

Converting and evolving a subject heading list into a thesaurus

Maria Teresa Guaglianone, Giovanna Aracri and Maria Taverniti
National Research Council, Institute of Informatics and Telematics, Rende, Italy

Abstract

Purpose – The objective of this paper is to describe the evolution of the available subject heading list, i.e. the CC Soggettario (Carabinieri Corps Soggettario), towards a thesaurus, that is CCThes (Carabinieri Corps Thesaurus), to support subject indexing and retrieval of the documentary heritage held by the Historical Office of the General Command of the Carabinieri Corps. This work follows the need to implement a controlled vocabulary compliant with the state-of-the-art standards.

Design/methodology/approach – The methodology implements the practice of reengineering available vocabularies, following standardised guidelines for thesaurus development. The conversion process includes the balance maintenance of what has been achieved in the CC Soggettario and the enrichment of the semantic structure in the thesaurus by using both deductive and inductive methods.

Findings – The main result of this study is a thesaurus compliant with ISO 29964-1:2011 recommendations, which improves information retrieval performances and interoperability with other vocabularies and applications. It generally has a mono-hierarchical structure with the possibility of admitting, as an exception, the poly-hierarchy for a few concepts. An introductory user guide has been created as a complementary tool to the CCThes.

Originality/value – This is an applied study which deals with Knowledge Organisation System (KOS) reengineering and outlines this process using a pragmatic approach. The paper strength lies in providing the description of performed activities and conveying a set of resources to approach KOS reengineering practice. The study is also relevant for the preservation and diffusion of a part of the social memory and identity of Italy.

Keywords Knowledge organisation systems, Controlled languages, Indexing, Information retrieval, Thesaurus construction, Cultural heritage

Paper type Research paper

1. Introduction

To properly convey to users what a document is about, as well as assigning a classification and improve document and information retrieval, it is appropriate to represent its content by means of some keywords taken from controlled vocabularies, that is authoritative lists of terms to be used in indexing, which are predefined and supported by international standards (De Keyser, 2012).

Indexing is mainly based on the usage of a subject heading list or a thesaurus. Both are Knowledge Organisation Systems (KOSs), with differences in the development and in the representation of semantic relationships (Chatterjee, 2017). The former is based on less strict guidelines especially deriving from explicit theory and practice, the latter is designed and

© Maria Teresa Guaglianone, Giovanna Aracri and Maria Taverniti. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at [Link to the terms of the CC BY 4.0 licence](https://creativecommons.org/licenses/by/4.0/)

The paper is the result of the collaboration among the Historical Office of the General Command of Carabinieri Corps, the Informatics and Telematics Institute – secondary branch of Rende (Cs) – of the National Research Council (IIT-CNR) of Italy and the Department of Culture, Education and Society of the University of Calabria (DiCES-UniCal).

Author contributions: Conceptualization, M.T.G., G.A. and M.T.; methodology, M.T.G., G.A. and M.T.; investigation, M.T.G., G.A. and M.T.; writing – original draft, M.T.G. (paragraphs 3-4-5) and G.A. (paragraphs 1-2); writing – review and editing, M.T.G., G.A. and M.T. All authors have read and agreed to the published version of the manuscript.



constructed by following international standards specifications, able to guarantee consistency and quality and to favour interoperability (Dextre Clarke, 2019).

These vocabularies control the form of index terms, provide guidelines for their use and specify the relationships between terms. The achievement of such goals can occur in different ways and with targeted strategies, for example subject heading lists mainly use an alphabetical presentation, whereas thesauri also provide a systematic organisation, given by the semantic network. In fact, they contain different types of semantic relationships and more specific and rich terminology. Moreover, they allow for an additional explicit visualisation of the relational structure in the form of categorised lists or graphic displays, which support indexers and users (Chatterjee, 2017), helping to have a general comprehension of a domain of knowledge and outline inter-relationships between concepts (Aitchison *et al.*, 2000).

This paper focuses on the evolution of a subject heading list, i.e. the CC Soggettario (Carabinieri Corps Soggettario), towards a thesaurus, serving as an indexing and retrieval tool for the documentary heritage held by the Historical Office of the General Command of the Carabinieri Corps which is a fundamental piece of the Italian history. The Carabinieri Corps is one of the Italian police forces, with general functions of control, civil police and public security service. At the same time, it is the fourth Italian Armed Force, and it contributes to the homeland defence: by taking part in international peacekeeping missions; by performing military policy activities; by ensuring security in Italian diplomatic and consular offices around the world. [1] The ultimate purpose of this conversion is to update and enhance the content of the CC Soggettario and to implement the CCThes as more efficient indexing tool, which enables users: to provide both on-site and remote access to resources; and to promote culture by rendering available and accessible the knowledge embedded in the vast historical heritage preserved by the Carabinieri Corps. Finally, the evolution of the CC Soggettario towards the CCThes arises from the need to be compliant with the state-of-the-art standards in controlled vocabulary development and use. Conformity to a standard has as a direct consequence the possibility to enrich the range of subjects that can be used for indexing and classification by including explicitly and conventionally formalised semantic relationships, which ensure terminological control and interoperability.

The paper has an applied slant, and it is organised to provide the reader with a detailed and pragmatic description of the performed activities. The reflections can be used as a sort of toolkit by the community approaching the KOS reengineering practice. As known, there are several studies about how subject heading lists and thesauri should be developed, and how these latter tools can evolve into ontologies. Therefore, the theory on how to convert subject heading lists into thesauri can be fathomed from the literature. Nevertheless, this paper intends to explicitly clarify the necessary steps and the methodological choices of the conversion process. This perspective can be useful to the current state of knowledge. The paper is organised as follows: firstly, some initiatives of terminology standardisation and harmonisation in the military domain are presented; then, the CC Soggettario, as the starting tool of the study, is described; subsequently, the methodological path that led to the conversion process is described; lastly, some of the results achieved so far are shown.

2. Background

Reengineering KOSs is a common approach to reuse existing vocabularies instead of redesigning them from scratch (Soergel, 2009). Several initiatives have focused on converting glossaries (Hilera *et al.*, 2010), thesauri (Wielinga *et al.*, 2001) or subject headings (Rout, 2018) into ontologies to improve subject retrieval and to encourage the use of ontology-based systems. The improvement of a simple structure by means of the identification and formalisation of conceptual relationships between entities (such as terms, concepts, resources and data) is quite a common practice in several domains. The aim is to improve information retrieval and to achieve higher levels of interoperability. This trend is also consolidated in the

military domain where several ontology-based applications have been developed over the years (Bowman *et al.*, 2001). It is a strategic domain that demands effective and unambiguous communication to reduce misunderstanding and inefficiency.

