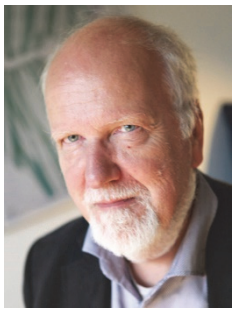
Keyword†

Marco Lardera* and Birger Hjorland**

* Coop Consorzio Nord Ovest, Pavia, Italy, <larderamarco@hotmail.com>

** University of Copenhagen, Faculty of Humanities, Department of Communication, South Campus, building 14, 2. Floor, Karen Blixens Plads 8, 2300 Copenhagen S, Denmark, < birger.hjorland@hum.ku.dk>

Marco Lardera has a Master's Degree in Philosophy. He has worked at scientific libraries in the University of Pavia, where he has developed a strong interest in knowledge organization issues. He has contributed to the development and promotion of the online DDC-based tool SciGator. He is currently employed in the IT field, working with databases and search engines.



Birger Hjørland holds an MA in psychology and PhD in library and information science. Since March 2020, he has been Professor Emeritus at the Department of Communication, University of Copenhagen and formerly professor in knowledge organization at the Royal School of Library and Information Science/Department of Information Studies (2001-2020) and at the University College in Borås (2000-2001). He is a member of the editorial boards of Knowledge Organization, Journal of the Association for Information Science and Technology and Journal of Documentation. His h-index on 2020-114 is fifty in Google Scholar and twenty-nine in Web of Science.

Lardera, Marco and Birger Hjorland. 2021. "Keyword." *Knowledge Organization* 48(6): 430-456. 117 references. DOI:10.5771/0943-7444-2021-6-430.

Abstract: This article discusses the different meanings of 'keyword' and related terms such as 'keyphrase', 'descriptor', 'index term', 'subject heading', 'tag' and 'n-gram' and suggests definitions of each of these terms. It further illustrates a classification of keywords, based on how they are produced or who is the actor generating them and present comparison between author-assigned keywords, indexer-assigned keywords and reader-assigned keywords as well as the automatic generation of keywords. The article also considers the functions of keywords including the use of keywords for generating bibliographic indexes. The theoretical view informing the article is that the assignment of a keyword to a text, picture or other document involves an interpretation of the document and an evaluation of the document's potentials for users. This perspective is important for both manually assigned keywords and for automated generation and is opposed to a strong tendency to consider a set of keywords as ideally presenting one best representation of a document for all requests.

Received: 15 November 2020; Accepted: 4 June 2021

Keywords: keyword, keyphrase, human indexing, automated indexing

† Derived from the article titled "Keyword" in the ISKO Encyclopedia of Knowledge Organization, Version 1.0 published 2020-11-17. Article category: KOS Kinds.

1.0 Keyword and related terms

The term "keyword" is one among many related terms, that are sometimes considered synonyms. Soergel (1974, 31-4), for example, considered 'keyword' synonym with 'descriptor', 'clueword', 'cueword', 'index term', and partly also 'subject heading'. Wikipedia (2020) considers 'index term', 'subject term', 'subject heading' and 'descriptor' as keywords.[1] Before we directly address the meaning of keyword in Section 1.6.4, we have therefore chosen to consider a range of related terms. We will argue that 'keyword' is one of many kinds of terms used in information science for describing documents, and that many of these terms have specific meanings, and thus should *not* be considered synonyms. The understanding of these different meanings assumes an understanding of various kinds of indexing and retrieval mechanisms. A summary of conclusions about the concepts presented is given in table 2 at the end of this Section 1.

1.1 Term, index term, free-text term and uniterm

1.1.1 Term

The *Oxford English Dictionary* (2020b) has no entry for the noun *term* but defines *terminology*: "The system of terms belonging to any science or subject; technical terms collec-

tively; nomenclature. Also: the scientific study of the proper use of terms." This is in agreement with how 'term' is defined by Wiktionary (2020, sense 5), a: "word or phrase, especially one from a specialised area of knowledge".

Jacquemin and Bourigault (2003, 600-1) characterized "the classical view" of terms as the "dominant approach to termhood [which] stems from the General Theory of Terminology which was elaborated by E. Wüster in the late 1930s, with the Vienna Circle". This positivist view

"assumes that experts in an area of knowledge have conceptual maps in their minds. This assumption is misleading and unproductive because experts cannot build a conceptual map from introspection. Terminologists constantly refer to textual data.

According to a definition that is better suited to corpus-based terminology, a term is the *output* of a procedure of terminological analysis. A single word, such as *cell*, or a multiword unit, such as *blood cell*, are terms because they have been manually selected as such."

In information retrieval (IR), "term" has, however, a much broader meaning. In full-text indexing the term-document matrix is a table that describes the frequency of terms that occur in a collection of documents (Baeza-Yates and Ribeiro-Neto 2011, 62-3). In this context "term" is every word in a document and in a collection of documents (excluding possible stopwords). This is an extremely important model, and thus "term" (and not, for example, "keyword" or "index term") in this broad meaning is the dominating terminology in IR.

In the same context, Manning, Raghavan and Schütze (2008, 21-2; italics in original) use the token/type distinction to define 'term':

> A *token* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A *type* is the class of all tokens containing the same character sequence. A *term* is a (perhaps normalized) type that is included in the IR system's dictionary. The set of index terms could be entirely distinct from the tokens, for instance, they could be semantic identifiers in a taxonomy, but in practice in modern IR systems they are strongly related to the tokens in the document. However, rather than being exactly the tokens that appear in the document, they are usually derived from them by various normalization processes.

Anderson and Pérez-Carballo (2005, 1.58, 18-19) emphasized that a term may consist of more than one word:

> A term is a word or a phrase representing a single concept or multiple concepts that are tightly bound together in the context of a particular IR database [...]. Some concepts need more than one word to express them, for example, information science or venetian blind. Some terms could be divided into two separate terms, but they are used so commonly together in a consistent order that they are considered a single bound term or compound term.

