

Controlled data vocabularies

Standardized terms for effective information management

Matthew Richard, HELCOM Data Manager

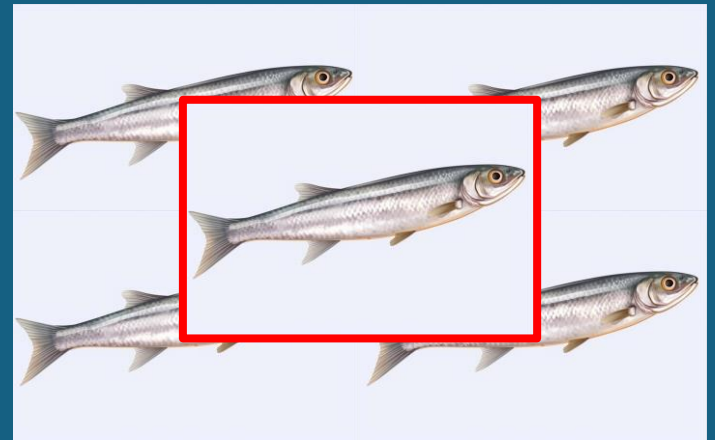
What are controlled data vocabularies?

Shared language: a controlled vocabulary is simply a list of agreed-upon words, terms or concepts that everyone uses consistently to mean the same thing.

- Car / automobile / auto → **pick one**
- Herring / Baltic herring / *Clupea harengus* → **pick one**

Otherwise, the computer thinks they're three different things, even if humans know they're the same.

Controlled vocabularies therefore **reduce ambiguity and improve data retrieval**



Life without controlled data vocabularies

No vocabulary = everyone puts food in the fridge wherever they want, labelled however they feel.

With vocabulary = Milk always goes on the same shelf, with the same label.

Result:

- You find what you need faster
- Less spoiled data
- Less duplication

Vocabularies **improve data quality and consistency.**



Examples of existing vocabulary servers

EU vocabularies – provide wide range of different types of controlled vocabularies used across EU institutions. INSPIRE directive defines themes and code lists for spatial datasets.

GEMET – GEneral Multilingual Environmental Thesaurus, developed by EEA. Widely used in data standards (e.g. INSPIRE datasets).

The NERC Vocabulary Server – publishes standardized term lists used in oceanography, climate, and environmental sciences. Operated by the British Oceanographic Data Centre.

ICES vocabularies – the authoritative reference for ICES-managed data streams. Maintains its own controlled vocabularies, actively used in marine and fisheries data flows.

WoRMS – expert-validated classification of marine species names, synonyms, and taxonomic hierarchies.

Examples of existing vocabulary servers

Countries and territories – lists concepts associated with names of countries and territories.

Marine species – an authoritative classification and catalogue of marine names.

Human activities – EU EMODnet categories of human activities, structures or organisational boundaries in or near the marine environment.

HELCOM 'HUB' Underwater Biotope and Habitat Classification System – provides a framework for defining biotopes in the whole Baltic Sea.

EU Vocabularies

EU Vocabularies

[Publications Office](#) | [EU law](#) | [European data](#) | [EU tenders](#) | [EU research results](#) | [EU Whoiswho](#) | [EU publications](#)[Publications Office](#) > [EU Vocabularies](#) > [Controlled vocabularies](#)[Share](#) [? H](#)[Home](#) | [Controlled vocabularies](#) ▾ | [Models](#) ▾ | [Business collections](#) ▾ | [Online tools](#) ▾ | [Releases](#) | [Help](#) ▾

Controlled vocabularies



Controlled vocabularies provide a consistent way to describe data. They are standardized and organized arrangements of words and phrases presented as alphabetical lists of terms or as thesauri and taxonomies with a hierarchical structure of broader and narrower terms.

Authority tables

In order to harmonise and standardise the codes and the associated labels used in the Publications Office and on interinstitutional level in the context of the data exchange between the...

[Learn more >](#)

Code lists

In order to harmonise and standardise the codes and the associated labels used in the Publications Office and on interinstitutional level in the context of the data exchange between the...

