Controlled vocabularies for the social sciences: what they are, and why we need them

Short article appeared on the ODISSEI website on October 3rd, 2022 [link]

Author	Organisation	ORCID
Angelica M. Maineri	ODISSEI https://ror.org/03m8v6t10	https://orcid.org/0000-0002-6978-5278

Brief summary: Resources like controlled vocabularies, taxonomies and thesauri are essential to achieve interoperability. The aim of this short article is to provide definitions of the terms, in order to understand their use in the context of annotating the metadata behind the ODISSEI Portal.

Introduction

"I know what I meant to mean!", a frustrated Gloria (Sofia Vergara's Modern Family character, season 6, episode 7) exclaims after being mocked for misusing a figure of speech. These types of misunderstandings are common in human interactions, but most of the time they can be disambiguated, for instance by explaining a certain term, using a more widespread synonym, or describing the context in which that term is used.

There are instances in which we need to do the same with machines. For instance, in the data management landscape, the Interoperability principle (the 'I' in FAIR) relies on the idea that even the machine 'understands what we mean', i.e. the machine must be able to achieve a basic understanding of terms and concepts [1]. In this short article, I will describe some resources that are used to help humans and machines make sense of metadata: controlled vocabularies, taxonomies and thesauri. Ironically, these terms are not always used correctly, and their applications are wider than it is presented here. The aim of this short article is to provide definitions to understand their use in the context of annotating the metadata ingested into the ODISSEI Portal, a repository that combines metadata from a wide variety of research data sources into a single interface. Thanks to the annotations enabled by controlled vocabularies and related resources, the ODISSEI Portal will allow advanced semantic queries to support findability and interoperability.

Controlled vocabulary

Most of us are familiar with a vocabulary, or a collection of terms with their definitions. What are the features of a controlled vocabulary (CV) then? A controlled vocabulary - or terminology - can be defined as "a **normative** collection of terms, the **spelling of which is fixed** and for which **additional information** may be provided such as a definition, a set of synonyms, an editor, a version, as well as a licence determining the condition of use." [2]. There are some important elements in this definition:

- *Normative* means that there are policies to establish roles and responsibilities, such as who maintains and curates the list, and how to propose changes (e.g. adding a term) (see also [3]).
- *Fixing the spelling* allows to reduce ambiguity and the risk of duplication [4]. Think of FAIR and fair.
- *Additional information* clarifies the meaning of the terms and the conditions of use. This additional information is not mandatory, but is helpful for those who want to (re)use the CV.

In their non-hierarchical form, whereby no formal relationships are specified among terms, CVs can be used to label digital resources, to facilitate retrieval of said resources, and to filter the results of a search.

Example: DDI vocabularies

The Data Documentation Initiative (DDI) is an international standard for describing the data produced in the Social, Behavioural and Economic (SBE) Sciences domain. Within the DDI alliance, many controlled vocabularies are being developed to describe the (meta)data. The CVs consist of lists of terms with their definitions. A Controlled Vocabulary Working Group is in charge of reviewing the comments, additions, and edits proposed by the community.

Take for instance the DDI Controlled Vocabulary for Mode of collection: the list represents the most common ways to collect data in the SBE sciences. It includes an unambiguous 'Value of the Code', but also a 'Descriptive Term of the Code', alongside a definition. Using such a vocabulary helps avoiding mistakes such as using 'PAPI' to refer to self-administered paper questionnaires.

