Vocabularies and Retrieval Tools in Biomedicine: Disentangling the Terminological Knot

Klaar Vanopstal · Robert Vander Stichele · Godelieve Laureys · Joost Buysschaert

Received: 21 October 2008 / Accepted: 6 October 2009 © Springer Science + Business Media, LLC 2009

Abstract Terms like “thesaurus”, “taxonomy”, “classification”, “glossary”, “ontology” and “controlled vocabulary” can be used in diverse contexts, causing confusion and vagueness about their denotation. Is a thesaurus a tool to enrich a writer’s style or an indexing tool used in bibliographic retrieval? Or can it be both? A literature study was to clear the confusion, but rather than giving us consensus definitions, it provided us with conflicting descriptions. We classified these definitions into three domains: linguistics, knowledge management and bibliographic retrieval. The scope of the terms is therefore highly dependent on the context. We propose one definition per term, per context. In addition to this intra-conceptual confusion, there is also inter-conceptual vagueness. This leads to the introduction of misnomers, like “ontology” in the Gene Ontology. We examined some important (bio)medical systems for their compatibility with the definitions proposed in the first part of this paper. To conclude, an overview of these systems and their classification into the three domains is given.

Keywords Information retrieval · Medical terminology · Medical coding systems · Taxonomy · Thesaurus · Ontology · Controlled vocabulary · Classification

Introduction

Terms such as thesaurus, taxonomy, ontology and controlled vocabulary, and even glossary, dictionary and lexicon at first sight seem to be unambiguous terms. However, they are used in different ways in different contexts, causing continual confusion. Moreover, the distinction between the terms themselves is not always straightforward.

A look at the information about the term ‘death’ in three different thesauri (see Table 1), tells us that not all thesauri give the same kind of information:

The Unesco Thesaurus [1] includes information such as narrower terms (NT), broader terms (BT), related terms (RT), other language equivalents (SP, FR) and related terms (RT). In Roget’s Thesaurus [2], by contrast, other information is given: function, derivations, and related terms. The ICPC2-ICD10 thesaurus, a system used for medical classification which links concepts of ICPC2 to ICD-10 concepts, gives the classification codes R99 (ICD-10) and A96 (ICPC) for ‘death’. It is clear that these thesauri differ considerably in their structure and scope. Does this mean that for one of them, the denomination “thesaurus” is not—or less—apt?
The main problem is that terms such as “taxonomy”, “thesaurus”, “ontology” and “controlled vocabulary” are used in many different contexts, including linguistics, bibliographic information retrieval (IR) and knowledge management, including medical coding. As Kagolovsky and Moehr [3] point out, “information retrieval” has no common definition, due to the different research backgrounds of the authors who use the term. Kagolovsky and Moehr propose the following definition, citing Harter and Hert [4]: a system that “retrieves documents, or references to them, rather than data”. This definition corresponds to what we will call in this paper bibliographic retrieval. Medical registration systems, on the other hand, are established in the first place to represent and store information—rather than documents—and in the second place to later retrieve and re-use that information.

The first section of this paper gives an overview of the different fields in which the terms “glossary”, “lexicon”, “dictionary”, “taxonomy”, “classification”, “thesaurus”, “ontology” and “controlled vocabulary” can be used. On the basis of these observations, definitions will be suggested and recommendations made for a more consistent and unambiguous use of the relevant terminology. In the second section, these insights will be applied to the biomedical domain, where these issues are particularly relevant. To conclude, an overview of the existing tools in the three dimensions (linguistics, knowledge management—including medical coding—and bibliographic retrieval) is presented.

### Domains of application of the terms

As mentioned above, terms such as taxonomy, thesaurus, ontology, controlled vocabulary etc. can be defined in various ways depending on the domain of application. We will discuss three domains, namely linguistics, knowledge management—including medical coding systems—and bibliographic retrieval.

There are several linguistic tools which can help to find the right terms, or to find an explanation or definition for a certain term, viz. dictionaries, lexicons, glossaries, thesauri and controlled vocabularies. These systems (can) have a purely linguistic function. However, thesauri and controlled vocabularies can also be used for the retrieval of documents or data.

A second domain which will be discussed here, is that of the storage and retrieval of knowledge. We especially focus on medical coding systems, such as ICPC and ICD. Medical coding systems can be described as classifications or nomenclatures of health- and medicine-related phenomena. These concepts are structured and usually given a code which indicates the place of the concept in the nomenclature, as can be seen in Fig. 1.

Bibliographic retrieval can be defined as the science of searching a database for journal or magazine articles, containing citations, abstracts and often full texts or links to the full texts. The underlying structures to search for articles in databases include taxonomies,
(K35-K38) Diseases of appendix

- (K35.) Acute appendicitis
  - (K35.0) Acute appendicitis with generalized peritonitis
  - (K35.1) Acute appendicitis with pylephlebitis
  - (K35.2) Acute appendicitis, unspecified

- (K36.) Other appendicitis

- (K37.) Unspecified appendicitis

- (K38.) Other diseases of appendix
  - (K38.0) Hyperplasia of appendix
  - (K38.1) Appendicular congestion
  - Paracolitis
  - Stercolith
  - (K38.2) Diverticulum of appendix
  - (K38.3) Fistula of appendix
  - (K38.4) Other specified diseases of appendix
    - Intussusception of appendix
  - (K38.9) Disease of appendix, unspecified

Fig. 1 Extract of the ICD10 classification: “diseases of appendix”

The term ‘glossary’ originates from the Latin word *glossarium*, a collection of glosses. ‘Gloss’, in its turn, originates from the Greek word *glossa* (γλώσσα) which denotes the explanation of a specialized expression or difficult word. Hence, ‘glossary’ can be defined as a list of terms in a particular field of knowledge, with definitions or explanations.

Glossaries are usually arranged alphabetically. The terms in monolingual glossaries usually refer to LSP (Language for Specific Purposes) and are furnished with definitions. These definitions generally apply to one domain only, and thus rarely include variant meanings. In practice, however, these definitions are often omitted in multilingual glossaries.

