# Taxon profile specification

This document aims to open the discussion about different choices that we have to make with respect to the way we should describe a taxon in Bioschemas.org.

Feel free to comment, discuss, propose alternate solutions, raise other issues etc.

See the current [mapping file](https://docs.google.com/spreadsheets/d/1hHE7SHz9qt6Eg6j6DQgI9BwhrlGOwkO68fzek78m62I/edit#gid=1483018794).

## 

## Specification scope

As a first approach, this specification proposes to describe the most simple elements of a taxon: its scientific names, vernacular names, taxonomic rank and parent taxon.

Other information, such as conservation statuses, traits, occurrences, etc., may be specified in the future if needed, or they may simply be left up to the data provider to decide on the way to describe or not such information.

## Third party vocabularies

For equivalent terms available from several well known ontologies (DarwinCore, TDWG, NCBITaxon, Wikidata etc.), which term should we pick up?

Discussions within the Bioschema.org community converge towards the idea that each specification group should propose a single approved term for each type and each property, that represents the group consensus. But having multiple options would open the door to greater inconsistency.

Current specification: we use terms from DarwinCore, and the TaxonConcept TDWG ontology which are well known to the **biodiversity community.**

#### Wikidata

Wikidata provides a nice set of terms to describe taxa, and gets increasing adoption and visibility from the **Web community**. But Wikidata remains a versatile source (the counterpart of being collaborative) and scientists may deem it not sufficiently validated.

*Leyla*> Wikidata is/will be adding Bioschemas markup so even if we go for a different vocabulary, there would be a way to interact with WikiData.

*Franck*> We could still use relatively well established Wikidata properties (has taxon rank, has parent taxon) and/or classes (taxonomic ranks: species, genus, family etc.) when relevant.

#### NBCI Organismal Classification

*Leyla*> NCBITaxon. Maybe worth to see how the mentioned ones map to each other.

*Franck*> We can use some properties (e.g. has\_rank) and classes (e.g. taxonomic ranks). But are there any classes/properties from NCBI that would not be provided by any other ontology?

## Taxon names

Taxon names is a hard topic by itself: names history record, names status (accepted, synonyms…), nomenclatural codes, etc. The point of Bioschemas.org is not to get into too much detail, but to set up a simple way of identifying taxa through webpages.

Schema.org's properties *name* and *alternameName* can be used to denote the taxon names (currently accepted name and synonyms), however a question is whether these should contain a short name (latin or Linnean name), or a full name including the authority and date.

For now, the specification recommends schema:name for the short name, and *dwc:scientificName* for the full name. By the way, which DwC property best conveys the idea of the taxon' s current full name: *dwc:scientificName? dwc:acceptedNameUsage*?

So, should we try to link to such DwC properties due to the wide adoption of this vocabulary? Or should we stick to what's in schema.org?

*Carl Boettiger>* I strongly recommend we make use of taxonomic identifiers to associate a particular name with a given authority (NCBI, ITIS) rather than relying on names alone. Identifying something as `dwc:scientificName` is almost useless without connecting it to a particular authority, as different authorities have different names which they consider the "accepted" name vs the "synonym". (...) We can use Wikidata properties to name third party identifiers: NCBI id, EOL id, TAXREF id etc.

## Taxonomic rank

There is no such thing as a taxonomic rank in schema.org. DarwinCore property *dwc:taxonRank* only allows literal values, which is an old heritage of the DarwinCore CSV format. It should be possible to use other properties whose object are URIs such as *tc:rank* from the TDWG ontologies, so as to use a controlled vocabulary to name ranks. The problem is that many datasets have defined their own URIs for taxonomic ranks. So

* should we allow all and any of them?
* should we pick up one of them?

For now, the specification allows both properties *dwc:taxonRank* and *tc:rank* but does not restrict the range of *tc:rank*. What about using the Wikidata taxon rank property <https://www.wikidata.org/entity/P105>, and associated ranks like species <https://www.wikidata.org/entity/Q7432>? Or use tc:rank with Wikidata ranks?

## Parent taxon

BioChemEntity's property *isContainedIn* is a good candidate to denote the relationship between a taxon and its parent in the classification, i.e. between a species and its order, an order and its family etc. Another option would be to use DarwinCore property *dwc:parentNameUsageID*.

This however questions how we name each Taxon entity:

* should each taxon entity define its URI ("@id" in JSON-LD)? Such a URI could in turn be used by its descendants to refer to their parent.
* Should we use webpage URLs, such as those in property *url*?
* We may also use URIs from other classifications being represented as linked data.

The current specification mimics what has been done in the Gene case:

"isContainedIn": {  
 "@type": [ "BioChemEntity", "Taxon" ],  
 "identifier": "191588",  
 "url": "https://inpn.mnhn.fr/espece/cd\_nom/191588?lg=en",  
 "name": "Delphinapterus",  
 "dwc:scientificName": "Delphinapterus Lacépède, 1804",  
},