Code List

Value of the Code	Descriptive Term of the Code	Definition of the Code
Interview	Interview	A pre-planned communication between two (or more) people - the interviewer(s) and the interviewee(s) - in which information is obtained by the interviewer(s) from the interviewee(s). If group interaction is part of the method, use "Focus group".
Interview.FaceToFace	Face-to-face interview	Data collection method in which a live interviewer conducts a personal interview, presenting questions and entering the responses. Use this broader term if not CAPI or PAPI, or if not known whether CAPI/PAPI or not.
Interview.FaceToFace.CAPIorCAMI	Face-to-face interview: Computer-assisted (CAPI/CAMI)	Computer-assisted personal interviewing (CAPI), or computer-assisted mobile interviewing (CAMI). Data collection method in which the interviewer reads questions to the respondents from the screen of a computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the same device. The administration of the interview is managed by a specifically designed program/application.
Interview.FaceToFace.PAPI	Face-to-face interview: Paper-and-pencil (PAPI)	Paper-and-pencil interviewing (PAPI). The interviewer uses a traditional paper questionnaire to read the questions and enter the answers.
Interview.Telephone	Telephone interview	Interview administered on the telephone. Use this broader term if not CATI, or if not known whether CATI or not.
Interview.Telephone.CATI	Telephone interview: Computer-assisted (CATI)	Computer-assisted telephone interviewing (CATI). The interviewer asks questions as directed by a computer, responses are keyed directly into the computer and the administration of the interview is managed by a specifically designed program.

Figure 1. Screenshot from DDI Controlled Vocabulary for Mode of collection

Such controlled vocabularies are particularly useful to describe data consistently, i.e. using the same terms to describe the same feature across different datasets. However,

the application of the DDI CVs extends beyond the DDI itself, since such CVs could be used, for instance, to annotate Methods sections in journal articles.

Controlled vocabularies can also be structured, which means that some sort of relationship is established among the terms in the vocabulary. In the example above, for instance, PAPI is a subclass of face-to-face interview, which is a subclass of Interview. Well-known forms of structured CVs include taxonomies and thesauri.

Taxonomy

A taxonomy is a classification scheme [5], often represented by a tree structure. Unlike an unstructured CV, which is usually sorted alphabetically, in a taxonomy terms are nested into one another. The advantage of such a structure is that you can climb ladders of the hierarchy, and group information - this is especially useful when a large number of terms is involved. For instance, a cumulative search function may be designed around a taxonomy whereby one first selects the macro category, then further specifies the subcategories. On the other hand, if the results of a search are too narrow, a taxonomy can be helpful to expand the results to a more comprehensive group.

Example: ISCO classification

The European Skills, Competences, Qualifications and Occupations (ESCO) classifications group widely known taxonomies in the Social Sciences, such as the International Standard Classification of Occupations (ISCO). ISCO is a classification of occupations, each with its own description, and organised hierarchically. For instance (see Figure 2) General office clerks constitute a subgroup of General and keyboard clerks, which in turn constitute a subgroup of Clerical support workers.