Because of the issue pointed out by Anderson and Pérez-Carballo, traditional, "classical databases" distinguish between "word indexing", "phrase indexing" and combined word and phrase indexing (whether #information science should be found under #information and #science as separate words, or only under 'information science' as a phrase or under both the words and the phrase). [2]

1.1.2 Index term.

Baeza-Yates and Ribeiro-Neto (2011, 61-2; italics in original) wrote that *index term* has different meanings in search engine designing versus library and information science:

> *An* index term *is a word or group of consecutive words in a document. In its most general form, an index term*

	Document 1	Document 2	Document 3	Document N
Term 1	0	3	3	1
Term 2	5	4	1	0
Term 3	3	3	2	4
Term 4	0	4	4	1
Term 5	1	1	2	1
Term 6	2	1	5	4
.				
.				
Term i	0	1	2	0

Table 1. A model of the term-document matrix used in information retrieval.

> *is any word in the collection. This is the approach taken by search engine designers. In a more restricted interpretation, an index terms* [sic] *is a preselected group of words that represents a key concept or topic in a document. This is the approach taken by librarians and information scientists*.

However, as we saw in Section 1.1.1 in full-text IR (including search engine designing) it is "term" rather than 'index term', that is the common expression. We therefore suggest that "index term" is limited to the meaning, which Baeza-Yates and Ribeiro-Neto defined for library and information science: a preselected word or group of words that represents a key concept or topic in a document (synonym to keyword, as presented below). 'Index term' is a hypernym for subject heading, descriptor, derived index term, uniterm, keyword and tag.

1.1.3 Free text term

Anderson and Pérez-Carballo (2005, 1.61, 19) give this definition of "free-text term":

> Often shortened to free text, free-text term usually refers to the use of uncontrolled words or terms from natural language text for indexing or searching. When one searches the actual text of a document, one is searching the free-text terms that are found in the document. The difference between free-text terms and just terms is that sometimes terms may be standardized, at least a little, with respect to format, and they may also have links with the most common synonyms or equivalent terms, even if they are not controlled to the extent of descriptors.

Free-text terms are thus opposed to controlled terms (either lightly normalized or derived from a controlled vocabulary (CV), such as a thesaurus).

1.1.4 Uniterm

"Unit term" was coined by Taube, Gull and Wachtel (1952) for use in post-coordinate indexing and has often since been called 'uniterm'. It was a kind of minimal processing of natural-language terms, in which, for example, *s* for the plural form of nouns was removed. It represented a low-level form of indexing, like modern keywords.[3] It was studied by the Cranfield Experiments (Cleverdon 1962), in which the authors compared the Uniterm system to the classic library classification schemes (UDC, faceted classifications, and an alphabetical system). The results, although controversial, showed that a uniterm based search strategy scores better than classic systems in both precision (fraction of the retrieved items that are relevant) and recall (percentage of the relevant items retrieved), which came as something like a shock for the documentation profession at the time. The term *uniterm* has not, however, reached a clear definition. Costello (1961, 20) wrote: "An intensive study of the literature on Uniterm indexing has revealed that since Dr. Mortimer Taube originally published details of his system there have developed nearly as many different definitions and interpretations of uniterms and uniterm indexing as there are uniterm systems and documentalists responsible for their operation and maintenance". It seems that 'uniterm' may today be considered a synonym for 'keyword'.

1.2 Heading and subject heading

1.2.1 Heading

'Heading' is the general term for information used in printed catalogs and bibliographies as an access point, which includes 'subject heading' as a narrower term. Anderson and Pérez-Carballo (2005, 20, §1.62):

> In displayed indexes (indexes that are designed for visual inspection by humans as opposed to non-displayed indexes that are searched by computer algorithm), index terms are combined into headings consisting of multiple terms. It is possible to have index headings with only single terms, but headings of two or more terms are more meaningful, because the lead term is modified or amplified or described by the subsequent term or terms. The subsequent term or terms create a context for the first, or lead, term.

Reitz (2004):

> Heading: The name of a person, corporate body, or geographic location; the title proper of a work; or an authorized content descriptor (subject heading), placed at the head of a catalog entry or listed in an index, to provide an access point. In library cataloging, genre/form terms are also used. In AACR2, form of entry is subject to authority control. See also: main heading and subheading. In Dewey Decimal Classification (DDC), a word or phrase used as a description of a class, given in the schedules in conjunction with the class number, for example, "Library and information sciences" for which the class notation is 020.

1.2.2 Subject heading

Subject heading is a subcategory of heading used about information provided as a subject access point (SAP).[4] The term is associated with subject heading systems such as Li-

brary of Congress Subject Headings (LCSH), Physics Subject Headings (PhySH) (Smith 2020), and many more.

Reitz (2004):

> Subject heading: The most specific word or phrase that describes the subject, or one of the subjects, of a work, selected from a list of preferred terms (controlled vocabulary) and assigned as an added entry in the bibliographic record to serve as an access point in the library catalog. A subject heading may be subdivided by the addition of subheadings (example: Libraries--History--20th century) or include a parenthetical qualifier for semantic clarification, as in Mice (Computers). The use of cross-references to indicate semantic relations between subject headings is called syndetic structure. The process of examining the content of new publications and assigning appropriate subject headings is called subject analysis. In the United States, most libraries use Library of Congress subject headings (LCSH), but small libraries may use Sears subject headings. Compare with descriptor. See also: aboutness and summarization.

Headings and *subject headings* are also used in the electronic environment although the concept originated in the print environment (and the idea of selecting or repeating some information at a specific place in a record is obsolete in the electronic world). However, subject heading systems such as LCSH are also used in electronic databases (e.g., in online public access catalogues -OPACs). A study by Larson (1991) indicated, however, that these subject headings may be harmful rather that fruitful, and that experienced users may learn to avoid them, a view that has been challenged.[5] The difference between subject headings and descriptors is, that the former is developed for pre-coordinate indexing, the latter for post-coordinate indexing (Indexing 2.2.3). Soergel (1974, 31-4) wrote, however, that although this usage corresponds to the original definition of 'descriptor' it has not gained currency in the profession.[6] ISO 5127:2017 fails to distinguish between descriptor[7] and subject heading[8].

It is here recommended to use the original meaning, where precoordinate vs post-coordinate indexing is the defining difference between "subject heading" and "descriptor".