Favouring the historical point of view, oriented to the enhancement of the archival heritage, in the following subsections the characteristics and scopes of the two categories of controlled vocabularies compared in this study, that are subject heading lists and thesauri, are primarily outlined, and then, an overview about some vocabularies in the military domain is provided.

2.1 Subject heading lists and thesauri: an overview

Subject heading list and thesaurus are documentary languages, and all their terms are normalised according to reference rules to control their form and meaning. More specifically, they are employed in document indexing to guarantee the coherent, formalised and concise representation of the conceptual content of the documents. As mentioned, there are some differences between them, but the most relevant one is that subject heading lists belong to the category of pre-coordinated languages, whereas the thesaurus is a post-coordinated language (De Keyser, 2012). In the former, terms used to express concepts are combined during indexing, hence the derived string is, in its entirety, related to the document providing a complete content overview; in the latter, terms are directly related to the document and not to each other (Spinelli, 2006) and, in turn, they can be linked during search by means of the use of the Boolean operator. In addition, the thesaurus can be visualised as a flat list of keywords, in an alphabetical and systematic order or visualised in both ways by making clear all the semantic relationships between terms and between concepts (Broughton, 2006).

Traditionally, subject heading lists are used for organising and searching documents in a library. They include a master list of subjects associated with bibliographic records which, in turn, contain the more relevant information on resources collected in a catalogue. The subject, as for author, title, publisher, etc. is a bibliographic metadata and gives information about document content rather than its structure. Hence, it provides a diversified access point (a preferred entry) to the bibliographic collection and may be represented by means of both natural language and conventional notations, usually through a classification system. According to the indexing granularity level, the subject or the range of subjects are expressed by means of a single code or a combination of codes. The terminological control function, if present, is accomplished by managing synonymic phenomena essentially by using cross-references. Subject heading lists, following the principles conceived by Cutter (1904), are also aimed to guide the user to search for all the documents dealing with a specified subject.

The Library of Congress Subject Headings (LCSH) [2] is a large and authoritative subject indexing language adopted around the world and maintained by the Library of Congress. It has an extensive coverage, even though some studies criticise its hierarchical organisation (Mazzocchi, 2018) and its quality level claiming the need of some improvements (Barros and Hjørland, 2024). It is available in several languages and because of its considerable number of headings it is continuously updating for guaranteeing a suitable information representation. It is used by multidisciplinary libraries, and it is rich with subject headings related to diversified fields of expertise. As outlined by Dykstra (1988), because of the conceptual structure adopted, often the LCSH is being assimilated to a thesaurus, even if it really is not. The rules for creating the subject headings and the relationships between them follow an *alphabetic-classed plan*, more explicitly in a thesaural structure and the references between headings are expressed by means of the thesaural labels. It is continuously updated: the first edition was printed between 1909 and 1914, the last edition, which is the 43rd, in 2021.

Concerning the thesaurus, it is a conceptual systematic model, and concepts are unambiguously designated by terms (International Organization for Standardization, 2011), which identify concepts rather than subjects (Dykstra, 1988). Concepts refer to a given

material or abstract reality, whereas subjects, as already explained, directly represent the topic of a document.

The thesaurus is a meaning-driven vocabulary and it implements a concept-based approach (Martínez-González and Luisa, 2019) based on the definition of explicit semantic relationships between pairs of concepts, in order to specify the semantic networks and to highlight which type of relationship occurs between them. In detail, the main objectives of a thesaurus concern: the identification of the descriptor, also named preferred term (PT), used to index the document and to recognise not accepted forms, that is non-descriptor or non-preferred term (NPT); the definition of hierarchical relationships to explain if a concept is broader or narrower than another and to explain if two concepts are members of the same class; the specification of other types of semantic relationships. All relationships between concepts are handled through the following standardised and paradigmatic semantic relationships:

- (1) An equivalence relationship is established between terms referring to the same concept, that are synonyms or quasi-synonyms.
- (2) A hierarchical relationship defines a membership relationship between two concepts. It is established between PTs following classification criteria, since the concepts are part of a network of hierarchical hypernyms and hyponyms.
- (3) An associative relationship binds together concepts that have a strong mutual implication for which it is not possible to establish hierarchical and/or equivalence relationships.

By means of semantic relationships and scope notes (SN), thesauri perform and achieve several tasks including manual and semi-automatic indexing, document and information retrieval, domain knowledge organisation and representation, metadata assignment, clustering and search expansion.

2.2 State of the art

In this paragraph several controlled vocabularies dealing with terminology standardisation and harmonisation in the military domain are presented. Their main objective is to release a set of normalised terms able to guarantee understanding during communicative events.

The NATO Glossary (North Atlantic Treaty Organization, 2021) is the main reference for controlled vocabulary designers in the military domain. It collects the terminology concerning the essential military affairs and communications and its purpose is to encourage a more shared and aware use of terminology. It is a bilingual resource (English and French) in which each preferred term is followed by additional information such as synonyms, abbreviations, related terms and date of acceptance. It consists of three parts: (1) organisation and procedures, (2) English terms and definitions and (3) French terms and definitions. To support information sharing and interoperability, and to improve the glossary's overall capacity to help experts in document production, it has undergone significant enhancements. The outcome of this update is *NATOTerm* [3], the official NATO Terminology Database made up of military and non-military terminology pertinent to NATO's objectives. The database can be queried, and each terminological entry includes the PT (that is the full form), the abbreviation if applicable, the definition if available, the term status and the source in both official languages. Furthermore, *NATOTerm* is used as hub vocabulary for improving interoperability in some terminological resources including those described below.