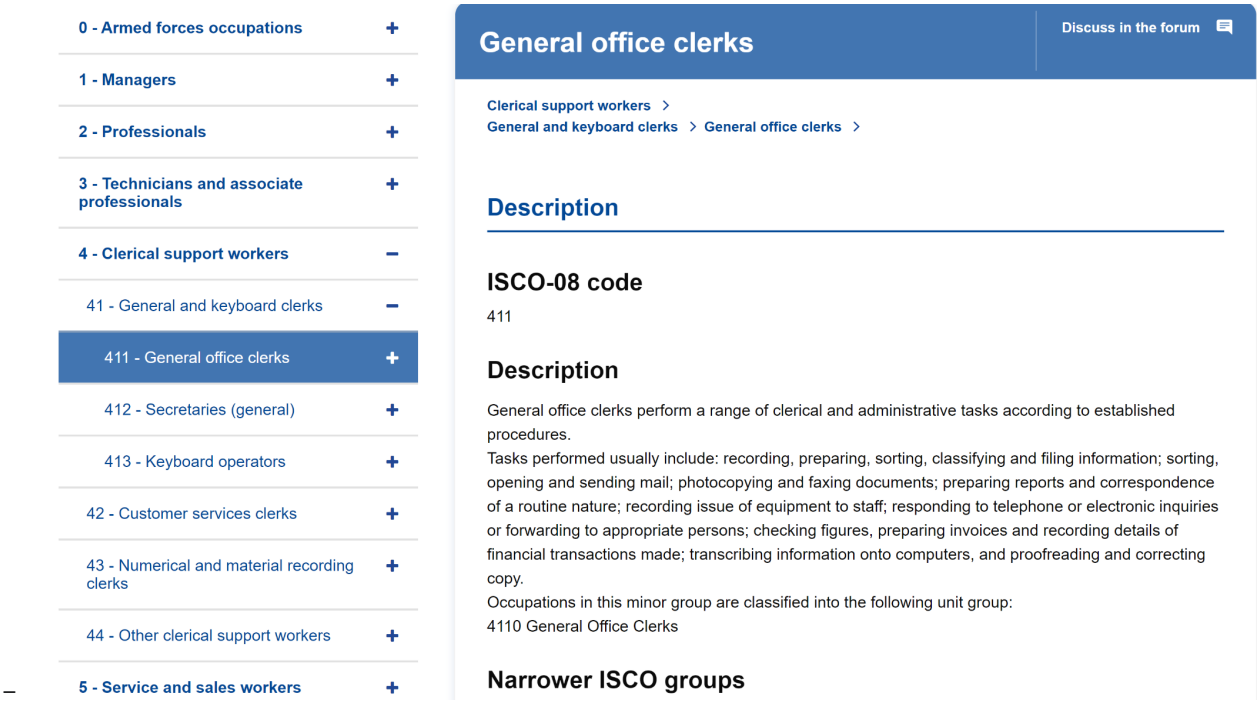


Figure 2: screenshot from European Skills, Competences, Qualifications and Occupations (ESCO)

