illustrate traits in the described taxon.
- **Epitype** (ICNafp): In botanical nomenclature only, a type designated to expand on the original holotype concept. There should be only one epitype.

- **Culture fromtype** (ICNafp): Also, sometimes designated ex-type, e.g., ex-holotype, etc. There can be multiple of these. The types of cultivable, microbial eukaryotes must be inactivated and one or preferably more living cultures is extracted from the type and maintained in culture collections. Since these are living, they do not have formal status as type, but since they will often be the source of DNA sequence data they are annotated and indicated as type material.
- **Reference material** or **reference strain** (not designated in any code): The reference material and reference strain qualifiers are internal GenBank terms used to capture any reference strain or material exclusively of types. These terms can include specimens or cultures that don't have nomenclatural standing but nevertheless could have name-bearing value. For example, Candidatus prokaryotic names, are names proposed for new species that have not been formally described by the ICNP will receive reference strain designations, since technically they do not possess types.

Additional types are listed under the controlled vocabulary for the INSDC: http://www.insdc.org/controlled-vocabulary-typematerial-qualifer

Data Model

Each entry in the INSDC databases maps onto an entry in the NCBI Taxonomy database at the level of species or below (an exception is made for patent entries). Since October 2018 the NCBI Taxonomy has been migrated from an SQL server relational database to a system that incorporates a system of databases, focused around a central resource called **NameBank**. This provides a framework for data including additional information that is not public in NCBI Taxonomy and contextualizes it.

Each NCBI Taxonomy entry or **TaxNode** includes: a **primary Name** with a number of **secondary names**. Each entry or node has a public stable, unique identifier, the taxonomy identifier (**TaxId**) as well as the presence of separate, stable unique name entity identifiers (**NameIds)** which are not public and managed in **NameBank**.

Each **TaxNode** has the following:

1 **TaxId**

This is shared by all names for a specific **TaxNode**. Each **TaxNode** has a stable, unique numerical identifier, the taxonomy identifier (**TaxId**). Each **TaxId** has a labelled **primary name** (a **formal** or **informal name**) which shows up on the NCBI records.

In publications this can be standardized as a **primary name** with its **TaxId** displayed as "NCBI:txid" followed by a number, e.g.

Homo sapiens NCBI:txid9606

This information is also displayed in the **NCBI TaxBrowser**.

2. **NameBank Entity Id**

NameBank includes several separate, stable and unique name entity identifiers for **secondary names** and their properties, managed in **NameBank**. Formal **secondary names** under the NCBI system will consist of a Latin binomial (consisting of the genus name and species epithet) as well as its authority (the person(s) who described the species) and a year of its description. In the NCBI Taxonomy this is the **current name** for the TaxNode labelled as *Homo sapiens:*

Homo sapiens Linneaus, 1758 (with a NameBank Entity Id, which is not displayed publicly)

Relational terms make it possible to indicate relationship between for **secondary names** and **primary names**. For instance, the first name attached to a species description is indicated as the **basionym**. The currently

accepted name is indicated as **current name** (which could also be the first name). These terms also differentiate between different synonyms: **heterotypic synonym** and **homotypic synonym**. Homotypic (or objective synonyms), are names based on the same type (see discussion of type material). Heterotypic (or subjective) synonyms are based on different types that were considered distinct taxa when they were first proposed, but subsequently considered to belong to the same taxon, documented in a publication or other authoritative source. The **relational terms**, specifying relationships between the different names are indicated in Table 1.

Table 1. Relational terms in NCBI Taxonomy.

Name category	Description	Number per TaxNode/TaxId
primary name	This is the designated label for the TaxNode	One per TaxNode
Formal relational terms (relying on definitions in the codes of nomenclature)		
current name	The current name chosen out of all synonyms for the TaxNode. Will often overlap with primary name (except in a few cases) or, more rarely, basionym.	Up to one per TaxNode (where indicated)
basionym	The originally described name, attached to the type material	Up to one per name (where indicated), one to more per TaxNode
homotypic synonym	Names generated after the basionym (e.g. by moving it to a different genus), but sharing the same type	None to several per TaxNode
heterotypic synonym	Names with a different basionym and type from those mentioned above	None to several per TaxNode
Informal relational terms, public		
acronym	Mainly used for viruses	None to several per TaxNode
equivalent	Used for informal names which a related but not synonyms	None to several per TaxNode
includes	Used for informal names which forms subset of a name	None to several per TaxNode
in-part	Used for formal names which forms subset of a name	None to several per TaxNode – can be duplicate across TaxNodes
blast name	Informal name for groups of organisms	Up to one per TaxNode (where indicated)
common name	Informal names in common usage – these are not comprehensively added	None to several per TaxNode
genbank acronym	Name type that ensures an acronym name type is displayed prominently in flat files	Up to one per TaxNode but excluding other genbank name types (where indicated)
genbank synonym	Name type that ensures a second synonym is displayed prominently in flat files	Up to one per TaxNode but excluding other genbank name types (where indicated)
genbank common name	Name type that ensures a vernacular name is displayed prominently in flat files	Up to one per TaxNode but excluding other genbank name types (where indicated)
Informal relational terms, not public		
misspelling	Used for searches only	None to several per TaxNode
unpublished name	Used for searches only	None to several per TaxNode

Importantly, it should be noted that taxonomic information in the NCBI Taxonomy is not complete and several entries will only contain a primary name and current name. This reflect changes where minimal information was available when an entry was first made. Where additional information becomes available it is added to expand existing entries.

Access

The NCBI Taxonomy can be viewed and interacted with in three main ways.