1.3 Descriptor

Another concept strongly related with *keyword* is *descripto*r. The first publication using this word is traced to Calvin Mooers (1950)[9], who used it to define a new method for specifying the subject of a document. According to Mooers, the descriptor technique is based on principles that makes it different from traditional subject indexing: taking into consideration the specificity of the user population who interact with a given set of documents, focusing on the retrieval capabilities rather than describing the content of the document and using an idea-related vocabulary unlinked from the literal words contained in the text.

Mooers found that a good descriptor set contains about 200-400 terms only: adding more of them may decrease the retrieval effectiveness. Thus, it can be created and maintained rather quickly by a subject-matter expert. (Which is not, however, in accordance with most modern thesauri).[10] Mooers (1972; 2003) further found that the most critical aspect of using descriptors is the necessity of a knowledgeable librarian able to understand the concept of "idea-words" separated from the actual words contained in the document. Mooers emphasizes that most of the people who tried to work with descriptors were unable to do that, consequently causing the failure of the method in the librarian community, even though users reported to appreciate the information retrieval capabilities of descriptors. He concluded (2003, 821):

> In epilogue, the descriptor method is largely a failure because it proved to be beyond the capabilities of the persons who chose to enter the service profession of librarianship in which descriptors were to be used.

In the recent literature the word "descriptor" has been used in a more general sense, indicating a term in a CV used to describe a concept that is linked to all the other related terms.[11] Anderson and Pérez-Carballo (2005, 1.60, 19) explain this meaning as following:

> The term descriptor is usually reserved for a term that is part of a controlled indexing language. Such indexing languages are often listed in a thesaurus. For each concept included in the indexing language, one descriptor will be chosen to represent the concept, and all other terms that can be used for the same concept are linked to the descriptor by means of cross references. Thus, if a thesaurus uses the descriptor lawyer, then it might not use the terms attorney, barrister, solicitor, or counselor-at- law. Each of these alternative terms would be linked to the preferred descriptor lawyer and would be given the status of un-used synonymous or equivalent terms.

However, as described in 1.2.2, the main difference between subject headings and subject heading systems on the one hand and descriptors and thesauri on the other is that the first are designed for pre-coordinate indexing techniques (in typed or printed catalogs or indexes) while the last are based on post-coordinate indexing (which are often dominating in electronic databases).[12]

1.4 Concept symbol

Citation techniques introduce the possibility of a new kind of highly valuable "keyword", the meaning of which can be found in the citation context of citing papers. This idea was introduced by bibliometrician Henry G. Small (1978), who described citations as "concept symbols" and as researchers' "tools-of-the-trade" and emphasized that they are interpretations of works. Citations may thus influence how other researchers view the importance of a given document /which may be different from how the author of the document saw its importance and expressed this, for example, in the title of the document (see the example about an article by Ronald Fisher in Hjørland (2017, 60) https://www.isko.org/cyclo/subject#2.6). Small also emphasised that when a document is cited by a large group of researchers, its meaning can be standardized by the citations it receives. These citations may form a consensus, which in the research community becomes a symbol of the meaning and importance for the document; for example, the reference "Hulme (1911)" will, by many researchers in the field of knowledge organization, stand as the concept symbol for the article, which introduced the principle of literary warrant even if this term was not used by that article.

This bibliometric thinking has provided a fundamentally new way of looking at keywords and the subject of documents.

1.5 Tag

Furner (2010, 1858-9) established that a tag is to be considered a kind of an index term:

> Tagging is the activity of assigning descriptive labels to useful (or potentially useful) resources. In effect, the labels that are assigned by agents to resources are the names selected by those agents for the categories, classes, or concepts into which the resources are deemed by the agents to fall. The tagger who assigns the tag "cat" to a digital photograph of a cat is simply specifying that the photograph is one of the resources that the tagger wishes to place in a category or class named "cat."
> In the parlance of mid- to late-twentieth-century information science, "cat" is an *index term*, and the activity of assigning index terms (words, phrases, codes, etc.) to resources (books, journal articles, Web pages, blog entries, digital photos, video clips, museum objects, etc.) has long been known as *indexing*, whether undertaken by people or machines.

Tags, in the most common meaning, are user-generated keywords, implemented by many online platforms (including the majority of the social media such as Flickr and Twitter), that let users describe the resources using their own personal vocabulary, sometimes with the help of a recommendation system able to suggest potentially relevant keywords to the tagger (Jaschke 2008).

Many studies have discussed the benefits of this bottom-up approach, including the capability of the crowd to enrich existing classifications with more up-to-date terms but at the same time folksonomies risk increasing the noise level, tagging the same concept with different terms, therefore reducing the search and retrieval possibilities (Adler 2009).

Holstrom (2019) underlines that the peculiar nature of tags lies in the fact that casual indexers tagging a document are performing *subject analysis* but not *subject representation*, since they don't map their natural language tags to a controlled language with clear and explicit semantic relationships.

Nevertheless, social tagging has been proposed as a valid replacement for organizing both web contents and library catalogs. Comparisons between tags and CVs has been made with mixed results, suggesting that the two approaches need to be considered as complementary and not mutually exclusive (Syn 2009).

1.6 Word, stopword, n-gram, keyword and keyphrase

1.6.1 Word

A word is a linguistic unit that is difficult to define. Uhlenbeck (2003, 377) wrote:

> Words: Linguistic units, probably because of their pragmatic and functional character, are notoriously difficult to define. For the sentence as well as for the word, many definitions have been proposed; but so far none has gained general acceptance (for surveys of word definitions, see Togeby 1949, Krámský 1969, and Juilland and Roceric 1972). This lack of consensus among linguists stands in sharp contrast to the general agreement of native speakers everywhere, who seem convinced that they have words at their disposal for daily use in actual speech.

In natural language processing stemming techniques and lemmatization are used to identify words derived from a common root and appearing in a variety of forms or determining the lemma for each word form that occurs in text. The lemma of a word encompasses its base form plus inflected forms that share the same part of speech.

In information retrieval a sequence of characters surrounded by blanks or punctuation are normally regarded as a word (but see also Section 1.6.3 N-gram). In bibliographical records a given field may be "word indexed" or "phrase

indexed" (or both).[13] The descriptor #child custody is indexed by words with the expressions "child" and "custody" as index terms. It may be phrase indexed with "child custody" as an index term. In the last case the blanks are ignored when the expression is represented in the inverted file of the database. Often "word" does not include notations such as classification codes but only considers sequence of characters that form parts of natural language.