MiliMarco is based on the model set provided by *NATOTerm* plus other information, such as conceptual category, semantic relationships, part of speech, term type, etc. It is the output of a project funded by the Spanish Ministry of Economy and Competitiveness aimed at converting a glossary into a knowledge base, and it is a bilingual (Spanish – English)

terminological knowledge base of concepts and semantic relationships precisely regarding the military area (Faber and Haraüz, 2019). The Italian National Forces have also made efforts to standardise their own terminology (Stato Maggiore della Difesa, 2009) in accordance with NATO Glossary to promote cooperation and mutual understanding within the Italian National Forces and between the Italian Defence and NATO. Regarding security, it is important to point out that the Intelligence Glossary was developed to make available to policy makers a suitable tool able in providing detailed knowledge on threats and dangers that could compromise the Italian Republic's national security (Presidenza del Consiglio dei Ministri, 2019). The Department of Defence (DoD) Dictionary of Military and Associated Terms (Department of Defense, 2016), developed by the Armed Forces of the United States, is similar in scope to the NATO Glossary. It provides a set of approximately 25,000 terms that support information exchange and interpretation within documents to improve semantic interoperability and reduce information loss and scattering in document planning. Terms are organised into three main categories: Military Terminology, Policy and Joint Doctrine and Strategic Effect Terms and all of them must follow a strict procedure to be included in the Dictionary. All the terms selected to represent a concept have to perform several criteria and are validated by DoD Terminologists. In the DoD Dictionary the term should be preferably expressed in the singular form with few exceptions; abbreviations and acronyms are not considered as primary entries, but they are encompassed in the main entry by means of "Also called", whereas the associative link is indicated as "see also".

The Library of Congress Subject Headings (LCSH), among the KOSs for indexing and cataloguing, must be mentioned. Some aspects of the military domain are here outlined by means of the specific class CLASS U – MILITARY SCIENCE organised in further nine subdivisions. As concerns the Italian national context, the more authoritative reference for those approaching subject indexing is the Nuovo Soggettario, a multidisciplinary thesaurus employed to support the subject indexing by the Italian libraries' system. It is used by the Italian National Bibliography (BNI) to record the bibliographic description of a document. The Nuovo Soggettario has been developed and is currently managed by the National Central Library of Florence. It is freely available on the web, constantly updated and, its own terminological, conceptual and relational set, increases considering the *literary warrant* principles, through the existing literature (Biblioteca Nazionale Centrale di Firenze, 2021). In fact, it also collects several concepts about military domain, some of which introduced by means of a specific node label [*Attività di ambito militare, di aggressione, di difesa*]/*Military, aggression, defence activities*] referred to the macro-category *Attività/Activity*. The Nuovo Soggettario has been a constant reference in the development of the CCThes since it is: compliant with international standards on subject indexing; in line with the semantic web best-practices; suitable for describing various kinds of resources; interoperable with other KOSs.

In this study, the aim is also to provide visibility and access to the information contained in the heritage preserved at the Historical Office of the General Command of the Carabinieri Corps, established in 1965 as the historical section of the Operations Office, whose mission is to collect and process significant data concerning activities achieved by Corps. To satisfy this aim, and for statistical data analysis for historical and communication purposes, the Transmission Command, directorates and General Command offices periodically provide copies of reports, notifications and relevant acts for documentation purposes (Carbone, 2017).

3. Methodology

This section examines the methodological approach used to realise the transition from the CC Soggettario to the CCThes, whose construction follows practices of reengineering and enhancing available vocabularies, such as subject heading lists, based on the standardised guidelines for thesaurus development of the ISO 29964-1:2011.

The CCThes aims to support the Carabinieri Corps in subject indexing improvement, entailing some important advantages: (1) reduction of semantic ambiguity; (2) enhancement of communication exchange within the community of practice; (3) more shared use of terminology by indexers and users; (4) more efficient and rigorous access to information; (5) improvement of information retrieval; (6) accomplishment of semantic interoperability and information reuse.

3.1 Carabinieri Corps Soggettario

Considering that the CCThes construction has as basis the existing CC Soggettario, this subsection focuses on its description, since it is the starting tool of this study. It is available in paper format with a brief user guide in the “Introduction”, and it consists of 1,058 terms, alphabetically ordered, and distinguished as follow:

- (1) 601 subject headings: graphically identified by means of the capitalisation convention and used to describe the historical and cultural heritage of the Carabinieri Corps, e.g. *ALFIERE/STANDARD BEARER*;
- (2) 457 other entries: reported in lowercase, that cannot be employed as subject headings and sometimes guide indexers to the choice of the suitable subject headings, e.g. *Avversità climatiche v. OPERAZIONE SOCCORSO (a.R.S.)/Climatic Adversities see RESCUE OPERATIONS (press review)*.

In addition to subjects, the CC Soggettario includes a set of symbols, abbreviations and other conventions aimed at specifying how subject headings should be employed for subject indexing. Subjects consist of a flat list of terms, in which the semantic connection between subjects is expressed through two different kinds of cross-references:

- (1) “see”, expressed by means of the abbreviation “v.”, is used to establish a relationship between an entry not suitable for indexing and the corresponding more appropriate one, e.g. *Alluvione v. OPERAZIONE SOCCORSO/Flood see RESCUE OPERATION*). There is no mutuality between the two concepts involved in this type of relationship;
- (2) “see also”, expressed by means of “v.a.”, is used to represent a strong semantic connection between subjects, e.g. *REINTEGRAZIONE NEL GRADO v.a. RIABILITAZIONE MILITARE/REINSTATEMENT IN RANK see also MILITARY REHABILITATION*). This relationship is mutual, so that the two subjects have the same relation towards one another.

To better understand how the CC Soggettario evolved in the CCThes and how subjects have been semantically organised, it is important to take a trace of the most important conventions used to index documents and to further detail their topic. In fact, to specify a subject considered not sufficiently suitable to fully express the topic of the document, in the CC Soggettario the key-subject may be followed by:

- (1) Another subject in brackets which explains the peculiarity (point of view) of the document. For example, the subject *Guerra mondiale 2^a (Africa Orientale)/World War II (East Africa)* refers to the documents which deal with those facts and events that occurred during the Second World War in East Africa;
- (2) The plus sign (+) which however indicates the need to provide a more specific subject because the key one is not sufficiently exhaustive to specify the document topics. Therefore, if a document is indexed by means of *STAZIONI +/STATIONS +* it will be necessary to further specify the general concept *STAZIONI/STATION* by indicating the name of the station *STAZIONE DI SALERNO/SALERNO STATION*.

3.2 Conversion process and methodological issues

In the conversion process, some general rules have been considered, where possible, to make CCThes construction easier and to balance both the maintenance of what has been achieved in the CC Soggettario and the enrichment of the semantic structure, according to thesaurus construction principles. Such preliminary requirements try: to maintain the original denomination of terms and, at the same time, fix any mistakes found; to reflect the CC Soggettario original structure and, contemporarily, enrich CCThes architecture and content; to avoid redundant and inconsistent information. Therefore, the CCThes is intended as an exclusive indexing and searching tool, albeit it respects the semantic paradigms and fundamental concepts of the CC Soggettario.