[Learn more >](#)

Thesauri

A thesaurus is a controlled and structured vocabulary where concepts are represented by labels. In the context of the EU Vocabularies, a thesaurus is a multilingual equivalent of the previous...

[Learn more >](#)

ATTO tables

ATTO is an internal application of the Publications Office for managing translations. The acronym comes from the French term 'Atelier des Tables de Traductions de l'Office des publication...

[Learn more >](#)

Alignments

Alignments are common in database interoperability projects and tasks however, from the perspective of the semantic web we are referring to ontology alignments when a variable...

[Learn more >](#)

Taxonomies

A taxonomy is a set of controlled vocabulary terms organised into a hierarchical structure. Each term in a taxonomy is in one or more parent/child relationship to other terms in the...

[Learn more >](#)

Purpose of data vocabulary

Maintaining terms in one single place.

Publishing vocabulary for internal or public use.

Referring to vocabulary terms when needed (publications, documents, webpages, dataset names, attribute names, unique attribute values, metadata record names, keywords).

Using vocabulary during data collections for data quality check (validation).

It's not about telling people what words they're allowed to use. It's about agreeing on what the computer should understand. Humans can still speak freely. Computers need one agreed meaning.

HELCOM and PROTECT BALTIC use shared vocabularies to avoid misunderstandings across countries and systems.

How to use terms of the vocabulary

Each term has a webpage describing the term (i.e., name tag with home address)

Each term has a URL (link), which should be used when the term is mentioned

(term + persistent identifier)

Vocabulary comes with programming interface (API), which should be used to automatically update terms in datasets, lists, websites:

- Update tasks on request,
- Scheduled (e. g. monthly) updates.

Using existing vocabularies

Find suitable vocabulary:

- Vocabulary contains suitable terms,
- Maintained by trusted organization.

<https://www.marinespecies.org/aphia.php?p=taxdetails&id=1268>

Refer to the term with URL:

- For publications, documents, metadata descriptions:
 - Use the link functionality - [Peanut worm](#).
- For dataset names and attribute names:
 - Use the link functionality when describing dataset or attribute in the metadata record.
- For attribute values within the dataset:
 - Create a separate attribute (next to existing one) and paste a link to the term in that attribute.

The screenshot displays the WoRMS website interface for the taxon **Sipuncula**. The page title is "WoRMS taxon details" and the taxon name is "Sipuncula". The page provides a comprehensive overview of the taxon's classification and related information.

Classification: Biota (Kingdom) → Annelida (Phylum) → Sipuncula (Order)

Authority: Stephen, 1965

Status: accepted

Rank: Order

Parent: Annelida

Synonymised names: Sipunculida - unaccepted (synonym)

Direct children (17): A list of 17 families, including Antillesomatidae, Aspidosiphonidae, Gollingidae, Phascosomatidae, Siphonosomatidae, and Sipunculidae, with their respective authorities and dates.

Environment: marine, brackish, fresh, terrestrial

Fossil range: recent + fossil

Original description: Stephen, A. C. (1965). A revision of the classification of the phylum Sipuncula. *Annals and Magazine of Natural History*. (See note: As Sipuncula [details] Available for editors [request])

Taxonomic citation: WoRMS (2026). Sipuncula. Accessed at: <https://www.marinespecies.org/aphia.php?p=taxdetails&id=1268> on 2026-03-14

Taxonomic edit history: Date: 2024-12-21 15:54:06Z, Action: created, by: [redacted]

How to setup and maintain

Knowledge of terms is needed.

A list of vocabularies with terms.

A database for vocabulary terms.

A website to view vocabularies and terms (e. g. <https://vocab.helcom.fi>).

An API to access vocabularies and terms programmatically.

An administration website to maintain vocabularies and terms.

Responsible people to maintain vocabularies and terms.

Backend development knowledge is needed to implement vocabulary server.

In summary...

Controlled vocabulary: A list of agreed words that computers don't misinterpret.

Why it matters: Same word = same meaning = usable data.

Without it: Search breaks. Integration breaks. Trust breaks.