Glossaries can be integrated into a book or a website, but they can also be stand-alone lists. They can be used as, but are not, per se, controlled vocabularies (see infra). They can be monolingual (e.g. Wikipedia’s *Glossary of medical terms related to communications disorders*¹ or the Dutch RIZIV glossary²), bilingual (e.g. the TERMISTI glossaries of abortion³ and autism⁴ terms) or multilingual (e.g. Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages⁵). Definitions are often omitted in bilingual and multilingual glossaries.

The term glossary is used interchangeably with lexicon and dictionary. This presumed equivalence, however, leads to a blurring of the conceptual boundaries of the terms. Ananiadou [⁵] defines ‘lexicon’ as a list containing “the lexical elements (either as full forms or as canonical base forms), together with additional linguistic information about them, which is required for further morphological, syntactic, and semantic processing.” She adds that lexicons are not fully standardized, which allows its makers to model them so that they best suit their own purposes. We will adopt Ananiadou’s definition.

Dictionaries, both monolingual and multilingual, can refer to general language or to a specialized terminology. They often give limited morphological and grammatical information (e.g. gender, part of speech, plural form) and sometimes also a phonetic transcription, next to a definition. Bi- and multilingual general language dictionaries usually provide a translation—or several translations used in different contexts—, collocations and idiomatic expressions. Conversely, specialized multilingual dictionaries usually offer translations, and very little further information. An example from the *Wörterbuch für Industrie und Technik* (French–English/English–French) [⁶]:

Reprofilage n.m. Neuprofilierung n.f. Bâtiments et travaux publics⁶

Summarizing, it can be said that the boundaries between the terms glossary, lexicon and dictionary have blurred to some extent. However, we define ‘glossary’ as “a list of words or terms with their explanations”, ‘lexicon’ as “a list of words or terms, together with linguistic information about them” and ‘dictionary’ as “a list of words or terms with limited linguistic information, usually a definition, and, in the case of bi- or multilingual dictionaries, one or more translations”.

**Thesauri**

The word ‘thesaurus’ is derived from the ancient Greek ‘thesauros’ (θησαυρός), or ‘treasure’. In the sixteenth century, its meaning was narrowed to ‘treasure of words’, like a dictionary or an encyclopaedia. The word ‘thesaurus’ fell into disuse for some time, but revived with the release of Roget’s *Thesaurus of English Words and Phrases* in the nineteenth century. Roget adopted an onomasiological approach—providing the word for a given idea—in his

---

⁶ The first column refers to the French term, the second to the German translation and the third column refers to the corresponding domain.
thesaurus, whereas most dictionaries were, and still are, characterised by a semasiological approach, i.e. they describe the referential meaning denoted by words. Roget did not organize his thesaurus alphabetically, but systematically, i.e. according to ideas or concepts.

The purpose of an ordinary dictionary is simply to explain the meaning of words; and the problem of which it professes to furnish the solution may be stated thus:—The word being given, to find its signification, or the idea it is intended to convey. The object aimed at in the present undertaking is exactly the converse of this: namely,—The idea being given, to find the word, or words, by which that idea may be most fitly and aptly expressed [7].

A thesaurus can thus be a purely linguistic tool, which provides a standard language of a particular field of knowledge and contains information about nuances of concepts. This type of thesaurus is referred to by Kilgarriff and Yallop [8] as the ‘Roget-style thesaurus’. Its objective is to improve the effectiveness of communication: the relationships outlined in the thesaurus help to fine-tune style or to obviate misunderstandings.

Later, in the mid-twentieth century, the term experienced another shift in meaning, adopting the information retrieval aspect (see infra).

Controlled vocabulary

A controlled vocabulary is a set of terms which provides a standard language for a specific domain. It consists of two types of terms: preferred terms, which are designed to control a domain-specific language, and non-preferred terms used as “access vocabulary”, “lead-in” or “entry” terms. The use of preferred and non-preferred terms is illustrated by Wodtke [9]:

In our restaurant we had the preferred term, “first course”, and all the terms our patron might use, “starter, first course, hors d’oeuvres, appetizer”, neatly tucked into our head. So if a patron wanted an appetizer of smoked salmon, we would write in the check “first course: smoked salmon”.

A controlled vocabulary can be used as a prescriptive terminology, as a means to ensure language hygiene and/or consistency in the use of terminology. The Plain English Campaign⁷ is an independent British organization which helps businesses, local governments and government departments to improve their communication by providing editing services, training courses and glossaries. They also published a controlled vocabulary, The A to Z of alternative words, which is a list of words with their simpler alternatives designed for writers of all text types to ensure readability.

Knowledge management and medical coding

Taxonomies and classifications

A literature search for the term taxonomy proves that Garshol [10] is right in saying that the term has been “used and abused to the point that when something is referred to as a taxonomy it can be just about anything” and that the basic denominator is that of an “abstract [hierarchical] structure”.

Taxonomy is derived from the Greek words τάξις (tàkis), ‘order’ and νόμος (nómo), ‘rules, law’ and is often described in a very broad sense, as “the science of classification of organisms” [11]. However, the term taxonomy can also be defined in terms of its structural characteristics: “a taxonomy provides a classification structure that adds the power of inheritance of meaning from generalized taxa to specialized taxa” [12]. This inheritance implies that subclasses take over characteristics of their ancestor classes. Agro [13] and Beck [14] also use the term in the sense of a hierarchical structure which represents (a part of) reality. Dictionaries such as Oxford English Dictionary and Merriam-Webster and other reference works such as WordNet and Roget’s Thesaurus differentiate between the two meanings, i.e. taxonomy as a science and taxonomy as a hierarchical representation of reality. Sterkenburg [15] combines both meanings in his definition: “study of the theory, practice and rules of classification of terms, objects and concepts”.

The term taxonomy originated in biology, where it referred to the classification of the names of organisms. It was the Swedish scholar Carolus Linnaeus who combined the loose principles of the existing taxonomies into the ‘Linnaean taxonomy’ (Systema Naturae 1735). In this hierarchical classification, nature was divided into kingdoms, phyla (for animals) and divisions (for plants), classes, orders, families, genera and species. In the figure below (Fig. 2), modern humans (homo sapiens) are defined according to the Linnaean taxonomy.

