ITIS Darwin Core Archive Layout and Data Application

how DwC-A is prepared from ITIS data

David Mitchell

3/6/2019

This document targeted for ITIS data users who want to understand how Darwin Core Archive (DwCA) files are derived from ITIS.

ITIS DwCA Record Layout

Field Name	Description	Required	Data Type	Specifications
taxonID	A taxonomic serial number, TSN, assigned for occurrences of Taxonomic Units at each level of the hierarchy through genus and for all polynomial infrageneric Taxonomic Units	Yes	INT	Unique identifier for valid/accepted and invalid/not accepted scientific names
acceptedNameUsageID	The TSN of the valid/accepted name used to link a synonym (invalid/ not accepted name) to corresponding name	Yes when taxonomicStatus is something other than accepted or valid, otherwise null	INT	Unique identifier for valid/accepted scientific names
parentNameUsageID	The TSN of the valid/accepted name that is the direct parent of the subject occurrence in taxonID	Yes when taxonomicStatus is accepted or valid, otherwise null	INT	Unique identifier for valid/accepted direct parent
scientificName	Full scientific name derived concatenated name elements including taxon authorship if applicable	Yes	Text	Max 218 UTF characters
scientificNameAuthorship	Taxon authorship, formatted per the governing nomenclatural code	No	Text	Max 100 UTF characters
kingdom	Kingdom in which the valid/accepted name is classified	Yes	Text	Max 35 UTF characters
phylum	Phylum or Division in which the valid/accepted name is classified	Yes for names under the rank of Phylum/Division, otherwise null	Text	Max 35 UTF characters
class	Class in which the valid/accepted name is classified	Yes for names under the rank of Class, otherwise null	Text	Max 35 UTF characters
order	Order in which the valid/accepted name is classified	Yes for names under the rank of Order, otherwise null	Text	Max 35 UTF characters
superfamily	Superfamily in which the valid/accepted name is classified	Yes for names under the rank of	Text	Max 35 UTF characters

		Superfamily, otherwise null		
family	Family in which the valid/accepted name is classified	Yes for names under the rank of Family, otherwise null	Text	Max 35 UTF characters
genus	Genus in which the valid/accepted infrageneric name is classified, or the first term of the scientific name for species and below	Yes for names under the rank of Genus, otherwise null	Text	Max 35 UTF characters
subgenus	Subgenus in which the valid/accepted name is classified	Yes for names under the rank of Subgenus, otherwise null	Text	Max 73 UTF characters
specificEpithet	The second term of the scientific name for species and below	Yes for species and subspecific taxa, otherwise null	Text	Max 35 UTF characters
infraspecificEpithet	The third term of the scientific name for subspecies and below	Yes for names under the rank of Species, otherwise null	Text	Max 35 UTF characters
taxonRank	The rank name for a level in the taxonomic hierarchy	Yes	Text	Max 15 UTF characters
taxonomicStatus	The status of the use of the scientificName	Yes	Text	Max 19 UTF characters Values: accepted valid synonym homotypic synonym proParteSynonym misapplied
modified	Derived effective date of the record	Yes	Date	yyyy-mm-dd
namePublishedIn	The publication containing the original description	No	Memo	Derived from another nomenclatural entity.
scientificNameID	A taxonomic serial number, TSN, assigned for occurrences of Taxonomic Units at each level of the hierarchy through genus and for all polynomial infrageneric Taxonomic Units	Yes	INT	Unique identifier for valid/accepted and invalid/not accepted scientific names
verbatimTaxonRank	Subspecific rank indicator	No	Text	Max 7 UTF characters

				Values: ssp. var. subvar. f. subf.
taxonRemarks	Notes about the taxon or name	No	Memo	Commentator and comments
higherClassification	Pipe delimited hierarchy from the trunk down to branch immediately superior to scientificName	No	Text	e.g. Bacteria | Negibacteria | Proteobacteria | Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Cronobacter | Cronobacter dublinensis

1.0 Overview

To facilitate the exchange of taxonomic data within the biodiversity informatics community the Global Biodiversity Information Facility created Darwin Core Archvie (DwCA), a data standard with features that permit exchange of taxonomic data. ITIS data can be downloaded in DwCA format, and this document defines how ITIS data are used to create the dataset. ITIS database elements and their data definition can be found in the document ITIS Data Model (https://itis.gov/pdf/ITIS_ConceptualModelEntityDefinition.pdf), and DwCA terms and their definition can be found in the Darwin Core Terms Quick Reference Guide (http://rs.tdwg.org/dwc/terms).