1.6.2 Stopword

'Stopword' is a term that has been put in a stopword list because it is common and has been considered non-relevant for describing the subject of the documents in a given collection. That said, just as there is subjectivity in the choice of any kind of index terms and keywords, different stop word lists may be optimal in a given context. Dolamic and Savoy (2010) reported experiments with various stopword lists in different languages and demonstrated that even the inclusion or exclusion of the English definite article *the* in stopword lists may influence search results. A stopword can be considered the opposite of keyword, as keywords are selected because of their importance in describing the subject of a given document (as presented in Section 1.6.4).

1.6.3 *N*-gram

An *n-gram* is a contiguous sequence of *n* items from a given sample of text or speech. The items can be phonemes, syllables, characters, words or base pairs according to the application. In the present context it can be said that instead of words as units, any sequence of characters (be they letters, numbers, punctuations, or spaces) may be used. *N* can be any whole number (n=1 is a unigram; n=2 is a bigram; n=3 is a trigram; n=4 a "four-gram" etc.) In this way the somewhat arbitrary splitting of a text in words, can be replaced by a splitting of any sequence of characters of any length (or combining multiple lengths of *n*-grams). This is a technology increasingly used in computational linguistics, information retrieval and other fields (see, e.g., Cohen 1995).

1.6.4 Keyword and keyphrase

Keyword is a term with many different meanings, being used in various fields such as linguistics, computer science, information retrieval, library science and knowledge organization. Sometimes "keywords" are used for words that are central in a culture (or the analysis of that culture), see, for example, Williams (2015). The *Oxford English Dictionary* (2020a) provides the following definitions:

1. A word or concept of great significance.
2. A word used in an information retrieval system to indicate the content of a document.
3. A significant word mentioned in an index.

In library and information science (LIS) keywords are subject access points (SAP) and metadata. Their roles are determined in relation to different theories of indexing and information retrieval. Some sources define keywords as limited to free text terms, for example, the ISO 5127:2017 standard consider only the title and text of a document as a keywords source:

> *Significant word (1) <orthographic word> (3.1.5.18) taken from the title (3.7.4.01) or the text (3.2.1.05) of a document (3.1.1.38) to represent all or part of the content.*

The same understanding is reflected in Anderson and Pérez-Carballo (2005, 1.61, 19)[14] and in Wellisch (1995).[15] However, Feather and Sturges (1996, 341) did not demand that a keyword is taken from document, but provided the following definition: "A word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document".[16] East Carolina University Libraries[17] defined keywords as natural language terms in opposition to terms from controlled vocabularies. However, the term "keyword" can be a source of confusion, as explained by Gross and Taylor (2005, 213): "In 2005, most online catalogs can search every field in a record, although moving from catalog to catalog can be quite confusing, with the definition of "keyword search" being quite different as to which fields are included in that search". Larson (1991, 199) provides the following explanation for the confusion:

> The MELVYL system provides access to the database through both keyword and "exact" indexes and provides integrated authority control for personal and corporate name searches. Keyword indexes permit searching on any word from particular field or set of fields in the MARC record. The "exact" indexes provide searches with left-to-right matching of specific fields in the MARC record with optional right truncation. Both keyword and exact indexes ignore case, punctuation, and a small set of stopwords in matching, and indexes may be combined in command mode searches using Boolean operators.

In this article, keyword can be understood as including a single word from a CV (which is made searchable as such in a database). This corresponds to a mention by Anderson and Pérez-Carballo (2005, 222 §12.75):

> [K]eyword searches using Library of Congress subject headings: A third way of searching is the 'keyword' approach. It should be invoked by the intelligent OPAC in several instances. In this first example, a searcher has used the 'exact' approach to 'jazz music' but wants more options. The next step would be to apply the 'keyword-in-main-heading approach'" (retrieving all subject headings which include 'jazz").

As argued by Larson (1991, 199) a keyword can be considered different from the 'exact' indexes which provide searchers with left-to-right matching of specific fields in databases. Larson mentioned MARC record, but this can of course be generalized to any type of bibliographical databases. Keyword is by some researchers meant to include phrases. Hartley and Kostoff (2003, 433), for example, wrote: "While "key words" is common usage, these descriptors should strictly speaking be called "key words and phrases", since multiword phrases can be used as descriptors in most publications". Sometimes "multiword keywords" are suggested (e.g., Thomaidou and Vazirgiannis 2011), but here the term "keyphrase" is suggested as a better alternative, as suggested by Siddiqi and Sharan (2015, 18): "Both single words (keywords) and phrases (key phrases) may be referred to as 'key terms'" and

> A keyphrase connotes a multi-word lexeme (e.g., computer science engineering, hard disk), whereas a keyword is a single word term (e.g., computer, disk). Using single words, as index terms, can sometimes lead to misunderstanding. For example, in phrases like 'hot dog', the constituent single words does [sic!] not have their regular meanings and are thus quite misleading if used as individual indexing terms. Also, they may be too general, e.g., words 'junior' and 'college' are not specific enough to distinguish 'junior college' from 'college junior'. Also, when selected from a controlled vocabulary, keyphrases reduce the problems associated with synonymy and polysemy in natural language.

It follows that a keyword can be both:

- A kind of index term: as such it can be a word from a controlled terminology made searchable as a separate word, or it can be free terms assigned by indexers (including authors and taggers as indexers, or keywords obtained by citation links as in Keyword Plus as described below).
- A free-text term used for searching, that can come from either document titles, abstracts, full-text, or any other element in a document made available in a search system.

In the information retrieval field, a keyword (or a keyphrase) is a word (or a phrase) representing the topic or the content of a document, used for retrieving it from a source of information, such as a database. Although some authors also use the word *keyword* in a broader meaning, considering as a keyword every word in a document by which you can perform a full-text search, Firoozeh et al. (2020) speak of "keyness properties" of words and phrases and distinguish three main types of keyness properties: informational, linguistic, and domain-based.

- Informational properties of keyness include principles of exhaustivity and specificity, and further minimality, impartiality, and representativity.
- Linguistic properties of keyness include well-formedness and citationess.
- Domain-based properties of keyness include conformity, homogeneity and univocity

According to Hjørland and Kyllesbech Nielsen (2001), keywords can be considered a type of SAP, searchable entities used to retrieve documents starting from the topic of a document, extending more traditional entry points such as title or abstract.