Linnaeus’ taxonomy, which is now called the alpha taxonomy or classical taxonomy is still a model for biological classifications.

The designations “taxonomy” and “classification” are used interchangeably, whereas they are not completely synonymous. Agro [13] and Van Rees [16] argue that taxonomies distinguish themselves from classifications in that they group concepts according to essential, internal attributes, i.e. according to relationships between the

---

⁷ http://www.plainenglish.co.uk/
concepts. Taxonomies, unlike classifications, are created from the bottom up, are based on actual content and guide users through a body of information. A classification, on the other hand, is a grouping of concepts according to arbitrary, external attributes. These external attributes can be colour, shape, geography, size, usability, etc. Classifications are created from the top down and are based on conceptual frameworks. Table 2 summarizes the characteristics of taxonomies versus classifications according to Agro and Van Rees.

Cann, however, uses other criteria to define the concepts of classification and taxonomy. He describes special versus general, analytical versus documentary and enumerative versus faceted classifications. Firstly, a classification describes either general knowledge, e.g. the Universal Decimal Classification (UDC) or a specific knowledge domain, e.g. the International Classification of Diseases (ICD). Secondly, a classification can be analytical or documentary. Analytical implies that physical phenomena are systematized into an understandable scheme. Cann also designates this type of classification as “taxonomies”. In his opinion, “taxonomy” and “classification” are not, as argued by Agro and Van Rees, co-hyponyms, rather “taxonomy” is hyponymous to “classification”, or a taxonomy is a “kind of” classification. Documentary classifications are used as information management and retrieval tools (e.g. UDC). Thirdly, classifications can be either enumerative or faceted. An enumerative classification lists certain classes and all their subclasses of interest, is created from the top down and allows for compound subjects. This type of classification is often called hierarchical, which is a common misunderstanding, as faceted classifications can also have a hierarchical structure. Faceted classifications are created from the bottom up and do not provide “ready-made class numbers for compound and complex subjects”. In enumerative classifications, there is usually only one path the user can follow to find his subject, i.e. from a broad category to the specific concept. In faceted classifications, the concepts are organized into classes according to several principles of division. An example of a faceted classification can be found in Springerlink’s organization of documents, where documents can be retrieved using different principles of division. The collection can be searched by the facets “content type”, “featured library” or “subject collection”.

Cann’s view (see Fig. 3) seems to be more solid and logical. Here, a classification is considered as a hyponym for all types of concept categorization. However, Cann still overlooks the fact that analytical classifications, or taxonomies, have also come to play a role in information retrieval, i.e. they have adopted the function of documentary classifications.

We propose a definition for “taxonomy” in data retrieval, based on ISO/IEC 11179-2: “a taxonomy provides a hierarchical classification structure that adds the power of inheritance of meaning from generalized taxa to specialized taxa”. Classification is a more general term which can be defined as the grouping together of concepts on the basis of shared characteristics. Both structures can be used in medical coding systems.

Ontologies

A closer look at the concept of ‘ontology’ shows that, again, its meaning depends on the domain or the (historical) context in which it is used: philosophy or information science. When used in the context of philosophy, Ontology is often written with an upper-case ‘O’, whereas ontology with a lower-case ‘o’—and with a plural form, ontologies—refers to a representation of reality or to an information retrieval system.

The term ‘Ontology’ is derived from the Greek words ὄν (being) and ὂν (science, study, theory) and literally translates into “the science of being”. This branch of metaphysics organizes, or attempts to organize the universe
and its components into a scheme with explicit formulation of their possible relations. Most dictionaries, such as LONGMAN Dictionary of Contemporary English [19], Oxford English Dictionary [20] and Merriam-Webster [21] define Ontology in this context. As a derived meaning used within the context of knowledge management, an ‘ontology’ can be described as a representation of what exists. Some ontologies, like SNOMED or OpenGalen, are more than just a representation of the concepts within a specific domain with their relationships; they are designed to use as a coding system or for clinical decision support.

Bibliographic retrieval

Taxonomies

With the advent of the Internet, taxonomies started covering other purposes than those described above: they now also function as metadata for information retrieval. These taxonomies are no longer used to retrieve only data, rather the concepts are used as keywords for tagging documents, or for referencing to these documents. Cann [17] refers to this type of taxonomy as “documentary classifications” (see supra). Their structure offers more transparent and more efficient search options, including explosion of the search term. Term explosion allows the system to search for information about not only the concept itself, but also about its narrower, hyponymic concepts.

Taxonomies can be included in thesauri and ontologies [14, 22], and taxonomies and thesauri are often bracketed together as one and the same concept. So what distinguishes the taxonomy from thesauri and ontologies? Basically, ‘taxonomy’ can refer to any hierarchical classification of elements of a group into subgroups according to specific criteria, often visualized as a tree. Its relationships are not specified, i.e. broader and narrower terms can designate the obvious subsumption relationship (parent/child), but also a mereologic relationship (part/whole). Taxonomies do not cover any relationships other than hierarchical. Thesauri and ontologies compensate for this lacuna and give explicit or implicit indications as to the nature of the relationship.

Thesauri

Peter Luhn (IBM) conceived the idea of using a thesaurus, which was previously a purely linguistic tool, for information retrieval. In the 1960s, the first thesauri for information retrieval were published. The Thesaurus of Engineering and Scientific Terms [23] sketched the broad outlines of the standard format for thesauri. In this period, thesauri evolved towards their current form, defined by ISO 2788 [24] as “the vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (for example as “broader” and “nearer”) are made explicit.” Controlled means that the vocabulary is predetermined and is used as a prescriptive terminology. This implies that the terminology of the subject field is subdivided into preferred terms—also called descriptors—and non-preferred terms or entry terms (see infra). A thesaurus is usually organized hierarchically, which means that the relationships ‘broader term’ and ‘narrower term’ are visible in a tree-like structure or made explicit by the abbreviations BT and NT respectively. ISO 2788 states that there are various ways in which the terms in a thesaurus can be displayed, the most common of which are alphabetical, systematic and graphic display. The standardized relationships in thesauri are the hierarchical, associative and the equivalence relationship. These are a priori relationships, which means that they are context-independent, rather than being inferred from the documents they describe.