The ITIS DwCA file is a self contained zipped (.zip) archive consisting of three files. The first file, taxa.txt, includes a header row the fields mapped to the ITIS standards as described herein. The second file, meta.xml, describes the core taxa.txt file in terms of the DwCA format. The third file, eml.xml, describes the common attributes of the records in taxa.txt

1.1 taxa.txt

The core taxa are included in the taxa.txt file. The file is a tab delimited text file, UTF-8 encoded, that always begins with the following tab delimited (tabs indicated by ->) header.

taxonID->acceptedNameUsageID->parentNameUsageID->scientificName->scientificNameAuthorship->kingdom->phylum->class->order->superfamily->family->genus->subgenus->specificEpithet->infraspecificEpithet->taxonRank->taxonomicStatus->modified->namePublishedIn->scientificNameID->verbatimTaxonRank->taxonRemarks->higherClassification

1.2 meta.xml

The description of the core taxa file, in terms of DwCA format, is provided in the meta.xml file. Unless ITIS updates the core taxa file by adding or removing fields, the meta.xml file remains the same regardless of the data content changes or additions to core taxa.

1.3 eml.xml

Another metadata file using the Ecological Metadata Language (eml) is provided with information about the source of the core taxa. The metadata included in eml.xml includes the auto-generated file name, the date of file creation, and the version of the ITIS software that created the download, and information about the source of the download. This information is extracted into a separate file so identical information won't have to be repeated for every core taxon record.

2.0 Data Definitions and Policies

2.1 taxonID

The Darwin Core term dwc:taxonID is the identifier for the set of taxon information. The ITIS DwCA implementation uses the Taxonomic Serial Number (TSN) from the ITIS field **`tsn`** (table taxonomic_units) to populate taxonID. An ITIS TSN is a unique, persistent, non-intelligent number serially assigned by ITIS to all scientific names as they are loaded into the ITIS database. In terms of ITIS, a TSN is assigned to all taxonomic units regardless of usage.

The field is non-nullable. It may repeat in a download when more than one record is needed to represent the one-to-many relationship between an invalid/not accepted name and the corresponding valid/accepted names.

2.2 acceptedNameUsageID

The Darwin Core term dwc:acceptedNameUsageID is the identifier for the currently valid or accepted taxon. The ITIS DwCA implementation uses the Taxonomic Serial Number (TSN) from the field **`tsn_accepted`** (table synonym_links) to populate acceptedNameUsageID.

The field is null when dwc:taxonomicStatus is ‘accepted’ or ‘valid’, and is non-null when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS an invalid/not accepted synonym can have more than one valid/accepted name. When this occurs the result will be multiple records differing only in values found in acceptedNameUsageID and possibly in the dwc:modified field and the hierarchy attribute fields.

2.3 parentNameUsageID

The Darwin Core term dwc:parentNameUsageID is the identifier for the direct parent of the taxon represented in dwc:taxonID. The ITIS DwCA implementation uses the Taxonomic Serial Number (TSN) from the field **`parent_tsn`** (table taxonomic_units) to populate parentNameUsageID.

The field is null when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’ and non-null when dwc:taxonomicStatus ‘accepted’ or ‘valid’.

In ITIS an accepted or valid name can only have one parent name, therefore there is ever only one parentNameUsageID for a valid or accepted name.

2.4 scientificName

The Darwin Core term dwc:scientificName is the full scientific name with taxon authorship if applicable. In ITIS scientific names are composed of 9 name elements (**`unit_ind1`**, **`unit_name1`**, **`unit_ind2`**, **`unit_name2`**, **`unit_ind3`**, **`unit_name3`**, **`unit_ind4`**, **`unit_name4`**, and **`taxon_author`**). Because DwCA only supports 4 name elements (dwc:genus, dwc:specificEpithet, dwc:infraspecificEpithet, dwc:scientifcNameAuthorship) some ITIS names cannot be represented in DwCA and are excluded. These include hybrid formulae (e.g. Abies concolor X Abies grandis, TSN 822690) and quadrinomials (e.g. Grimmia alpicola var. rivularis f. acutifolia Grout, TSN 550185 & Apocynum androsaemifolium ssp. androsaemifolium var. griseum (Greene) Bég. & Beloserky, TSN 184773).