1. Under the **NCBI Taxonomy Browser** two different kinds of web pages are supported. Hierarchy pages present the taxonomic classification, while taxon-specific pages summarize all of the information associated with a particular taxonomic entry. By default, the hierarchy displays three levels in the classification, but this can be changed (zero levels display the taxon-specific page). The hierarchy pages are also customized to display hotlinked counts of entries in other Entrez databases. Taxon specific pages will display the names and their relational terms associated with that entry (except for misspellings and unpublished names). The lineage can display a full or abbreviated classification. In addition to this, manually curated information is displayed as well. This includes type material and comments annotated by the taxonomic curators as well as relevant literature and type material information, with hotlinks as appropriate. An Entrez records table shows links to other Entrez databases, including links to records from type material, when they are available and annotated. At the bottom of the taxon specific pages additional layered information sets are found. For relevant organisms this can include **Genome information** with links to resource information in other databases. Another set of links are **External Information Resources** or **NCBI LinkOut**. This provides outside groups the ability to maintain links to externally curated resources after submitting information for their web-accessible resources (https://www.ncbi.nlm.nih.gov/projects/linkout/). Another block below this will display the modifiers such as **strain**, **isolate** or **culture collection** associated with the organism in GenBank sequence entries. Finally, additional search capabilities are supported in the **Taxonomy Browser** such as a wild card search.
2. **Taxonomy Entrez** supports Boolean queries that includes search fields common across all Entrez databases. This is a uniform indexing, search, and retrieval engine for names. It differs from the other Entrez databases in that the default search field is [all names] instead of [all terms] – unless you specify an explicit search field, **Taxonomy Entrez** assumes that you are searching with a name (or a **TaxId**). Each entry in the **Taxonomy Entrez** results represents a taxon and is identified by its **primary name**. The **primary name** for a taxon may be either formal (e.g. *Homo sapiens*) or informal (e.g. *Homo* sp. Altai, or uncultured fungus). Taxa above the species level may also be either formal (e.g. the order Mammalia) or informal (e.g. the clade node eudicotyledons). In addition, each taxon may be associated with **secondary names** with several relational terms – homotypic synonyms, basionyms, misspellings, common names and so on.

Boolean queries can be used in Entrez to answer many questions. For example, how many amphibian taxa have appeared in Entrez for the first time this year? As with any Entrez query, it is often useful to set up "what's new" email updates in MyNCBI for queries of interest. Note: the 'specified' property flags formal names at and below the species level and the 'has type material' filter will return all names with type material annotated, whether it links to records with this annotation or not.

Amphibia [subtree] AND specified [prop]

Amphibia [subtree] AND 2019 [date]

Amphibia [subtree] AND 2019/12/15:2020/01/15 [date]

Amphibia [subtree] AND 2020 [date] AND species [rank]

Amphibia [subtree] AND 2020 [date] AND species [rank] AND specified [prop]

Amphibia [subtree] AND 2020 [date] AND species [rank] NOT specified [prop]

Amphibia [subtree] AND has type material[filter]

3. A third option is the full text files of the complete database that gets updated and deposited every hour at the **Taxonomy FTP** site as taxdump files. There is now also a second version of the FTP taxdump files. The new version includes taxonomic lineages of taxa and information on type material. The most important files in both sets of FTP files are nodes.dmp (which maps **TaxIds** to their parent **TaxIds**) and names.dmp (which maps names to **TaxIds**). Other files in both sets are delnodes.dmp which lists **TaxNodes** that have been deleted from the database, as well as **TaxNodes** that were once public but are no longer linked to any public sequence entries. Also, merged.dmp maps secondary **TaxIds** onto primary **TaxIds** for taxa that have been synonymized in the database. There is now also a second version of the FTP taxdump files. The new version includes files with differently formatted taxonomic lineages, information on type material and host information (**typematerial.dmp**, **typeoftype.dmp**, **rankedlineage.dmp**, **fullnamelineage.dmp**, **taxidlineage.dmp**, and **host.dmp**). Details are in the readme files.

Links to all three options are shown below:

1. http://www.ncbi.nlm.nih.gov/taxonomy
2. http://www.ncbi.nlm.nih.gov/Taxonomy/Browser.wwwtax.cgi
3. ftp://ftp.ncbi.nih.gov/pub/taxonomy (previous unchanged version, still supported)

ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/ (new version, recommended for usage, with extra options)

Other tools for NCBI Taxonomy information

- The **name/id status report** can be found under Taxonomy Tools on the NCBI Taxonomy home page (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi). This allows the user to upload a list of names (or **TaxIds**) to get back a report on their status and if they are present, which is preferred scientific name is being used at NCBI. A command-line version of this function (taxident) is available in the NCBI C++ toolkit.
- **SourceCheck** reads in a set of Genbank accessions and returns associated metadata anid is now available as a standalone tool: ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/cmdline/
- The **Taxonomy Common Tree**: https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi
- A live summary of the numbers of **Taxonomy Statistics**, all public entries in the NCBI Taxonomy, broken down by user-defined taxonomic group, date, or rank is available: https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=statistics
- The **NCBI Tree Viewer** (TV): https://www.ncbi.nlm.nih.gov/tools/treeviewer/. More capabilities are available in the **NCBI Genome Workbench**: https://www.ncbi.nlm.nih.gov/tools/gbench/

A complete set of NCBI tools can be found here: https://www.ncbi.nlm.nih.gov/guide/all/

Further reading

Federhen, S. (2012) The NCBI Taxonomy database. Nucleic Acids Research, 40, D136-D143.

Federhen, S. (2015) Type material in the NCBI Taxonomy Database. Nucleic Acids Research, 43, D1086-D1098.

Ciufo, S., Kannan, S., Sharma, S., et al. (2018) Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol, 68, 2386-2392.

Sharma, S., Ciufo, S., Starchenko, E., et al. (2018) The NCBI BioCollections Database. Database : the journal of biological databases and curation, 2018, bay006.