When used in the context of information and library science, ‘thesaurus’ refers to a retrieval instrument, used to index and/or search documents. This is often the main or only purpose of present-day thesauri, and most authors [14, 24–30] define thesauri in this context. Chowdhury [28] describes the following main objectives of thesauri for information retrieval:

1. vocabulary control: a translation of natural language into a more constrained language
2. consistency between different indexers
3. limitation of the number of terms needed to label the documents
4. search aid in information retrieval

The historical and interdomain shifts—from the linguistic field to the field of information science—described above are reflected in the definitions given by Landau [31]:

1. A “storehouse” of knowledge such as exhaustive encyclopaedia or dictionaries,
2. Exhaustive lists of words from the general language, without definitions, arranged systematically according to the ideas they express.
3. A list of subject headings for a particular field of knowledge, arranged in alphabetic or classified order and used for information retrieval and related purposes.

Due to these shifts, the term ‘thesaurus’ carries several meanings, and it is thus recommendable to study the context and subject field in which the term occurs before drawing any conclusions as to its meaning.

There are several standards for thesauri. ISO 2788 was created for the design of monolingual thesauri and ISO 5964 [32] documents the design of multilingual thesauri. These standards, however, are outdated [33], as they only refer to printed thesauri. Both standards will be replaced by a new standard, ISO 25964, based on BS 87239 [27], the corresponding British standard. ANSI/NISO, the US standardization organization, created its own standard, Z39.19. These guidelines have a somewhat broader scope: they comprise all monolingual controlled vocabularies, including lists, taxonomies, thesauri and synonym rings. There is no single ‘worldwide’ standard, as the US and other standards (BS, ISO) departed from each other in previous editions. In an interview [34], Dr. Amy J. Warner10 stated that the new ANSI/NISO standard should be more compatible with the existing standards.

In conclusion, the term thesaurus can be used in different contexts, related to different fields of knowledge which came into existence at different points in time. When used in the context of information science, a thesaurus can be defined as a “controlled vocabulary, which is usually organized hierarchically and which includes standardized, a priori, hierarchical, associative and equivalence relationships between concepts” [24].

Controlled vocabularies

According to the ANSI/NISO Guidelines [26], a controlled vocabulary, which is a list of preferred and non-preferred terms, is—or should be—exempt of ambiguities, homonymy and polysemy and all terms should have “an unambiguous, non-redundant definition”. Controlled vocabularies can be used for consistent indexing and searching of information. For instance, using a controlled vocabulary in medical information retrieval can help health professionals to describe and classify medical information, optimizing the work of both searchers and indexers.

Compared to natural language, a controlled vocabulary has some weaknesses and some strengths, as stated by Aitchison et al. [25]. Its weaknesses include the relative lack of exhaustivity and specificity, the laboriousness of keeping it accurate and up-to-date and the cost of doing so. Moreover, this language has to be learned by the searcher and efficient exchange is often hampered by the incompatibility of the existing controlled vocabularies. Aitchison et al., however, add that over-exhaustivity may provoke a loss of precision. In addition, a controlled vocabulary can facilitate the search process considerably by expanding the query to its synonyms and excluding ambiguity. A controlled vocabulary is usually incorporated into a thesaurus, an ontology, a topic map, which, in turn, can be used in an information retrieval system.

Ontologies

In the late twentieth century, the term “ontology” adopted some new properties as it saw its introduction into information architecture and science. Most recent sources [14, 22, 26, 35–39] describe ontology in this field. Its best-known definition is that by Gruber [40]: “an explicit, formal specification of a shared conceptualisation”. An analysis of this definition is expedient, as it concentrates some important components. Firstly, ‘explicit’ means that the concepts included in the ontology are clearly defined, as are the constraints on their use. ‘Formal’ refers to the language of the ontology. A formal language is computer-readable: the computer ‘understands’ the relationships—also called ‘formal semantics’—within the ontology. This way, they can be used to support computer applications. Examples of formal representation languages for ontologies include RDF [41] (Resource Description Framework; cf. the Nautilus ontology [42]), F-Logic [43], or Frame Logic (e.g. FLORID [44]), KIF (Knowledge Interchange Format, e.g.), a later version of which—Common Logic—has been submitted to and approved by ISO, OIL [45] (Ontology Inference Layer), DAML+OIL, a combination of DAML (DARPA11 Agent Markup Language) and OIL, and OWL [46] (Web Ontology Language; e.g. Basic Clinical Ontology for breast cancer12), which combines OIL and DAML+OIL. Ontologies written in these formal languages can be used for inferencing or to support other software applications.

There are, however, ontologies which do not use a formal language, but a—restricted—natural language. Usserhold [47] argues that “we must rely on NL [natural language] definitions to be sure of what something means”.

The last components of the definition, shared and conceptualization, imply that this abstract model of phenomena in the world has been agreed upon by a group of users or experts.

---

9 The BS 8723 standard consists of five parts, the first two of which broadly correspond to ISO 2788, whereas the combination of part one and four have approximately the same scope as ISO 5964 (multilingual thesaurus). BS 8723–3 covers vocabularies other than thesauri, BS 8723–4 gives recommendations concerning interoperability of vocabularies and BS 8723–5 discusses exchange formats.

10 Project Leader for NISO’s Thesaurus Development Team.

11 DARPA stands for Defense Advanced Research Projects Agency.

12 http://acl.icnet.uk/~mw/MDM0.73.owl
As observed by Garshol [10], an ontology usually consists of concepts, relations and properties, but “exactly what is provided around this varies”. The basic elements of an ontology are concepts, grouped into classes. The actual object referred to by the concept, is an individual or instance. Relations between concepts and instances are often called roles. Attributes or properties are assigned to the concepts or instances.