As already described in Section 1.1.4 an early example of keyword-based information retrieval evaluation was represented by the Cranfield Experiments (Cleverdon 1962). The authors of those experiments compared the Uniterm system, a low-level form of indexing, similar to modern keywords, to the classic library classification schemes (UDC, faceted classifications, etc.). The results, although controversial, showed that a keyword-based search strategy scores better than classic systems in both precision (fraction of the retrieved items that are relevant) and recall (percentage of the relevant item retrieved).

Similar experiments are performed each year, since 1992, in the TREC (Text Retrieval Conference) tests, where multiple research teams compete in retrieving documents from large databases in the most efficient way (Sanderson and Croft 2012).

A different and more specific meaning for the word keyword is present in the field of search engine optimization (SEO), where a keyword is a term contained in the HTML code of a page with the purpose of improving the page rank in the search engine result pages. Even different is the concept of keyword in linguistics: words that allow to discriminate between two or more corpora of documents, identifying the unique elements of each one (Bekhuis 2015).

Finally, in the indexing and knowledge organization fields the most reasonable way to address the issue of defining what is a keyword is to consider both CVs, terms and free-text terms as potential keywords. Under this comprehensive definition, for instance, we can conceive as keywords the terms

contained in a thesaurus, if used for indexing, as well as the keywords provided by authors when submitting a manuscript to a journal. Free-text terms, however, should be considered in this regard only when significant. Since all the words in a document are, according to the definition discussed in this section, free-text terms (including the meaningless ones), we may recognize as keywords only the ones that bring with them a strong semantic value and are useful for expressing the subject of the document.

There is a certain degree of subjectivity in this definition of keyword because the assessment of what is "significant" is, within certain limits, a matter of personal judgement or theoretical view. A keyword can be understood as a word highlighted (or assigned) by somebody in a text. It seems obvious that different words or phrases may be highlighted according to the purpose or interest of the person doing the highlighting. Keywords are thus subjective, but the most fruitful understanding of this subjectivity is to relate it to common perspectives and interests because this is the best way (or the only way) of making general principles for selecting keywords. In other words: A keyword should not be considered as something, which a given word either is or is not. A keyword should be considered a word that from some perspective provides a useful description of a document.[18]

Term	Broad meaning (IR): Any word in a document which is candidate for indexing, including algorithmic full-text retrieval. Narrow meaning: Word or phrase from the terminology of a given field. (E.g., in the field of knowledge organization #classification and #controlled vocabulary are terms).
Index term	Broad meaning: Synonym for "term" (any word in a document used for or potential useful for indexing). Narrow meaning: A word or phrase used for indexing; a single word index term corresponds in searching to a keyword. In the narrow meaning index term is hypernym for 'subject heading', 'descriptor', 'derived index term', 'uniterm', 'keyword' and 'tag'.
Free-text term	Term as it appears in the documents to be indexed, as opposed to terms from controlled vocabularies.
Uniterm	Free-text terms with a low-level degree of normalization (e.g., "color" is a uniterm for each of the words "color", "colors", "colour", "colours").
Heading	The general term for information used in printed catalogs, bibliographies and browsable lists as an access point. Also used in the term "HTML heading" and in OPACs (here probably caused by their use of subject heading systems developed in the time of printed indexes). A heading may be single terms or a string of terms.
Subject heading	Subcategory of heading used as a subject access point (SAP). Normally from a controlled vocabulary (a subject heading system) with pre-coordinated terms (as opposed to descriptors).
Descriptor	A keyword from a controlled vocabulary designed for post-coordinative indexing (as opposed to subject heading).
Concept symbol	A symbol associated with a given meaning by a community of users. Specifically: The meaning associated with a given bibliographical reference by a set of citing papers. (E.g., "Hulme (1911)" as symbol for the text that introduced the concept 'literary warrant').
Tag	User-generated keyword, implemented by many online platforms, that let users describe the resources using their own personal vocabulary, sometimes with the help of a recommendation system able to suggest potentially relevant keywords.
Word	In information science: A sequence of characters surrounded by blanks or punctuation. More than one consecutive word can be treated as a unit (a phrase) in phrase indexing. (E.g., the term 'information retrieval' consists of two consecutive words, which can be 'phrase indexed' to function as one expression.
Stopword	A common word considered non-relevant for describing the contents of a document and included in a stopword list.
N-gram	In general: A contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, characters, words, or base pairs according to the application. In the present context: Any sequence of characters of a given length.
Keyword and keyphrase	A word, selected for its significance for describing the content of a document, whether or not the selection is done by man or machine, whether it is a controlled or uncontrolled term, or whether it is derived from or assigned to a document. A keyword may be derived from a subject heading, but *subject heading* and *descriptor* are terms related to specific kinds of indexing languages. A keyword is different from a full multiterm subject heading or index term, which is called a *keyphrase*. Both keywords and keyphrases are sometimes included in the term *key terms* .

Table 2. Summary of conclusions about the concepts presented.

2.0 Functions of keywords

It follows from the above given definition of "keyword" that the function primarily is to characterize documents to improve the findability of the documents described (or to find pages or passages within documents). As such, keywords represent one among many kinds of indexing languages.

In printed books, the indexes may improve the findability of the pages in which a particular subject or concept is discussed, but not where every word appears (which is more the task of concordances).[19] The difference between "word indexing", "concept indexing" and "subject indexing" is important, as first pointed out by Bernier (1980) and developed by Hjørland (2018, §3.6). In printed books keywords may also (in combination with tables of contents, section headings and internal references) support browsing, for example, by using keywords as margin notes, or by different kinds of typographical highlighting.

In printed journal indexes, pre-coordinated keywords (subject headings) were mostly used and, as argued by Milstead (1984, 187), pre-coordination is the only appropriate method in the print environment. Computers compare lists of document numbers posted to keywords to locate a combination of concepts, but this is not practical for human beings to do. (But sometimes printed indexes are made from post-coordinate indexes because the producers do not want to produce two different indexes).