When ITIS contains a hybrid indicator (X) in **`unit_ind1`** or **`unit_ind2`** the scientific name element it precedes is modified by placing a × (multiplication sign) in front of the name without a space. For example the scientificName for TSN 25311 is ×Sorbaronia C.K. Schneid., and for TSN 825231 Citrus ×aurantium aurantium L.

2.5 scientificNameAuthorship

The Darwin Core term dwc:scientificNameAuthorship is authorship information associated with the scientific name, formatted and applicable to scientific names per the governing nomenclatural code. The ITIS DwCA implementation uses the string value from **`taxon_author`** (table taxon_authors_lkp) to populate scientificNameAuthorship. The scientificNameAuthorship is nullable because not all names, especially names at the rank of Family and above, have associated authorship information.

2.6 Hierarchy and Name

2.6.1 kingdom

The Darwin Core term dwc:kingdom is the full scientific name, without taxon authorship, of the kingdom in which the name represented by dwc:taxonID is classified under when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **`rank_id`** 10 is Kingdom and represents the trunk of the hierarchy. The ITIS DwCA implementation uses the kingdom name under which the name is classified, and derives the kingdom name from **`unit_name1`**. Current values are Bacteria, Protozoa, Plantae, Fungi, Animalia, Chromista, and Protozoa.

2.6.2 phylum

The Darwin Core term dwc:phylum is the full scientific name, without taxon authorship, of the phylum (Bacteria, Protozoa, Animalia, Archaea) or division (Plantae, Fungi, Chromista) in which the name represented by dwc:taxonID is classified when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **`rank_id`** 30 is Phylum for names governed by International Commission on Zoological Nomenclature (ICZN) and International Code of Nomenclature of Bacteria (ICNB), and Division for names governed by International Code of Nomenclature for Algae, Fungi, and Plants (ICN). The ITIS DwCA implementation uses the phylum or division name under which the name is classified, and derives the phylum name from **`unit_name1`**.

2.6.3 class

The Darwin Core term dwc:class is the full scientific name, without taxon authorship, of the class in which the name represented by dwc:taxonID is classified when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **rank_id** 60 is Class and is available for all kingdoms. The ITIS DwCA implementation uses the class name under which the name is classified, and derives the class name from **unit_name1**.

2.6.4 order

The Darwin Core term dwc:order is the full scientific name, without taxon authorship, of the order in which the name represented by dwc:taxonID is classified when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **rank_id** 100 is Order and is available for all kingdoms. The ITIS DwCA implementation uses the order name under which the name is classified, and derives the order name from **unit_name1**.

2.6.5 superfamily

The Darwin Core standard does not support the term superfamily but it is included here because other importers, most notably Catalogue of Life partners, do. It may be adopted in the future, but until then superfamily should be ignored by importers who cannot accept non-DwCA terms. The term superfamily is the full scientific name, without taxon authorship, of the superfamily in which the name represented by dwc:taxonID is classified when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **rank_id** 130 is Superfamily and is available for kingdoms Bacteria, Protozoa, Animalia, and Archaea. The ITIS DwCA implementation uses the superfamily name under which the name is classified, and derives the superfamily name from **unit_name1**.

2.6.6 family

The Darwin Core term dwc:family is the full scientific name, without taxon authorship, of the family in which the name represented by dwc:taxonID is classified when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **rank_id** 140 is Family and is available for all kingdoms. The ITIS DwCA implementation uses the family name under which the name is classified, and derives the family name from **unit_name1**.

2.6.7 genus

The term dwc:genus is derived two different ways, depending upon the rank of the scientific name. For infrageneric names (names below genus and above species) the Darwin Core term dwc:genus is the full scientific name, without taxon authorship, of the genus which the name represented by dwc:taxonID is classified in when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’. For all specific and subspecific taxa, regardless of usage, dwc:genus is the first term of the scientific name. This means for valid and accepted names above species dwc:genus is a hierarchical attribute, and for valid and accepted names of species and below dwc:genus is the atomized first term in the name.