The thesaurus development process is known and includes some established tasks such as term collection, term adaptation and modification, choice of preferred or non-preferred terms, semantic relation and scope note definition (Aitchison *et al.*, 2000; Broughton, 2006). Considering that the CCThes has as basis the existing CC Soggettario, the methodological approach adopted includes an analysis of this tool. Such a preliminary study is essential to identify typical structures and to define and establish, as much as possible, conventions useful to guide the subsequent steps towards the conversion of both terms and semantic relationships. It has led to the identification of some recurring issues, essentially dealing with vocabulary control (singular and plural form, abbreviations, compound terms, etc.) (Aitchison *et al.*, 2000), and equivalent solutions. The objective is to reach high-quality and consistent results.

To satisfy this aim, the organisation of the activities in the form of a project planning framework is convenient, since it ensures control and coherence of the process (Figure 1).

Both deductive and inductive methods are adopted. At the initial stage of the conversion process deductive approach is applied, since the CCThes incorporates the CC Soggettario's term list, trying to define a domain basic conceptual structure. Subsequently, further relevant terms are incorporated, and the hierarchy is incremented, following the inductive approach. In fact, this latter method is to select a set of terms and include it in the CCThes once they are used in indexing new documents. The presence of new terms allows for the definition of further semantic relationships, which determine thesaurus structure expansion, making it more complete and more coherent (Aitchison *et al.*, 2000).

The CCThes is created through the TemaTres Controlled Vocabulary server [4], a web application which allows for the effective feasibility of the methodological choices, since it is a multi-user open source software to create and manage an online consultable thesaurus, giving the possibility: to provide quality assurance functions (e.g. by preventing duplicates,

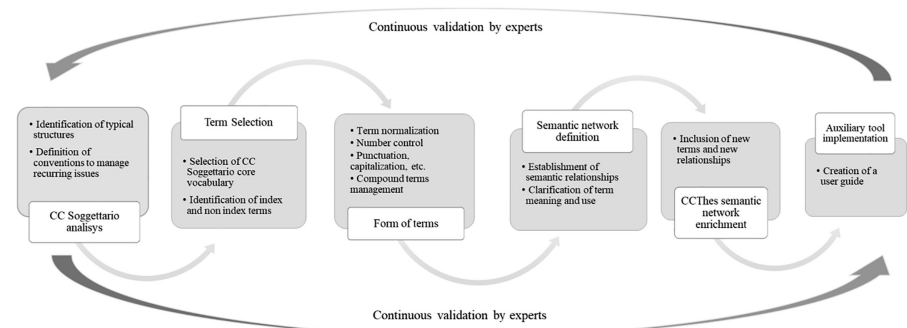


Figure 1.
Conversion process of
the CC Soggettario into
the CCThes

Source(s): Image by authors

free terms and illegal relations); to define new types of relationships and notes; to share and re-use formal and interoperable representations of knowledge.

3.2.1 Preferred and non-preferred terms identification. The CCThes construction process begins with the vocabulary selection and the identification of the most suitable terms for indexing the document collection. In our case the CC Soggettario is the master list, so it is necessary to face the management of subject and not subject headings of the CC Soggettario in their evolution towards PTs and NPTs of the CCThes. To point out that a concept has gone through some changes, the use of the Historical Variant relationship has been introduced to ensure continuity of use between the two indexing tools and to guarantee that documents previously indexed with the CC Soggettario continue to be accessible and retrievable by means of the CCThes. This relationship maps a CC Soggettario term, which is no longer preferred, to the new PT included in the CCThes establishing an equivalence relationship. It is expressed through the acronym UFHV (Used for Historical Variant), and it connects the PT of the CCThes and its historical variants coming from the CC Soggettario.

In the CC Soggettario these historical variants could be used or not used as subject headings, whereas in the CCThes, they always become NPTs and, therefore, not useable for indexing, but only for search purposes.

The example in [Figure 2](#) shows the case of the compound term *Guerra mondiale 2^a (Culquaber)/World War II (Culquaber)* that is made up of two related but different concepts, but only one of the two terms refers to and represents the subject. Thus, during the conversion process it has been divided in simple terms: the Battle of Culqualber, which happened during the Second World War, and the Second World War itself. The original compound term of the CC Soggettario is linked with the new PT *Battle of Culqualber* by means of the equivalence relationship and it is designated as Historical Variant. The relation with the concept Second World War, complementary to the main topic, is kept by establishing an associative relation.

Therefore, the core vocabulary of the CC Soggettario has gradually been selected, analysed and organised progressively designating PTs and NPTs to include in the CCThes, following the advice of experts at all stages of its compilation.

In the PT designation, ambiguity is avoided as far as possible. Hence, index terms are limited to one meaning within the reference domain. Moreover, the context provided by the structure in which the term is placed helps to clarify the meaning. Also, the PT itself is formulated in such a way that it conveys the intended scope in any language to any user who is familiar with the language in question ([International Federation of Library Associations and Institutions, 2009](#)).

3.2.2 Term normalisation. CC Soggettario's terms need a linguistic normalisation at a morphological and syntactic level (e.g. form of dates, ordinal numeral adjectives, proper names, compound terms, etc.). Morphological-syntactic conventions and rules are defined for the management of both the form of terms and the ways in which they can be combined in compound terms, recommending the use of simple and complex prepositions. The connection with CC Soggettario's term is maintained, as mentioned, by means of the equivalence relationship ([Figure 3](#)). The choice between UFHV or UF depends on whether the term belongs to the CC Soggettario or not: in the first case the UFHV relation is used to highlight the link with a subject heading listed in the CC Soggettario, whereas in the second case the UF relation is preferred, since it specifies an additional term that does not belong to the CC Soggettario, but which can still represent a valid entry to access the information conveyed by the PT.

The linguistic normalisation carried out within the CCThes considers both the most widely used conventions in the specific domain and what is expressed in the most consulted and authoritative dictionaries and indexing tools. The following sources are used for linguistic normalisation and terminological control: Nuovo Soggettario [\[5\]](#), AGROVOC Multilingual Thesaurus [\[6\]](#), EuroVoc [\[7\]](#), Treccani [\[8\]](#), Wikipedia [\[9\]](#), official website of the

Figure 2.
Example of
equivalence
relationship UFHV

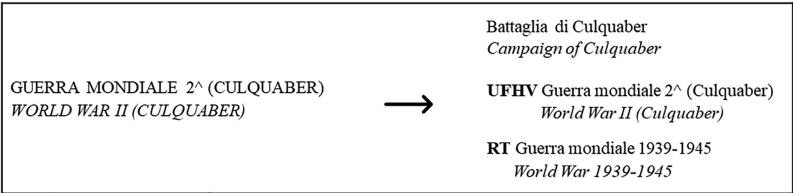
Carabinieri Corps – Curiosità section [10], Codice Ordinamento Militare (Ministero della Difesa, 2010).