Thesauri and taxonomies, and even glossaries are often considered bedfellows within the category of—simple—ontologies: they organize the concepts or terms of a knowledge domain, and all four can be used for indexing and searching information. An ontology, however, distinguishes itself from the other tools mainly by allowing more types of semantic relationships, which makes the ontology much more versatile, more powerful. In addition, an ontology usually structures its concepts not as a hierarchy, but as a network or a web.

Ontologies, as observed above, were initially conceived as a way to represent knowledge; they are “intended to support the vision of the semantic web through providing structured metadata about resources and a foundation for logical inferencing” [48]. They are aimed at giving a truthful reflection of reality, and this has repercussions on their further development for use in information retrieval.

In conclusion, the term ‘ontology’ is polysemous due to historical and interdomain shifts. Originally, it was the study of being, the outcome of which was a representation of what exists, or ‘an ontology’. This later became a schematic representation of fields of knowledge with concepts and their interrelationships. In information science, this structure is formalized and can be used for computer applications, including information indexing and retrieval.

Topic maps

Taxonomies, thesauri and ontologies were originally designed to represent knowledge. Later, especially with the advent of the Internet, they started being used as indexing vocabularies, facilitating information and document retrieval. Topic maps, on the other hand; were specifically designed for information indexing and retrieval and consist of a knowledge layer—comparable to an ontology—and a resources layer. The knowledge layer (called “topic space” in Fig. 4 [49]) is usually a semantic network deduced from the resources layer or pool and not—as an ontology—designed by experts as a representation of reality.

The distinction between ontologies and topic maps runs parallel to that between knowledge management and information management: ontologies cover only the knowledge itself, whereas topic maps also involve storing and tracking resources in which this knowledge may be found.

Fig. 4 Structure of topic maps

The idea of topic maps emerged in the early nineties when the Davenport Group met to discuss ways to merge indexes, glossaries, thesauri, cross references, etc. This new index was to reflect the structure of the knowledge it represented. Their efforts resulted in ‘topic navigation maps’, which were adopted as an ISO work item in 1996. In 2000, these topic navigation maps were renamed ‘topic maps’ and became a new ISO standard.13

Ontologies describe concepts—represented by terms—with their attributes and relationships and divide them into classes. These classes consist of concrete or abstract individuals or instances. Topic maps have subjects represented by topics and described by associations and occurrences. Topics are described in more detail by topic names and topic types, association types and occurrence roles (see also Pepper, 2000 [50]). In addition to this difference in structuring the knowledge layer, topic maps have some other important distinguishing characteristics, mainly concerning their development, initial purpose and

13 The definition of topic maps proposed in ISO/IEC 13250 is a circular definition, thus not helping to grasp the exact meaning of ‘topic maps’:
   a) A set of information resources regarded by a topic map application as a bounded object set whose hub document is a topic map document conforming to the SGML architecture defined by this International Standard.
   b) Any topic map document conforming to the SGML architecture defined by this International Standard, or the document element (topicmap) of such a document.
   c) The document element type (topicmap) of the topic map document architecture.
standards. The main differences and similarities are summarized in Table 3.

As observed above, the knowledge framework in ontologies is designed from scratch by a domain expert in order to support the vision of the semantic web. In topic maps, however, this knowledge layer is deduced from the documents or information contained in the resource layer. Pepper [50] and [51] Hummel [51] consider the separation into two layers and the standardized format respectively as the topic maps’ strengths. These qualities improve the navigational function of topic maps and their interoperability with other topic maps, and even with indexes, thesauri, taxonomies, ontologies and other traditional classification schemes. As confirmed by Garshol (2004) [10], "topic maps do not offer more, but other possibilities with regard to the knowledge represented, i.e. a flexible model with an open vocabulary”.

The format of topic maps is, as observed above, captured in an ISO standard, which also improves the efficiency and interoperability with other tools. Ontologies lack this standardization and are thus less suitable for exchange. The format of ontologies is not standardized, but many of their corresponding representation languages (XML, RDF, RDF Schema, and OWL) are.

Applications in the (bio)medical domain

The last decades have witnessed an information explosion in the (bio)medical domain, and with it the increasing need for solid vocabularies, terminologies and classification systems. They include—next to the numerous medical glossaries and dictionaries—the UMLS resources, the Gene Ontology, MeSH, SNOMED and OpenGALEN. The present section attempts to characterize these systems in terms of the definitions given above.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Differences between ontologies and topic maps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ontologies</strong></td>
<td><strong>Topic maps</strong></td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>An ontology is a representation of reality.</td>
</tr>
<tr>
<td><strong>Differences</strong></td>
<td>Is an organization of knowledge</td>
</tr>
<tr>
<td></td>
<td>Can be used as an information retrieval tool when the knowledge is linked to resources</td>
</tr>
<tr>
<td></td>
<td>Knowledge structure is designed by domain expert(s) and later linked to the documents or other resources</td>
</tr>
<tr>
<td></td>
<td>The knowledge layer is a representation of reality (within a specific domain)</td>
</tr>
<tr>
<td></td>
<td>The knowledge structure consists of concepts, classes, attributes, relations and individuals</td>
</tr>
<tr>
<td></td>
<td>Not a standardized format as such</td>
</tr>
</tbody>
</table>

Linguistic tools in the biomedical domain

Medical glossaries, lexicons and dictionaries

Wikipedia’s Glossary of medical terms related to communications disorders, the Ziekenhuis.nl woordenboek are examples of mono- and bilingual glossaries respectively. They cover terms from the field of medicine or social services, and comply with the definition of ‘glossary’ given in this article in that they are lists of terms, arranged alphabetically, with definitions.

The Specialist Lexicon, which is included the UMLS as one of the Knowledge Sources, meets the criteria for lexicons described in this article. It was designed for use in natural language processing (NLP) and is intended to be a general English lexicon that includes many biomedical terms. The linguistic information includes inflectional variants and derivations, acronyms, spelling variants and, when applicable, verb, noun or adjective complementation. An example of a lexical record can be seen in Fig. 5.