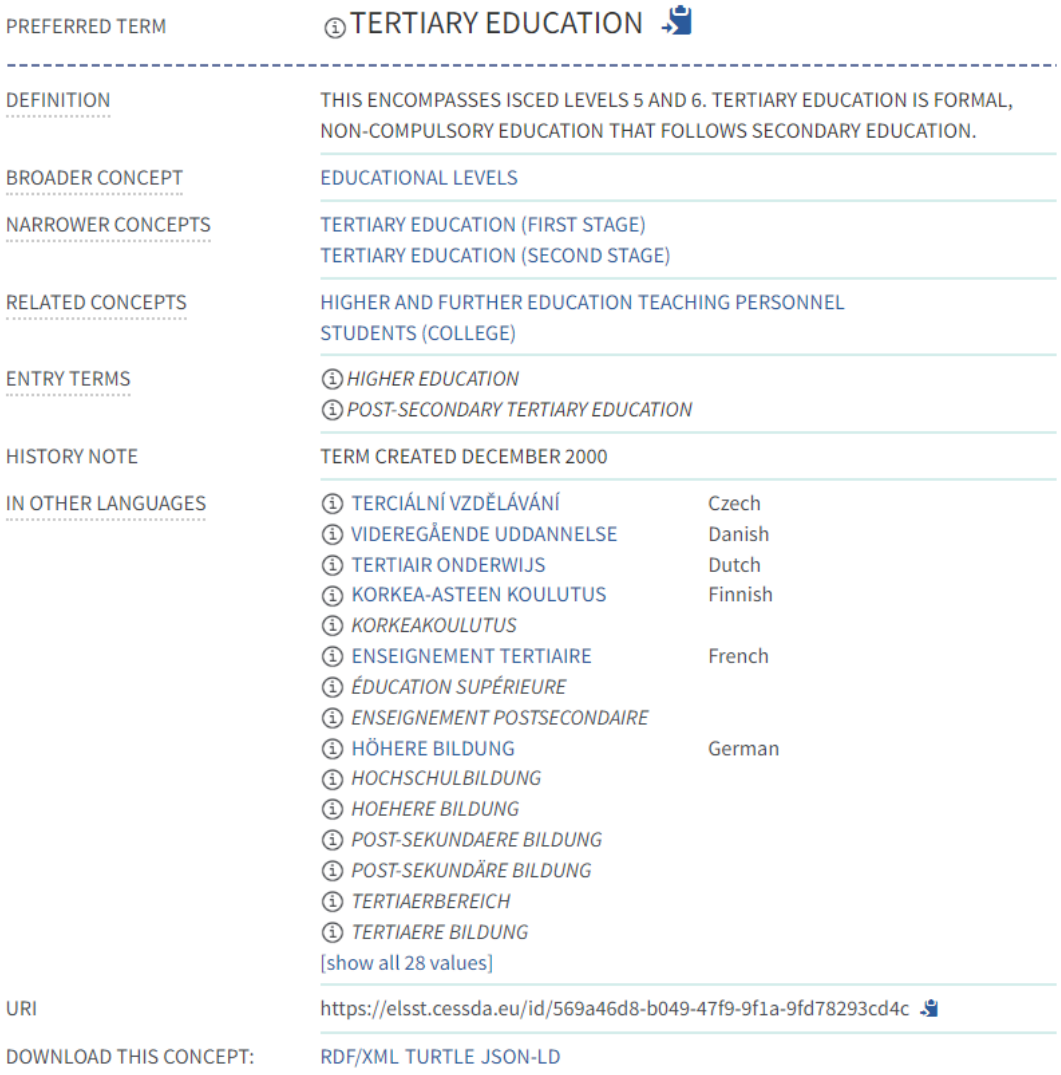
Thesaurus

A thesaurus is a structured CV in which relationships of hierarchy ("broader" "narrower", "contains"), association ("see also", "related to"), and equivalence ("a.k.a.") are established among the terms representing concepts [3,6]. The hierarchical relationships help classifying (similar to a taxonomy; in fact, a thesaurus can be considered as a more complex form of taxonomy), associations allow to link concepts that are closely related, and equivalence enables to represent links such as synonyms and translations.

Example: ELSST

The European Language Social Science Thesaurus (ELSST) represents a structured list of 3000 concepts relevant for the social sciences. For each term, whose spelling is specified in the "Preferred term" field, a definition is provided to describe the meaning of the term. Furthermore, terms are organised in terms of hierarchy ("see Broader

Concept", "Narrower Concepts"), association ("see Related Concepts"), and equivalence ("see Entry terms", "In other languages").



PREFERRED TERM	TERTIARY EDUCATION
DEFINITION	THIS ENCOMPASSES ISCED LEVELS 5 AND 6. TERTIARY EDUCATION IS FORMAL, NON-COMPULSORY EDUCATION THAT FOLLOWS SECONDARY EDUCATION.
BROADER CONCEPT	EDUCATIONAL LEVELS
NARROWER CONCEPTS	TERTIARY EDUCATION (FIRST STAGE) TERTIARY EDUCATION (SECOND STAGE)
RELATED CONCEPTS	HIGHER AND FURTHER EDUCATION TEACHING PERSONNEL STUDENTS (COLLEGE)
ENTRY TERMS	*HIGHER EDUCATION* *POST-SECONDARY TERTIARY EDUCATION*
HISTORY NOTE	TERM CREATED DECEMBER 2000
IN OTHER LANGUAGES	TERCIÁLNÍ VZDĚLÁVÁNÍ (Czech) VIDEREGÅENDE UDDANNELSE (Danish) TERTIAIR ONDERWIJS (Dutch) KORKEA-ASTEEN KOULUTUS (Finnish) *KORKEAKOULUTUS* ENSEIGNEMENT TERTIAIRE (French) *ÉDUCATION SUPÉRIEURE* *ENSEIGNEMENT POSTSECONDAIRE* HÖHERE BILDUNG (German) *HOCHSCHULBILDUNG* *HOEHERE BILDUNG* *POST-SEKUNDAERE BILDUNG* *POST-SEKUNDÄRE BILDUNG* *TERTIAERBEREICH* *TERTIAERE BILDUNG* [show all 28 values]
URI	https://elsst.cessda.eu/id/569a46d8-b049-47f9-9f1a-9fd78293cd4c
DOWNLOAD THIS CONCEPT:	RDF/XML TURTLE JSON-LD

Figure 3. Screenshot from ELSST

The power of CVs

Now that these concepts have been clarified, it is important to understand why we need these resources.