3.2.3 *Number control.* The analysis of the CC Soggettario highlights that the presence and use of terms in the singular or in the plural form is not always justified and coherent. The number control is necessary since the form of terms is discriminating in term categorisation and meaning understanding. “A fundamental aspect of nominal meaning is whether the entity is countable or not” (Anderson *et al.*, 2022), so that the noun countability principle is applied to cases of indiscriminate use of the singular or plural form. The singular form is used for abstract and uncountable entities and the plural form for concrete and countable entities (see Figure 4).

3.2.4 *Compound term management.* The CC Soggettario makes an extensive use of compound terms that also occur combined with other compound or simple terms enclosed in round brackets, which specify the point of view from which the document subject or topic is approached, for example *GUERRA MONDIALE 1^ (CARABINIERI PILOTI)/WORLD WAR I (CARABINIERI PILOTS)*.

The adopted solution to manage the extensive use of the CC Soggettario’s compound terms is the term splitting and the maintenance of the close connection and implication between terms (see Figure 5) through:

- (1) The use of semantic relationships, especially the associative one (RT);
- (2) The use of only one of the two terms;
- (3) The combination of both terms into a single compound entry;
- (4) The splitting of the concept, using the hierarchical relationship (BT-NT).



Source(s): Image by authors

Figure 3.
Examples of term
normalisation

CAMPAGNA AFRICA ORIENTALE 1935/1936 EAST AFRICA CAMPAIGN 1935/1936	Campagna d'Africa Orientale 1935-1936 Campaign of East Africa 1935-1936 UFHV Campagna Africa Orientale 1935/1936 Campaign of East Africa 1935/1936
GUERRA MONDIALE 1^ WORLD WAR I	Guerra mondiale 1914-1918 World War 1914-1918 UFHV Guerra mondiale 1^ World War I
MAZZINI	Mazzini, Giuseppe UFHV Mazzini

Source(s): Image by authors

Complex compound term management is the main way through which the CCThes acquires its own hierarchical structure since the result of this process is a preliminary term organisation in a semantic network.

3.2.5 Graphical convention management. As mentioned, the CC Soggettario has its own graphical convention, that is the “+” sign, which follows the subject headings to indicate that a specification by means of a narrower term, not listed in the CC Soggettario, is needed. In this case, the excessive discretion of the indexer could compromise terminological control and indexing consistency. Hence, additional new terms and hierarchical relationships are introduced in case of subject headings followed by the “+” sign, since it represents a concept specification. In case of a list of more specific terms, they can be inserted as branches of the taxonomy and used for indexing instead of the more generic terms. Indications are given on the use of the term in a cataloguing note (see [Figure 6](#)).

3.2.6 Cross-reference evaluation and conversion. Cross-references “v.” (see) and “v.a” (see also) in the CC Soggettario are not always used in a coherent way and they need to be normalised together with the other critical aspects previously identified. Therefore, they are evaluated and transformed in associative or equivalence relationships, depending on the case.

As intended in the CC Soggettario, the cross-references indicated with “v.” (see), which refer to the transformation of a not used subject heading towards the corresponding one to be used, should imply the specification of a synonymy relationship between an accepted subject heading and one or more unaccepted variants. However, this condition is not always valid and more often the appropriate relationship between the two terms is not the equivalence relationship, as expected, but the associative one ([Figure 7](#)).

The cross-references designated by “v.a” (see also) connect similar headings. This correlation in the CCThes is mainly expressed by means of the associative relationship, linking related terms (see [Figure 8](#)).

3.2.7 Semantic network enrichment. The semantic network of the CCThes has been enriched to make it as complete, consistent and coherent as possible. This operation is convenient to optimise the coverage level of the CCThes, avoiding gaps and solving inconsistencies in semantic structure. To achieve this aim, the CCThes has been integrated with:

- (1) Notes, which contain additional information about the term. They are based on the nature and purpose of the information they provide, and they are divided into:
 - Scope notes, which restrict or expand the meaning of a term, distinguish between two terms that have overlapping meanings in natural language, or provide guidance on the semantic use of a term. They can contain the term definition and

Uncountable entities	Countable entities
Addestramento <i>Training</i>	Pistole <i>Guns</i>
Disciplina <i>Discipline</i> (intended as observance of rules of conduct or behaviour)	Discipline <i>Disciplines</i> (intended as teaching and study subjects or branch of knowledge)

Source(s): Image by authors

Figure 4.
Examples of terms in
the singular and
plural form

Figure 5.
Examples of
compound term
management

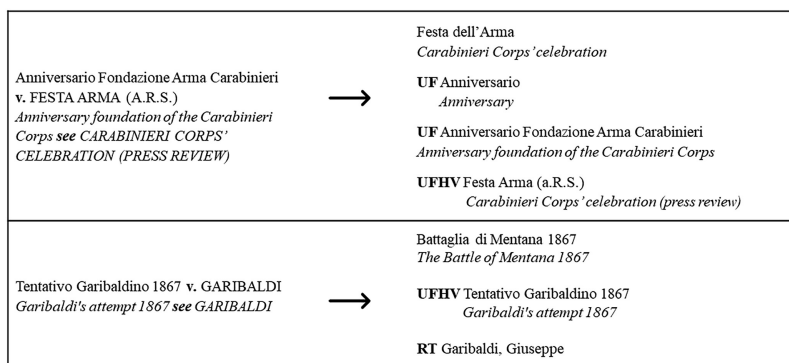
GUERRA MONDIALE 1^ (CARABINIERI PILOTI) WORLD WAR I (CARABINIERI PILOTS)	→	Carabinieri Piloti Carabinieri Pilots UF Carabinieri aviatori Carabinieri aviators UFHV Guerra mondiale 1^ (Carabinieri Piloti) World War I (Carabinieri Pilots) RT Guerra mondiale 1914-1918 World War 1914-1918
PREMIO (PRIVATO) PRIZE (PRIVATE)	→	Premi Prizes UFHV Premio (privato) Prize (private)
COMMISSARIATO (SERVIZIO -) COMMISSARIAT (SERVICE -)	→	Servizio di commissariato Commissariat service UFHV Commissariato (Servizio -) Commissariat (Service -)
BANDA MUSICALE (ALTRE ARMI E FF.AA.) MARCHING BAND (OTHER ARMS AND ARMED FORCES)	→	Bande musicali Marching bands UFHV Banda musicale (altre armi e FF.AA.) Marching band (other arms and Armed Forces) NT Bande musicali di altre armi Marching bands of other arms NT Bande musicali delle Forze Armate Marching bands of the Armed Forces

Source(s): Image by authors

Figure 6.
Examples of “+” sign
convention
management

STAZIONE + STATION +	Stazioni Stations UFHV Stazione + Station + NT Stazione di Salerno Stations of Salerno
REPARTO + DEPARTMENT +	Reparti Departements Cataloguig notes: specify the department name in addition UFHV Reparto + NT Reparti speciali Special departments NT Reparto Servizi Presidenziali Presidential Services Department

Source(s): Image by authors



Source(s): Image by authors

Figure 7.
Examples of cross-
references "v."

its source to clarify its meaning within the CCThes; the sources are the same used for terminological control purposes.