The Pinkhof geneeskundig woordenboek and the Dictionnaire d’infermeria are examples of a monolingual and a multilingual dictionary respectively. They give definitions and information on the origin of the word, which is generally Latin or Greek, and on gender.

The multilingual glossary of technical and popular medical terms in nine European languages

The Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages is a controlled vocabulary in the form of a glossary. Each ‘technical’ term in this glossary has a popular variant which should be considered as the preferred term in texts intended for patients. The glossary was initiated in the framework of the 92/27/EEC Directive, which made the inclusion of patient
information leaflets (PILs) in every medication package mandatory in the Member States of the European Community and stipulated that the leaflets had to be written in understandable language. As the use of terminology is often an important factor in the readability of these information leaflets, a glossary with popular variants for medical or technical terms was very useful. This controlled vocabulary was thus intended to help writers and translators make their PILs understandable for the general public.

The European multilingual thesaurus on health promotion

The European Multilingual Thesaurus on Health Promotion is a merger of three thesaurus projects in 12 languages and is used as a linguistic tool: it should stimulate the uniform use of terms related to health promotion and health education in Europe, as such a shared language supports the efficient exchange of information. This thesaurus is thus used as a controlled vocabulary, with preferred and non-preferred terms. The ISO standards 2788 and 5964 were used as construction guidelines, although it is not used for bibliographic retrieval.

Knowledge management and medical coding

The ATC classification

The ATC (Anatomical Therapeutic Chemical) classification is a system developed by the WHO for the classification of drugs and other medical products. Applying Cann’s view to this classification, one could state that this is a specific, documentary, enumerative classification. Specific, because it covers a part of the medical domain, namely medical substances. Documentary, because it functions as an information management and retrieval tool, and enumerative because it lists the classes and subclasses in a specific domain of interest and it is created from the top down.

The classification consists of 14 main classes, each one referring to another anatomical main group, e.g. nervous system (N). The next level is indicated by two digits and contains therapeutic subgroups, e.g. anti-parkinson drugs (N04). The third level, which is indicated by one letter, refers to the pharmacological subgroup, e.g. dopaminergic agents (N04B). The fourth level, again a letter, is a designation of the chemical subgroup, e.g. dopamine agonists (N04BC), and the last two digits indicate the chemical substance, e.g. pramipexole (N04BC05; see Table 4).

The ATC classification is mainly used to produce statistics about drug use, but also for the registration process of drugs.

The International Classification of Diseases and related health problems (ICD)

The International Classification of Diseases and Related Health Problems is published by the World Health Organization (WHO) and classifies diseases, signs, symptoms, complaints, social circumstances and causes of injury or disease. It is used in statistics, in automated decision support and in reimbursement systems. ICD-10, the tenth revision of ICD, is the most recent version of the classification. The first level of ICD-10 consists of 22 classes, each of which has several subclasses. The first letter in the code refers to the chapter, whereas the following digits specify the disease. For instance, in C18.7, C refers to malignant neoplasms, 18 refers to malignant neoplasms of the colon, and the numeric symbol after the decimal point further specifies the disease, in this case malignant neoplasm of the sigmoid colon.

ICD-10 is a specific, documentary and enumerative classification: it covers a specific domain, it is used to store and retrieve medical data and created from the top down.

The International Classification of Primary Care (ICPC)

The International Classification of Primary Care was designed by the WICC (WONCA International Classification Committee) for the classification of reasons for encounter (RFE), problems, diagnoses, interventions and the ordering of these data in an episode of care structure.

<table>
<thead>
<tr>
<th>ATC level</th>
<th>ATC code</th>
<th>ATC text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Anatomical main group</td>
<td>N</td>
<td>Nervous system</td>
</tr>
<tr>
<td>2 Therapeutic subgroup</td>
<td>N04</td>
<td>Anti-parkinson drugs</td>
</tr>
<tr>
<td>3 Pharmacological subgroup</td>
<td>N04B</td>
<td>Dopaminergic agents</td>
</tr>
<tr>
<td>4 Chemical subgroup</td>
<td>N04BC</td>
<td>Dopamine agonists</td>
</tr>
<tr>
<td>5 Chemical substance</td>
<td>N04BC05</td>
<td>Pramipexole</td>
</tr>
</tbody>
</table>
Chapter ten of the second version of ICPC has been converted into an electronic file, i.e. ICPC-2-E, is specifically designed for use in electronic patient records (EPR) and for research purposes. It is to be used together with the first nine chapters of ICPC-2. As ICD-10 is more fine-grained and allows for documentation at the level of individual patients [52], this classification was the perfect complement to ICPC-2. When ICD-10 was made available, together with its various translations, the WICC decided that all translations of ICPC were to relate to ICD-10, in order to allow for a better structuring of EPRs. For the Netherlands and the Dutch-speaking part of Belgium, this resulted in the ICPC-2/ICD-10 thesaurus (see infra).

ICPC-2 is a specific, documentary and enumerative classification which has a bi-axial structure (see Fig. 6). There are 17 main classes with an alpha code referring to the location of the complaint, and seven components with a two-digit numeric code, which organize each of these classes.

ICPC-2 is included in the UMLS (see infra).

ICPC-2/ICD-10 thesaurus

The ICPC-2/ICD-10 thesaurus is a continuation of the ICPC2/ICD10 Thesaurus, but with French translations added to it. The designation “thesaurus” is a misnomer in this case, as the system does not meet all the criteria described in the ISO standards for thesauri: it has no associative relationships. However, some terms do have synonyms or entry terms that lead the system to the correct concept. Like ATC, ICD and ICPC, this is a specific, enumerative, documentary classification used for medical coding.

Chapter ten of the second version of ICPC has been converted into an electronic file, i.e. ICPC-2-E, is specifically designed for use in electronic patient records (EPR) and for research purposes. It is to be used together with the first nine chapters of ICPC-2. As ICD-10 is more fine-grained and allows for documentation at the level of individual patients [52], this classification was the perfect complement to ICPC-2. When ICD-10 was made available, together with its various translations, the WICC decided that all translations of ICPC were to relate to ICD-10, in order to allow for a better structuring of EPRs. For the Netherlands and the Dutch-speaking part of Belgium, this resulted in the ICPC-2/ICD-10 thesaurus (see infra).