First of all, CVs ensure consistency, which fosters interoperability. Confusion between terms such as '(the) Netherlands' and 'Holland' (an incorrect yet widespread way to refer to the whole country), can hinder interoperability, because a machine may not understand that the same geographical area is referenced (unless otherwise specified). A solution is either to use a controlled vocabulary, and therefore pick the term - correctly spelled - from the list, or a thesaurus in which the different terms are linked through an equivalence relationship.

When relationships between concepts are specified like in a thesaurus, indexing and query recall can improve quickly. Think of a multilingual setting: when the terms used in the metadata are annotated using the thesaurus ELSST, a researcher is able to search a concept in their own language (e.g. 'onderwijs') and yet find relevant resources in other languages (e.g. on 'education').

Where to find CVs

Available registries of Controlled Vocabularies and linguistic resources include the EU's repository of Controlled Vocabularies and the Basic Register of Thesauri, Ontologies & Classifications (BARTOC). General purpose thesauri include the English lexical database WordNet and the multilingual resource BabelNet, which also lists entities such as cities and characters.

In the social sciences, there are CVs available to annotate data descriptions (e.g. DDI vocabularies to annotate mode of data collection, sampling strategy, etc), and subjects or topic classifications (e.g. CESSDA Topic classification). The CESSDA Vocabulary Service hosts many of the DDI vocabularies and CESSDA vocabularies. As concerns concepts, which are often used to search data or literature, resources are different by field. In the psychological field, the American Psychological Association maintains a thesaurus of psychological terms (>10,000 terms); the Leibniz Information Center for Economics publishes a thesaurus for Economics (> 20,000 entries). ELSST (see above, 3,000 entries) is perhaps the most comprehensive multilingual thesaurus for the social sciences, whereas the SAGE Social Science Thesaurus lists >60,000 entries (in English) extracted from SAGE encyclopaedias and other resources.

Conclusion

In the FAIR perspective, "controlling the vocabulary helps machines and humans to categorise information and helps reduce redundancy and errors." [4], or - in other words - to 'understand what we mean'. To be machine-actionable, these resources need to be expressed in specific formats, and concepts have to be assigned unique IDs - these topics will be explored further in a future blog post.

Do you have comments or questions, or do you know Controlled Vocabularies that should be added to the list? Let us know via fairsupport@odissei-data.nl.

Relevant links

- Guidelines on how to FAIRify vocabularies
- ODISSEI FAIR support
- ODISSEI Portal

References

[1] Jacobsen et al. (2020). FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2 (1-2): 10–29. doi: https://doi.org/10.1162/dint_r_00024

[2] Rocca-Serra, P. & Gray, A.G.J. (nd). FAIR Cookbook - Chapter 3 Introduction to terminologies and ontologies.
https://faircookbook.elixir-europe.org/content/recipes/interoperability/introduction-terminologies-ontologies

[3] Bowen, D. (2022). Controlled Vocabulary, Thesaurus, Ontology. https://www.dianebowen.net/controlled-vocabulary-thesaurus-ontology.html#:~:text=For%20a%20controlled%20vocabulary%2C%20an,with%20a%20kind%20of%20taxonomy..

[4] EMBL-EBI Training (2022a). Bioinformatics for the terrified: Controlled Vocabularies. https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/controlled-vocabularies/

[5] EMBL-EBI Training (2022b). Bioinformatics for the terrified: Taxonomy.

https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/controlled-vocabularies/taxonomy//

[6] EMBL-EBI Training (2022c). Bioinformatics for the terrified: Thesaurus.

https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/controlled-vocabularies/thesaurus/