- Historical notes, which indicate that the term chosen as PT in the CCThes had a different form, meaning or use in the CC Soggettario. They may contain various information: forms of the new PT which differ from those replaced or modified; if the PT was part of a compound term in the CC Soggettario and has been splitted in the CCThes; other historical information on the evolution and use of the term and concept over time within the reference domain.
 - Cataloguing notes, which support the choice and use of the most appropriate term to describe the subject of the document, by giving suggestions or instructions.
- (2) Additional terms and relationships coming from the tools used for terminological control and agreed with the representatives of the Historical Office of the General Command of the Carabinieri Corps. This is the case, for example, of the introduction of the full form of the term as a synonym of its corresponding acronym already included in the CC Soggettario or of the introduction of a more generic term to complete the description of a concept, integrate the semantic structure and reduce semantic ambiguity.

New terms and relationships are examined and accepted by domain experts responsible for the CCThes validation. Moreover, to ensure greater consistency and effectiveness, general concepts have gradually been identified and included in the hierarchy to serve as Top Terms (TT).

Such a resulting structure gives the tool a greater expressive power, since the system of notes, new terms and further relationships allows for the completion of classification tasks, by covering missing hierarchical levels, the contextualisation of terms and the enrichment and clarification of their meaning.

4. Results

The primary result of this study is a thesaurus available in RDF format (Skos-Core), which is interoperable with other vocabularies or applications and compliant with ISO 29964-1:2011 recommendations, and which meets the need to improve information indexing and retrieval performances. The main role of a thesaurus, in fact, is to help indexers in general comprehension of the subject area, by outlining inter-relationships between concepts and

providing definitions of terms (Aitchison *et al.*, 2000). The use of a thesaurus to index specific collections improves the quality of information retrieval in a particular knowledge domain.

The CCThes supports the subject indexing, search and retrieval of the various types of resources preserved at the Historical Office of the General Command of the Carabinieri Corps. It also allows for the harmonisation of concepts and terms, in order not to generate interpretative ambiguities in their use within the reference domain. Terms are organised in a way that the relationships between them are made explicit and the PTs are linked to any synonyms. Therefore, the study enhances the existing subject heading list, the CC Soggettario and reuses it to fulfil wider functions and purposes guaranteeing benefits in terms of cost and time reduction (Cardillo *et al.*, 2014).

The CCThes generally has a mono-hierarchical structure, i.e. the terms always have only one BT, with the possibility of admitting, as an exception, the poly-hierarchy for a few terms, i.e. the terms may have more than one broader term. It is recommended that each term within a thesaurus has at least one semantic relationship. Some terms without hierarchical relationships are temporarily present in the CCThes to maintain a link with the CC Soggettario and not to lose the terms deriving from it. Temporarily, these concepts are related to other concepts or terms by means of an associative or an equivalence relationship. Through the continuous enrichment of the CCThes, made by its application during experimental activities, all terms can be completely structured. This entails that every time a document needs to be indexed through a new term, this latter will become part of the CCThes, enriching the consistency of the semantic network.

Regarding the usability, the CCThes offers the possibility to browse the term list using: the alphabetical display, which provides the list of PTs and, if present, the equivalence relationships with the corresponding non-preferred ones, typographically distinguished from preferred ones by means of the italics graphic convention; and the systematic display, which shows the structure of the CCThes with all levels of hierarchy whereas the equivalence and associative relationships are not displayed.

The CCThes can be consulted and queried online, by using the credentials issued by the person in charge or his delegate at the Historical Office of the General Command of the Carabinieri Corps, and, also, offers the possibility to search by single term in the following ways:

- (1) Search by term, which returns a list of terms that contains the word used to execute the search, even in compound form. By selecting one of them, it is possible to visualise the systematised presentation of each concept, together with its relationships and notes. The semantic structure visualisation allows to further facilitate both the indexing and the searching phases.
- (2) Advanced search, which allows for the specification of additional search criteria. The most useful parameters for indexing include:
 - The search by words and combinations of words, in the entire list of terms of the CCThes or only in the notes. It is also possible to choose to show NPTs among the results.
 - The search by terms for which a particular type of note is present in the terminological record.

The CCThes consists of a total of 2,133 terms and includes the main typical elements of this kind of controlled vocabularies. Table 1 shows information about the number of terms, semantic relationships and notes contained therein.

Figure 9 shows the distribution of PTs and NPTs within the total amount of terms. Over 70% of terms in the CCThes are PTs, the remaining percentage indicates NPTs.

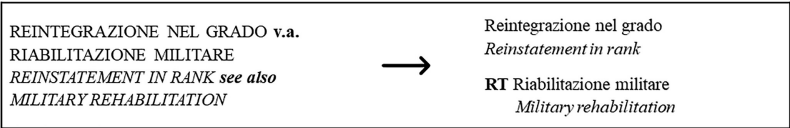
The semantic network in the CCThes presents the following distribution of associative relationships compared to hierarchical ones as shown in (Figure 10).

Concerning the global organisation of the vocabulary from the semantic point of view, Figure 11 shows the number of terms for each depth level in the semantic hierarchy. As a result, we have a quite homogeneous distribution with the majority of PTs having two levels of depth in the hierarchy.