ICPC-2 is a specific, documentary and enumerative classification which has a bi-axial structure (see Fig. 6). There are 17 main classes with an alpha code referring to the location of the complaint, and seven components with a two-digit numeric code, which organize each of these classes.

ICPC-2 is included in the UMLS (see infra).

ICPC-2/ICD-10 thesaurus

The ICPC-2/ICD-10 thesaurus is a continuation of the ICPC2/ICD10 Thesaurus, but with French translations added to it. The designation “thesaurus” is a misnomer in this case, as the system does not meet all the criteria described in the ISO standards for thesauri: it has no associative relationships. However, some terms do have synonyms or entry terms that lead the system to the correct concept. Like ATC, ICD and ICPC, this is a specific, enumerative, documentary classification used for medical coding.
(‘fully specified name’), and a ‘description’ comprising one preferred term and one or more synonyms.

Two main types of relationships are established in this ontology: hierarchical and attribute relationships. Hierarchical ‘is-a’ relationships are defined within one axis, whereas the attributes link concepts from different hierarchies. Attribute relationships include finding site, causative agent, occurrence, stage, etc.

The prerequisites for an ontology in information science are thus fulfilled: the SNOMED CT terminology represents knowledge from a specific domain (health care), is concept-oriented, and the definitions are formalized. Moreover, almost any semantic relationship can be expressed in this ontology.

OpenGalen

OpenGALEN is a multilingual terminology and coding system for the classification of surgical procedures, electronic healthcare records (EHCRs), clinical user interfaces, decision support systems, knowledge access systems, and natural language processing.

The OpenGALEN Foundation (Open Galen Foundation s.d.) defines ‘ontology’ as “the set of primitive, high level categories in a knowledge representation scheme together with any taxonomy which structures those categories”. In this view, the OpenGALEN system is an ontology indeed. However, it also fulfills the requirements of an information retrieval ontology in the strict sense: it represents the concepts of a specific domain with formalized relationships, making the ontology re-usable in other applications. Moreover, the ontology allows the expression of extensive semantic relationships, including “kind-of”, “part-of”, “connects”, “branch-of”, “serves” and laterality relations.

Bibliographic retrieval

The NCBI Entrez Taxonomy

The NCBI Entrez Taxonomy is a hierarchical structure which contains all organisms represented in GenBank, with at least one nucleotide or protein sequence. There are seven top classes, i.e. archea, bacteria, eukaryota, viroids, viruses, other and unclassified. The information provided for each concept is quite elaborate and includes an ID, a rank, a genetic code, synonyms, and information as to the location in the taxonomy (“lineage”; see Fig. 7).

The Entrez Taxonomy complies with the definition given above: it is a hierarchical classification structure in which meaning is passed from more generalized to more specialized taxa.

MeSH

MeSH is an acronym for Medical Subject Headings, a controlled vocabulary produced by the National Library of Medicine (NLM), geared specifically for information retrieval. MeSH is used for indexing and searching journal articles in MEDLINE and other resources from the NLM Catalog.

The MeSH vocabulary consists of preferred terms, or descriptors, and entry terms. However, MeSH is more than ‘just’ a controlled vocabulary, it is a fully fledged thesaurus. The equivalence relationship is established by entry terms, which can be synonyms, near synonyms, abbreviations, or alternate forms of the MeSH term. Besides the equivalence relationship, two other typical thesaurus relations, i.e. hierarchical and associative relations, are represented.

The concepts are structured into a hierarchy, the MeSH tree, with 16 main branches. Each descriptor can have multiple parents and can consequently appear in several places in the tree. This can be illustrated by looking at a specific example, e.g. the Wolfram syndrome. This descriptor appears under the following subcategories: Nervous System Diseases [C10], Eye Diseases [C11], Male Urogenital Diseases [C12], Female Urogenital Diseases and Pregnancy Complications [C13], Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16], Nutritional and Metabolic Diseases [C18] and Endocrine System Diseases [C19].

Each descriptor has a notation—one or several MeSH number(s)—which is an indication of the concept’s relationship to its neighbouring concepts. This type of notation is referred to by Aitchison et al. [25] as an "expressive notation" or "hierarchical notation" (as opposed to (semi-) ordinal, synthetic and retroactive notations). The length of

---

the number indicates the specificity of the term: the longer the number, the more specific the concept. Figure 8 shows that Eye [A09.371] is broader than Anterior Eye Segment [A09.371.060], which, in turn, is broader than Anterior Chamber [A09.371.060.067].

When applied in information retrieval, the MeSH thesaurus can be an extremely valuable tool. It allows explosion of the search terms, and in Entrez PubMed, the terms entered by the searcher are automatically mapped to the appropriate MeSH term [53]. Term explosion, as described above, is a technique which increases the search yield considerably by searching not only for the term itself, but also for its narrower terms.

When examined for compatibility with the definition of a thesaurus as an information retrieval tool, the MeSH thesaurus proves to fulfill almost all requirements. It is a controlled vocabulary, with its descriptors and its non-preferred entry terms, which lead the searcher to the descriptors. The MeSH tree is organized hierarchically and includes the standardized relations as described in ISO 2788 [24]—the hierarchical, associative and equivalence relationship. These relationships are a priori relationships, i.e., they exist independently of the contents of the articles indexed with MeSH terms. Moreover, each term has a scope note, which contain background information on the usage and scope of the term. Scope notes can contain a definition formulated by the MeSH project partners or copied from other sources, like dictionaries or biomedical publications.