In ITIS the distinction is important. For the majority of species and subspecific taxa binomial nomenclature assures the first term of the name is the genus in which the taxon is classified. However, there are cases in ITIS where valid and accepted name combinations are classified under an unmatched genus. This rare occurrence happens when taxa have been placed within a genus but new species combinations have not been published.

Hybrid genera in ITIS are represented in dwc:genus with a multiplication sign. For example ×Sorbaronia.

2.6.8 subgenus

The Darwin Core term dwc:subgenus is the full scientific name, without taxon authorship, of the subgenus in which the name represented by dwc:taxonID is classified when dwc:taxonomicStatus is 'accepted' or 'valid', or the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is ‘synonym’, ‘homotypic synonym’, ‘heterotypic synonym’, ‘proParteSynonym’, or ‘misapplied’.

In ITIS **rank_id** 190 is Subgenus and is available for all kingdoms. The ITIS DwCA implementation uses the subgenus name under which the name is classified, and derives the subgenus name from **unit_ind1**, **unit_name1**, **unit_ind2**, and **unit_name2**.

2.6.9 specificEpithet

The Darwin Core term dwc:specificEpithet is the second term of the scientific name for names at the ranks of species and below.

The ITIS DwCA implementation derives the specificEpithet from **`unit_ind2`** and **`unit_name2`**.

2.6.10 infraspecificEpithet

The Darwin Core term dwc:infraspecificEpithet is the third term of the scientific name for names at the ranks of subspecies and below.

The ITIS DwCA implementation derives the infraspecificEpithet from **`unit_name3`**.

ITIS maintains infraspecific rank indicators in **`unit_ind3`** and does not concatenate them with third term of scientific names to derive dwc:infraspecificEpithet.

2.16 taxonRank

The Darwin Core term dwc:taxonRank is the taxonomic rank of the name denoted in dwc:scientificName. The ITIS DwCA implementation derives the rank name per the **`rank_name`** (ITIS table taxon_unit_types) associated with the unique identifier for the kingdom specific level in the taxonomic hierarchy.

ITIS includes more ranks than currently recognized by DwCA. Noted below are the ranks ITIS uses to generate dwc:taxonRank that are not a part of the recommended controlled vocabulary.

Table 1 ITIS Ranks outside of the DwCA controlled vocabulary for taxonRank

Rank	Kingdoms
Infrakingdom	Animalia, Protozoa, Chromista, Plantae

Superphylum	Animalia
Superdivision	Chromista, Plantae
Infraphylum	Animalia, Protozoa
Infradivision	Chromista, Plantae
Parvphylum	Protozoa
Parvdivision	Chromista
Superclass	Animalia, Protozoa, Chromista, Plantae, Archaea, Bacteria
Infraclass	Animalia, Protozoa, Chromista, Plantae, Archaea, Bacteria
Superorder	Animalia, Protozoa, Chromista, Fungi, Plantae, Archaea, Bacteria
Infraorder	Animalia, Protozoa, Archaea, Bacteria
Superfamily	Animalia, Protozoa, Archaea, Bacteria
Race	Animalia
Stirp	Animalia
Morph	Animalia
Aberration	Animalia
Unspecified	Animalia

2.17 taxonomicStatus

The Darwin Core term dwc:taxonomicStatus is the status of the name denoted in dwc:scientificName. The ITIS DwCA implementation derives taxonomicStatus per the values in the fields **`usage`** and **`unaccept_reason`** (ITIS table taxonomic_units). Names that are not synonyms have a taxonomicStatus of valid in kingdoms Animalia, Archaea, Bacteria, Protozoa and accepted in kingdoms Chromista, Fungi, Plantae.

The ITIS DwCA implementation of dwc:taxonomicStatus for names with usage of invalid or not accepted are noted below.