Lastly, an introductory user guide, as a complementary tool to the CCThes has been created. It completes the CCThes (Aitchison *et al.*, 2000) and it is an element of connection and continuity between the previously used semantic subject indexing tool and the new one. Its

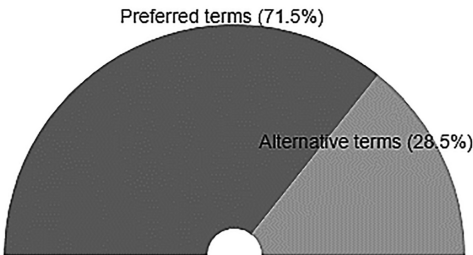
Preferred terms	1,525
Non-preferred terms	608
Hierarchical relationships	547
Associative relationships	1,398
Scope notes	37
Cataloguing notes	44
Historical notes	7

Table 1.
Number of terms,
relationships and notes
of the CCThes



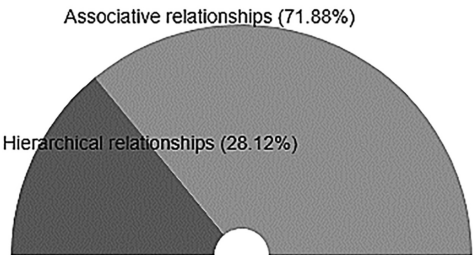
Source(s): Image by authors

Figure 8.
Example of cross-
references “v.a.”



Source(s): Image by authors

Figure 9.
Percentage of PTs
and PTs



Source(s): Image by authors

Figure 10.
Distribution of
associative and
hierarchical
relationships

function is to illustrate the main methodological choices made in the CCThes design and implementation, and to provide contextual information and examples useful for facilitating the choice of the most suitable indexing terms.

5. Discussion and conclusion

The evolution of the CC Soggettario towards the CCThes arises from a practical need, that is, to improve the description of the documentary heritage held and preserved at the archive and museum of the Historical Office of the General Command of the Carabinieri Corps. The aim is to support conceptual analysis and suitable index term identification and assignment, to favour document discoverability and retrieval. The CCThes can also benefit from all the potential advantages offered by digital technologies and environments. Those contributions include, for example: the enrichment of the structure functionalities based on the possibility to establish hyperlinks among all the structural elements (descriptors, non-descriptors, scope notes, etc.); the cost reduction of updating and maintenance, due to digitalisation of the thesaurus construction and publishing phases; the possibility of applying methods of reuse and interoperability at the time of planning and creating new tools (Arano, 2005).

The choice not to create the CCThes from scratch is based on the consideration that the conversion process of an available tool, with an already validated structure, can reduce the dependency on expert contribution, only involved in doubt solution and validation phase, and save resource, time and cost of the thesaurus construction. However, the process of conversion is more complex from the scientific standpoint, to adapt existing resources and to find rules and conventions to be used to facilitate the activity of the researcher and to improve efficiency. As a result, adopting and integrating existing resources with new terms and semantic relationships allow to increase the CCThes in semantic expressiveness and richness (Hilera et al., 2010).

The thesaurus has been chosen among other KOSs, since it has a clear semantic structure based on relationships and concept classification, which can facilitate consistency achievement in indexing and ensure that indexers and users choose the same terms in both resource indexing and searching phases (Vlachidis, 2020). Furthermore, its semantic structure is in line, having the same basis of knowledge classification and organisation, with ontology structure. It, also, enables the CCThes further development in a domain ontology, since its vocabulary can be transformed into the ontology classes or concepts, the relationships among terms can be converted into the relationship among the ontology classes, on basis of which, the domain ontology elements and relationships can be built (Ping and Yong, 2014).

As a future perspective, the CCThes could be further improved through its integration within a web-based document management platform that allows for archival heritage preservation and enhancement, by means of the organisation and indexing the documents,

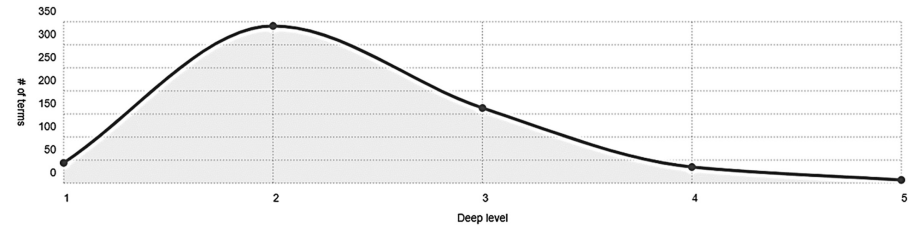


Figure 11.
Number of terms for
each depth level of
hierarchy

Source(s): Image by authors

making them searchable and accessible. In view of the possible future computerised and online management of documentation, it could be also possible to benefit from the advantages offered by the adoption of structured and controlled vocabularies, since in such a digital and collaborative scenario, they allow for fast, efficient and consistent indexing and searching processes. The CCThes could make available automated, parameterised and extendable search and it allows searching and indexing through mutually connected descriptors (Carlucci, 2002).

Notes

1. Carabinieri Corps' official website: <https://www.carabinieri.it/home>
2. Library of Congress Subject Headings PDF Files: <https://www.loc.gov/aba/publications/FreelCSH/freelcsh.html>
3. Official NATO Terminology Database: <https://nso.nato.int/natoterm/content/nato/pages/home.html?lg=en>
4. Tematres vocabulary server: <https://www.vocabularyserver.com/>
5. Nuovo Soggettario: <https://thes.bncf.firenze.sbn.it/ricerca.php>
6. Agrovoc multilingual thesaurus: <https://agrovoc.fao.org/browse/agrovoc/en/?clang=en>
7. EuroVoc multilingual and multidisciplinary thesaurus: <https://eur-lex.europa.eu/browse/eurovoc.html?locale=it>
8. Treccani online dictionary and encyclopaedia: <http://www.treccani.it/>
9. Wikipedia: <https://it.wikipedia.org>
10. Carabinieri Corps' official website – Curiosità section: <https://www.carabinieri.it/arma/curiosita/non-tutti-sanno-che/non-tutti-sanno-che>