In the electronic bibliographic databases post-coordinate indexing was introduced, allowing flexible combination of keywords ("descriptors"), but also introducing the problem known as "false drops".[20] These databases typically provide many kinds of SAP (e.g., descriptors from controlled vocabularies, free text terms in titles and abstracts and author supplied keywords). Much research has been done studying the relative contributions of such different kinds of subject access points (cf., Hjørland and Kyllesbech Nielsen 2001; McJunkin 1994).

In picture indexing, two approaches have been introduced (cf., Chu 2001): (1) "content based indexing" (the techniques of indexing and retrieving images based on, among other features, color, shape, and texture), and (2) "description based indexing" (techniques based on manually assigned captions, keywords, and other descriptions). In both cases keywords may be used, but in order to connect a keyword to a picture, the word cannot be derived as in texts, but must somehow be assigned on different levels of abstracting in analyzing the picture (from color and shape in the lowest level to artist style, such as impressionism, at the highest level of abstraction).

Besides these functions keywords and keyphrases have become important concepts in information technology, where they are used for text summarization, abstracting, ontology construction, recommender systems, text analyses, browsing[21], subtitling (e.g., of videos) and more. This is a highly active field of research in the second decade of the 21st century.

The authors of the present article write based on the understanding that in all contexts and for all purposes it is important to consider that keywords should not just be understood in a document-centered way as "semantic condensation" of the content of a document but should be understood in a request-oriented or policy-oriented way as an indication of the subject of a document (cf., Hjørland 2017, §2.4): A document does not have a subject, but is assigned one or more subjects depending on the purpose of the indexing. There can always be different perspectives on given text, and the assignment of keywords cannot avoid this subjectivity (more about this in Section 5). Therefore, keywords should not be considered more or less wrong or correct in relation to a given document but should be considered more or less functional to facilitate given interests.

2.1 Some indexes associated with the term "keyword"

Some kinds of indexes contain the word "keyword" in their name. A well-known example is the KWIC (KeyWord In Context) index, normally attributed to Hans Peter Luhn who demonstrated his system at IBM in 1958 and published an article about it in 1960.[22] KWIC is described by Manning and Schütze (1999, 31) as a "concordancing program", a program that can produce a concordance (cf., endnote 19). The KWIC index is one of the many kinds of "title derivative indexing techniques" (cf., Feinberg 1973) although also, for example, for full-text documents and for indexes in thesauri. KWIC indexes are probably the oldest kind of automatic indexing systems. The steps in creating a KWIC index are:

- Keywords contained in the title are identified using a list of stopwords[23] to exclude irrelevant words.
- All the possible rotations or cyclings[24] are generated around a central keyword (which is often emphasized, e.g., as in Figure 1 by using bold characters);
- The generated entries are then sorted in alphabetical order.

Naval Warfare in the	***Atlantic**, History of*
Naval Warfare, History of	***History** of in the Atlantic*
History of	***Naval** Warfare in the Atlantic*
History of Naval	***Warfare** in the Atlantic*

Figure 1. A typical KWIC index for the title *History of Naval Warfare in the Atlantic* (remark the unconventional presentation format with the lead term displayed in the middle of the column rather as at the left).

To deal with complaints raised against this format two other formats (KWOC and KWAC) were developed. In the KWOC index (Key Word Out of Context) the keywords are displayed outside the title, usually at the beginning of the line.

Atlantic,	*History of Naval Warfare in the Atlantic*
History,	*History of Naval Warfare in the Atlantic*
Naval,	*History of Naval Warfare in the Atlantic*
Warfare,	*History of Naval Warfare in the Atlantic*

Figure 2. A KWOC index demonstrating the same title.

Finally, a KWAC index ("Key Word Alongside Context", also known as "Keyword and Context" and "Key Word Augmented-in-Context"), as described by Anderson and Pérez-Carballo (2005, 261), preserves word pairs and phrases, but words preceding the keyword are no longer contiguous.[25]

Atlantic,	*History of Naval Warfare in the*
History,	*of Naval Warfare in the Atlantic*
Naval	*Warfare in the Atlantic, history of*
Warfare	*in the Atlantic, History of Naval*

Figure 3. A KWAC index demonstrating the same title.

Another kind of keyword index is the "Keywords Plus" used by the *Web of Science* database. It takes keywords from the references in a given article. The Clarivate Analytics homepage[26] writes:

> The data in KeyWords Plus are words or phrases that frequently appear in the titles of an article's references, but do not appear in the title of the article itself. Based upon a special algorithm that is unique to Clarivate Analytics databases, KeyWords Plus enhances the power of cited-reference searching by searching across disciplines for all the articles that have cited references in common.

The system is described in the literature by Garfield (1990a; 1990b), Garfield and Sher (1993) and Zhang et al. (2016). The last paper found (967): "Keywords Plus terms were more broadly descriptive [compared to author-assigned keywords]. Keywords Plus is as effective as Author Keywords in terms of bibliometric analysis investigating the knowledge structure of scientific fields, but it is less comprehensive in representing an article's content." Further (971): "Keywords Plus terms emphasized research methods and techniques, whereas Author Keywords tended to hone in on specific diseases and conditions," or generalized: Author keywords may be best at identifying documents about given topics and Keywords Plus best at identifying papers about specific research methods.

Such keyword-based indexes had the advantage of being very fast and inexpensive to be generated using computers. The capabilities of full-text search engines have made such indexes less necessary, and they have been in decline. However, the automatic generation of keywords and keyphrases from full text is, as shown later, a highly active research field today. Also, some attempts have been made in order to revitalize the KWIC technique as an instrument to enhance filtering of results in web search engine, displaying the most frequent context of the search query and letting the user select one of them (Käki 2006).

3.0 Classification of keywords in library and information science (LIS)

Hjørland (2011) made a distinction between four kinds of indexing based on the dichotomies of human versus computer-based indexing and derived versus assigned indexing (or extractive versus abstractive methods). These four kinds may also be used to classify kinds of keywords:

- *Human-based derived keywords*: Keywords extracted by a human from parts of the document (text, abstract, title or others) on the bases of the indexer understanding about which words may be relevant in the given context.
- *Computer derived keywords*: Keywords extracted from the document by a computer program (for more details see the paragraph "Automatic generation of keywords").
- *Human-based assigned keywords*: Keywords assigned by a human for describing a document, either:
  - terms from a controlled vocabulary (or rather, according to Section 1.6.4: words from CVs made searchable as keywords). [27]
  - free terms conceptualized by the indexer.
- *Computer assigned keywords*: Keywords assigned by a computer using various algorithms. Similarly, to the category above, the keywords can be either:
  - keywords derived from a CV
  - free terms assigned by an algorithm.