Table 2 Derived dwc:taxonomicStatus from ITIS unacceptability reason

unacceptability_reason	**dwc:taxonomicStatus**
Chromista, Fungi, & Plantae	
homonym (illegitimate)	heterotypicSynonym
horticultural	synonym
invalidly published, nomen nudum	synonym
invalidly published, other	synonym
misapplied	misapplied
orthographic variant (misspelling)	homotypicSynonym
other, see comments	synonym
pro parte	proParteSynonym

rejected name	synonym
superfluous renaming (illegitimate)	homotypicSynonym
synonym	synonym
unspecified in provided data	synonym
Animalia, Archaea, Bacteria and Protozoa	
homonym & junior synonym	synonym
junior homonym	heterotypicSynonym
junior synonym	heterotypicSynonym
misapplied	misapplied
nomen dubium	synonym
nomen oblitum	heterotypicSynonym
original name/combination	homotypicSynonym
other, see comments	synonym
pro parte	proParteSynonym
subsequent name/combination	homotypicSynonym

unavailable, incorrect orig. spelling	homotypicSynonym
unavailable, literature misspelling	homotypicSynonym
unavailable, nomen nudum	synonym
unavailable, other	synonym
unavailable, suppressed by ruling	synonym
unjustified emendation	homotypicSynonym
unnecessary replacement	homotypicSynonym
unspecified in provided data	synonym

2.18 modified

Darwin Core Archive doesn't include a term for datestamping a name. The ITIS implementation of DwCA is extended with the Dublin Core Metadata Initiative term dcmi:modified, which is the effective date of each name in the file. The date for each name is derived by comparing the update_date value from the ITIS taxonomic_units table with the greatest update_date value from the tables associated with taxonomic_units record, but with the exclusion of dates from the ITIS table vern_ref_links and their linked vernacular references. Whichever value is greater becomes the date for dwc:modified.

Hierarchy updates in ITIS that will impact the hierarchy attributes of dwc:scientificName will not be reflected in dcmi:modified. This way ITIS can make upper hierarchy changes to a branch without having the modification date cascade down to all the leaves.

2.19 namePublishedIn

The Darwin Core term dwc:namePublishedIn is the reference details of the publication containing the original description of dwc:scientificName. The ITIS DwCA implementation derives namePublishedIn by concatenating fields from the ITIS table publications when the link between the name and the publication has a positive value in the **`original_desc_ind`**. The publication fields concatenated to form reference are **`reference_author`**, **`actual_pub_date`**, **`title`**, **`publication_name`**, **`pages`**, **`publisher`**, **`pub_place`**, **`ISBN`**, **`ISSN`**, and **`pub_comment`**.

Every ITIS name is associated with one or more references, but not all names have an original publication associated with them. When an original publication is absent dwc:namePublishedIn is null.

2.20 scientificNameID

The Darwin Core term dwc:scientificNameID is the identifier for the nomenclatural details of the scientific name. Because ITIS does not have a separate indicator for a taxonomic assertion, the ITIS DwCA implementation uses the Taxonomic Serial Number (TSN) just as in dwc:taxonID.

Duplicating the ITIS TSN in taxonID and scientificNameID ensures the output's usefulness for subscribers with varying import requirements.

2.21 verbatimTaxonRank

The Darwin Core term dwc:verbatimTaxonRank is the rank of the dwc:scientificName as it appears in the original record. In order to provide full atomization of the ITIS elements in dwc:scientificName the ITIS DwCA implementation uses verbatimTaxonRank to provide the subspecific markers (e.g. ssp., f.) from **`unit_ind3`** (ITIS table taxonomic_units).

2.22 taxonRemarks

The Darwin Core term dwc:taxonRemarks contains notes about the taxon or name. The ITIS DwCA implementation derives taxonRemarks by concatenating the two ITIS fields **`commentator`** and **`comment_detail`** (ITIS table comments). A single name

in ITIS may contain multiple comments. For names with more than one comment the commentator + comment_detail are combined into dwc:taxonRemarks.

2.23 higherClassification

The Darwin Core term dwc:higherClassification is a pipe delimited hierarchy from the trunk down to the branch immediately superior to the dwc:scientificName when dwc:taxonomicStatus is 'accepted' or 'valid', or the hierarchy for the name represented by dwc:acceptedNameUsageID when dwc:taxonomicStatus is 'synonym', 'homotypic synonym', 'heterotypic synonym', 'proParteSynonym', or 'misapplied'.

The ITIS DwCA implementation uses the verbatim hierarchy from kingdom to parent name of dwc:scientificName.

The values in higherClassification will be consistent with the higher taxon records that relate to each other by dwc:parentNameUsageID.