Greenberg [54], however, mentions a slight difference between thesauri for information retrieval and subject headings: thesauri generally tend to support post-coordinate searching, whereas subject headings have a pre-coordinated syntax. In pre-coordinated vocabularies, combinations of concepts are made at the indexing stage by the indexers, rather than at the stage of query formulation by the user. This means that the searcher can select very specific, unambiguous and “ready-made” queries instead of combining single-concept terms. Compare, for example, the pre-coordinated MeSH term “Physiological effects of drugs” and the terms “physiological”, “effect” and “drugs” in post-coordination. The advantages of pre-coordination described in [55] include proximity searches, where the searcher uses the relationships between concepts to select the best query. Pre-coordinated terms can be very useful for browsing, as they enable hierarchical displays. Some of the disadvantages stated in [55] are that pre-coordination requires human manual construction, an expensive and time-consuming task. Another disadvantage of pre-coordination might be that some end-users who are not familiar with this method of searching, might experience some problems. Post-coordination implies that concepts will have to be combined at the searching stage using Boolean operators.

Subject headings have multi-word terms, and often use inverted word order. MeSH can thus be defined as a thesaurus with the syntax of a subject heading list.

Controlled vocabularies

Controlled vocabularies used in bibliographic retrieval are usually incorporated into another structure, like a thesaurus (MeSH) or an ontology (the UMLS knowledge sources combine several controlled vocabularies).

The UMLS knowledge sources

The UMLS (Unified Medical Language System) Knowledge Sources combine three of the vocabulary systems described above: a thesaurus (the Metathesaurus), a lexicon (the SPECIALIST Lexicon) and an ontological structure (the Semantic Network).

The Metathesaurus consists of a large number of source vocabularies, including MeSH, SNOMED CT, the Gene Ontology, and other controlled vocabularies. Partly as a consequence of this combination of vocabularies, the Metathesaurus has a polyhierarchical structure. The Metathesaurus can be used in a wide range of applications, including information retrieval, and it becomes more powerful when used in combination with the SPECIALIST Lexicon and the Semantic Network.

The SPECIALIST Lexicon covers both the English general language and concepts from the field of biomedicine. It provides syntactic, morphological and orthographic information about the terms included in the lexicon.
A third component of the UMLS Knowledge Sources is the Semantic Network, which consists of Semantic Types, or broad subject categories, and Semantic Relations between these Semantic Types. This tool enables a consistent categorization of the concepts in the Metathesaurus.

The combination of the Knowledge Sources could be regarded as an ontology, as it represents knowledge from a specific field, with its concepts and extensive relationships. Furthermore, the Semantic Relations are expressed in a formal language. The combination of Semantic Types and Semantic Relationships makes this knowledge source much more versatile than the average thesaurus. However, not all relationships in the Metathesaurus are described formally. There are source vocabularies which do not use a formal representation language (e.g. MeSH) in the Metathesaurus.

A medical ontology is being developed by the Lister Hill National Center for Biomedical Communications, a research division of the U.S. National Library of Medicine. This ontology will combine the UMLS with SNOMED-RT, GALEN and MEDLINE citations and will represent a "model for proximity between medical concepts".16

The Gene Ontology

The Gene Ontology (GO) is a controlled vocabulary developed by the Gene Ontology Consortium for the annotation of gene products in model organisms. This vocabulary consists of three separate hierarchies, each representing concepts from a different subdomain: cellular components, molecular functions and biological processes. It has a polyhierarchical structure, i.e. narrower term or hyponym can have more than one broader terms or hypernyms, and it has a simple RDF syntax.

Despite its name, the GO is not an ontology as described in this article. Two types of relationships are present in this controlled vocabulary, namely the hierarchical is-a and part-of relationships and the equivalence relationship. The term 'ontology' here refers to the fact that knowledge about a specific domain is represented, including the relationships between the concepts.

Smith et al. [56] give an overview of the requirements for the Gene Ontology to become a cost-effective and semantically consistent system. These changes would convert the Gene Ontology into a system with the relational characteristics of a true ontology. However, making these changes would raise many difficulties. As a result, the Gene Ontology will probably remain in its current form, i.e. a controlled vocabulary.

Topic maps

Beier and Tesche [57] developed a medical information retrieval system, using the Medical Subject Headings (in English and German) and their classification as the knowledge layer, and the resources layer includes AHCPR Guidelines, journal articles and selected internet sites. This is a federated search system, i.e. a system which simultaneously searches several databases and/or web resources. The query entered by the user is automatically expanded with the topic name (the preferred term), synonyms, translations and definition.

The interface (see Fig. 9) clearly shows the typical topic map structure of the resources layer and the superimposed knowledge layer. Between both layers, some extra MeSH information (MeSH code, definition and annotations, synonyms and translations) is displayed, in order to help the user find the right topic name for his or her search. The user can select the resources in which he wants the engine to search.

This topic map complies with the ISO standard and with the description of topic maps given above, except that the knowledge layer was not deduced from the resources.

Conclusion

There is a need for consistent terminology in the domains of linguistics, knowledge management and information retrieval, as in most fields of knowledge. Terms such as taxonomy, classification, thesaurus and ontology are often used interchangeably, resulting in definitions which are formulated from different perspectives.

Not only are the terms used in different ways, their scope may also change. When terms are adopted in other fields—a shift which often has a historical aspect—this may cause some confusion.

Unambiguous definitions are proposed for each of the terms in question, depending on the context they are used in, and criteria are presented for a more consistent use of the various competing designations. Some of the best-known vocabularies pertaining to biomedical linguistics, knowledge management and bibliographic retrieval are reviewed and examined for their compatibility with the definitions given in this article. We concluded that the use of the designations ‘ontology’ or ‘thesaurus’ in the biomedical domain—as elsewhere—is not always consis-

16 http://lhncbc.nlm.nih.gov/lhc/servlet/Turbine/template/research,langproc,MedicalOntology.vm
tent. More specifically, we found that the ICPC-2/ICD-10 thesaurus and 3BT are not thesauri, but bicoded classifications and that the Gene Ontology is not really an ontology but a controlled vocabulary.

Table 5 below gives an overview of the systems in biomedicine in a two-dimensional structure: according to their domain of application (linguistics, knowledge management—including medical registration—and bibliographic-
clical retrieval) and the kind of vocabulary (taxonomy, classification, thesaurus, controlled vocabulary, ontology or topic maps) they represent.

References