This classification, therefore, takes into consideration two dimensions: the origin of the keyword (derived or assigned) and the typology of the agent (a human being or a machine) (Figure 4).

For both the assigned and the derived approaches the task is to find the best possible word(s)[28] to represent the contents of a document (from a certain perspective and for a certain

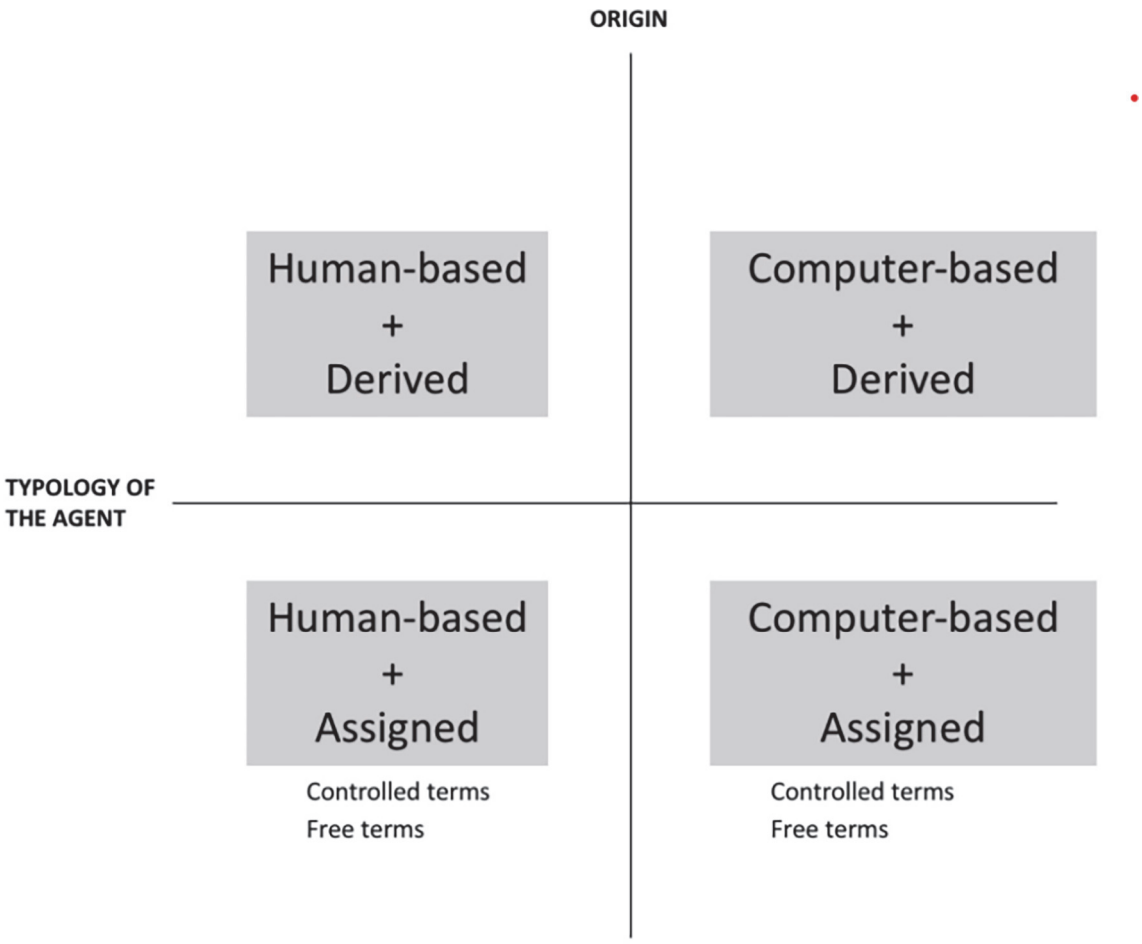


Figure 4. Classification of keywords.

purpose). The derived approaches are more restricted in that only terms appearing in the documents can be used. Assigned indexing using a CV are correspondingly restricted by the terms in the CV. This distinction is less important compared to the deep problem: To express the content of a document in a way that increases the findability of the document for those users that might benefit from using it.

Human-based keywords, regardless of being assigned or derived, can also be classified according to the identity of the actor (subject) that associate them to a document:

- Author keywords
- Indexer keywords
- Reader/third-part keywords

Holstrom (2019) further differentiated professional indexers, domain experts and casual indexers and expressed the view that each of these categories of indexers have different strengths and weaknesses (see further the discussion in Section 5).

4.0 Automatic generation of keywords

There is a huge literature on automatic keyword extraction, which will here only be presented very selectively and briefly (some such as KWIC -indexing were already presented in Section 2.1). In the literature, the main[29] approaches suggested (e.g., Siddiqi and Sharan 2015; Bharti et al. 2017) may be classified as (1) statistical approaches (2) linguistic approaches (3) machine learning approaches (4) domain specific approaches.[30] (In addition, there are hybrid approaches, and there is often strong overlap among the methods used in specific applications).

Automatic generation of keywords does not necessarily mean that keywords are automatically assigned to documents. In many cases, the generated keywords are only used as a suggestion for the user, that can use his human intellect to decide what keywords to accept and what others to discard (so-called "tag-recommendation"). See, for example, Subramaniyaswamy and Chenthur Pandian (2012).

4.1 Statistical approaches

Among the earliest work in this set of approaches was Luhn (1957) who suggested that a term found repeatedly in a document was appropriate as index term for that document. Luhn thus based automatic indexing on scores of in-document counts. A next step was taken by Edmundson and Wyllys (1961), who developed a "term significance formula" which, in addition to considering word frequency within a given document, also used a "reference" frequency, which measured words' relative frequency within a document collection. Spärck Jones (1972 and 1973) developed the term frequency–inverse document frequency (TF-IDF) score to calculate the importance of indexing terms. It is based on the idea that terms with a high frequency in a document, but which are rare in other documents have the highest value as indexing terms. The application of TF-IDF was further developed by Salton and Yang (1973) and Salton et al. (1975).