References

- Aitchison, J., Gilchrist, A. and Bawden, D. (2000), *Thesaurus Construction and Use: A Practical Manual*, 4th ed., Europa Publications, London.
- Anderson, C., Bjorkman, B., Denis, D., Doner, J., Grant, M., Sanders, N. and Taniguchi, A. (2022), *Essentials of Linguistics*, 2nd ed., eCampusOntario, Toronto.
- Arano, S. (2005), "Thesauruses and ontologies", *Hipertext.net*, No. 3, available at: <http://eprints.rclis.org/8972/2/12.pdf> (accessed 27 December 2023).
- Barros, T.H.B. and Hjørland, B. (2024), "Domain analysis versus facet analysis", *Knowledge Organization*, Vol. 51 No. 1, pp. 19-26.
- Biblioteca Nazionale Centrale di Firenze (2021), *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto*, 2nd ed., Associazione Italiana Biblioteche, Roma, Biblioteca nazionale centrale di Firenze, Firenze.
- Bowman, M., Lopez, A.M. and Tecuci, G. (2001), "Ontology development for military applications", *Proceedings of the SouthEastern Regional ACM Conference*, Atlanta, GA, March 16-17, 2001.
- Broughton, V. (2006), *Essential Thesaurus Construction*, Facet, London.
- Carbone, F. (2017), *Tra carte e caserme. Gli archivi dei Carabinieri Reali (1861-1946)*, Ministero della difesa, Ufficio storico del V Reparto dello Stato maggiore della difesa, Rome.
- Cardillo, E., Folino, A., Trunfio, R. and Guarasci, R. (2014), "Towards the reuse of standardized thesauri into ontologies", in de Boer, V., Gangemi, A., Janowicz, K. and Lawrynowicz, A. (Ed.s.), *WOP 2014 5th Workshop on Ontology and Semantic Web Patterns Co-located with ISWC2014 Riva del Garda*, Italy, pp. 26-37.

- Carlucci, C. (2002), *I termini della formazione: il controllo terminologico come strumento per la ricerca*, Isfol, Rome.
- Chatterjee, A. (2017), Controlled Vocabulary C.V., in Chatterjee, A. (Ed), *Elements of Information Organization and Dissemination*, Elsevier, pp. 151-169.
- Cutter, C. (1904), *Rules for a Dictionary Catalog*, Government Printing Office, Washington.
- De Keyser, P. (2012), *Indexing from Thesauri to the Semantic Web*, Chandos Publishing, Oxford.
- Department of Defense (2016), "Department of defense dictionary of military and associated terms", available at: https://irp.fas.org/doddir/dod/jp1_02.pdf (accessed 27 December 2023).
- Dextre Clarke, S.G. (2019), "Thesaurus for information retrieval", *Knowledge Organization*, Vol. 46 No. 6, pp. 439-459.
- Dykstra, M. (1988), "LC subject headings disguised as a thesaurus", *Library Journal*, Vol. 113 No. 4, pp. 42-46.
- Faber, P. and Haraüz, P.L. (2019), "Frame-based terminology applied to military science: transforming a glossary into a knowledge resource", *Lexicography*, Vol. 6 No. 1, pp. 105-131, doi: [10.1007/s40607-019-00060-y](https://doi.org/10.1007/s40607-019-00060-y).
- Hilera, J.R., Pagés, C., Martínez, J.J., Gutiérrez, J.A. and de-Marcos, L. (2010), "An evolutive process to convert glossaries into ontologies", *Information Technology and Libraries*, Vol. 29 No. 4, pp. 195-204, doi: [10.6017/ital.v29i4.3130](https://doi.org/10.6017/ital.v29i4.3130).
- International Federation of Library Associations and Institutions (2009), *Guidelines for Multilingual Thesauri*, IFLA, The Hague.
- International Organization for Standardization (2011), *ISO 25964-1 Thesauri and Interoperability with Other Vocabularies. Part 1: Thesauri for Information Retrieval*, ISO, Geneva.
- Martínez-González, M. and Luisa, A.D.M. (2019), "The support of constructs in thesaurus tools from a Semantic Web perspective: framework to assess standard conformance", *Computer Standards and Interfaces*, Vol. 65, pp. 79-91, doi: [10.1016/j.csi.2019.02.003](https://doi.org/10.1016/j.csi.2019.02.003).
- Mazzocchi, F. (2018), "Knowledge organization system (KOS)", *Knowledge Organization*, Vol. 45 No. 1, pp. 54-78, doi: [10.5771/0943-7444-2018-1-54](https://doi.org/10.5771/0943-7444-2018-1-54).
- Ministero della Difesa (2010), "Decreto Legislativo 15 marzo 2010 n. 66", *Gazzetta Ufficiale Serie Generale*, n.106 del 08-05-2010 – Suppl. Ordinario n. 84, available at: <https://www.gazzettaufficiale.it/eli/gu/2010/05/08/106/so/84/sg/pdf> (accessed 27 December 2023).
- North Atlantic Treaty Organization (2021), *NATO Glossary of terms and definitions (English and French)*, AAP-06 – Ed: 2021, NATO Standardization Office, Brussels, Belgium.
- Ping, L. and Yong, L. (2014), "On transformation from the thesaurus into domain ontology", *Proceedings of the 2012 2nd International Conference on Computer and Information Application (ICCIA 2012)*, Atlantis Press, Paris.
- Presidenza del Consiglio dei Ministri (2019), "Glossario intelligence – Il linguaggio degli Organismi informativi", available at: <https://www.sicurezza.gov.it/sisr.nsf/wp-content/uploads/2019/06/glossario-intelligence-2019.pdf> (accessed 27 December 2023).
- Rout, R. (2018), "Enhancement of subject searching in OPAC using ontology", *Revista Espacios*, Vol. 39 No. 44, pp. 1-15.
- Soergel, D. (2009), "Knowledge organization system development", available at: <https://www.dsoergel.com/UBLIS514DS-08.2a-1Reading4SoergelKOSOverview.pdf> (accessed 27 December 2023).
- Spinelli, S. (2006), "Introduzione all'indicizzazione", available at: <http://web.dfc.unibo.it/buzzetti/IUcorso2008-09/mdidattici/spinelli-indicizzazione.htm> (accessed 27 December 2023).
- Stato Maggiore della Difesa (2009), "Glossario nazionale dei termini e delle definizioni", available at: https://www.difesa.it/SMD_/Staff/Sottocapo/UGID/Dottrina/Pagine/Glossario_termini_definizioni.aspx (accessed 27 December 2023).

Vlachidis, A. (2020), *Discuss Similarities and Differences of Ontologies and Thesauri. Their Role to Modern Information Retrieval Systems*, available at: <https://www.webronika.com/wp-content/uploads/2021/03/19075478-INST0038-.pdf> (accessed 27 December 2023).

Wielinga, B.J., Schreiber, A.Th., Wielemaker, J. and Sandberg, J.A.C. (2001), "From thesaurus to ontology", *Proceedings of the 1st International Conference on Knowledge Capture (K-CAP '01)*, Association for Computing Machinery, New York, NY, pp. 194-201.

Corresponding author

Maria Teresa Guaglianone can be contacted at: mariateresa.guaglianone@cnr.it

For instructions on how to order reprints of this article, please visit our website:

www.emeraldgroupublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com