Whereas most approaches have been based on the "bag of word"[31] principle (disregarding word order and grammar), some approaches have utilized information of word placement in documents. The use of word co-occurrences for selection keywords or keyphrases takes one step in thus direction. See, for example, Bullinaria and Levy, 2007; Garg and Kumar, 2018 and Lancia 2008.

There are, however, more direct applications of document structure that have been applied for keyword extraction. Siddiqi and Sharan (2015) presented seven approaches (see endnote [32]) but unfortunately without references to further information. It seems obvious for future research to combine the field of genre studies (or composition studies) with empirical research on keyword selection.

4.2 Linguistic approaches

This set of approaches uses various linguistic analyses and tools for keyword detection and extraction, such as lexical analysis, syntactic analysis, and discourse analysis. Different types of words and word sequences do not have the same value as keywords. For instance, nouns and adjectives are more likely to appear in keywords than other kinds of words such as adverbs and determiners, as they tend to provide more information about a given text. Therefore, extraction tools may exploit morphological and syntactical features of textual units. Firoozeh et al. (2020, 272):

> Extraction methods rely not only on plain words (tokens) but also on their lemmatized forms, their parts of speech (POS tag), and some of their morphological features, such as gender or number (singular, plural). [...] When the extracted keywords are to be presented to human users, they are mainly retrained without any inflection so as to satisfy the citationess [sic] property.

See further about linguistic approaches in Jacquemin and Bourigault (2003).

4.3 Machine learning approaches

Machine learning is a kind of artificial intelligence. Often three forms of machine learning are distinguished: Supervised learning, semi-supervised learning, and unsupervised learning.

4.3.1 Supervised learning-based approaches

Supervised learning methods acquire knowledge from data that a human expert has explicitly annotated with category labels or structural information. Supervised learning requires a large amount of expensive training data. Witten et al. (1999) developed the algorithm *KEA* (keyphrase extraction algorithm) based on author-assigned keywords, which the algorithm learned from and had to demonstrate both on the same documents and on new documents which had not been assigned any keyphrases by their authors. What the algorithm learned was to extract phrases based on their features. Two features were calculated for each candidate phrase: (1) TF/IDF (the frequency of a phrase's use in a particular document compared with the frequency of that phrase in the corpus) (2) first occurrence (the number of words that precede the phrase's first appearance, divided by the number of words in the document. Based on these, and many more technical choices, the authors found that KEA on the average can match between one and two of the five keyphrases chosen by the authors in the collection. This was considered a good performance because KEA must choose from thousands of candidates and because it is highly unlikely that even another human would select the same set of phrases as the original author.

4.3.2 Semi-supervised learning-based approaches

These are approaches that mediate between supervised and unsupervised learning. Li et al. (2010), for example, proposed a semi-supervised keyphrase extraction approach, which explored title phrases as the source of knowledge, because phrases in the titles of documents are often appropriate as keyphrases. The position of a term has been a quite effective feature as a phrase extraction method (Witten et al. 1999). Therefore, terms located in the title are ranked higher. Li et al. used *Wikipedia* to compute the semantic relatedness between terms and thereby suggest keyphrases related to title terms.

4.3.3 Unsupervised learning-based approaches

These are approaches that do not require expert human annotation of examples: rather, an algorithm identifies its own keywords by clustering unlabeled examples into coherent groups. Many different unsupervised approaches have been applied for keyphrase extraction. Among them are graph-based approaches. An overview of these are given in Beliga et al. (2015). On p. 3 they explain:

> A graph is a mathematical model, which enables the exploration of the relationships and structural information very effectively. [...]. For now, in short, document is modelled as graph where terms (words) are represented by vertices (nodes) and their relations are represented by edges (links) [...]. The edge relation between words can be established on many principles exploiting different scopes of the text or relations among words for the graph's construction.

Among the relations mentioned are co-occurrence relations, syntax relations and semantic relations. The authors conclude:

> Graph-based methods for keyword extraction are simple and robust in many ways: (1) they do not require advanced linguistic knowledge or processing, (2) they are domain independent and (3) they are language independent. Such graph-based KE techniques are certainly applicable for various tasks: text classification, summarization, search, etc. Due to the aforementioned benefits it is reasonable to expect that graph-based extraction will attract the attention of the research community in the future. It can be expected that many text and document analyses will incorporate graph-based keyword extraction.

Some approaches use natural language processing technologies to derive keywords, for example, syntactic analyses of text. Kathait et al. (2017) selected nouns and adjectives and combined them with statistical methodologies. Witten et al. (1999) describe an algorithm, KEA, that identifies candidate keyphrases using lexical methods, calculates feature values for each candidate, and uses a machine-learning algorithm to predict which candidates are good keyphrases.

4.4 Domain specific approaches

Section 3.3.1 mentioned that supervised machine learning mostly requires domain knowledge in an initial state. A separate set of approaches was mentioned by Siddiqi and Sharan (2015, 19):

> Domain specific approaches: Various approaches can be applied to a specific domain corpus, which exploit the backend knowledge related to the domain (such as ontology) and inherent structure of that particular corpus to identify and extract keywords.

Frank et al. (1999) used a machine learning approach to extract keyphrases from sets of documents. They found that different keyphrases are used in different subject areas and that the exploitation of this domain-specific information significantly improves the quality of the extracted keyphrases. Gazendam et al. (2010) and Medelyan and Witten (2006) used domain-specific thesauri to improve the quality of automatic keyword extraction.

The literature on domain specific approaches is meager (and from this perspective it is somewhat surprising that Siddiqi and Sharan (2015) included it in their survey). From the perspective of the present authors, it seems rather obvious that keyword selection must be related to different "paradigms" or metatheoretical perspectives in fields of knowledge. It is common knowledge, at the least in the humanities, that histories of a given domain cannot be neutral, but are always made from a perspective (whether or not the author acknowledges it or is aware of it). The ways different histories of music, for example, select terms used for periodization, vary. Such knowledge must, to the degree that it can be formalized, be usable, also in relation to the automatic generation of keywords.

5.0 Keywords selected by different kinds of actors

Keywords may be, for example, author-generated, indexer-generated, editor generated, reader-generated, generated by algorithms (or be absent). Holstrom (2019) called the focus on who (subject or person) performed the key-word attribution for "an Actor-Based Model for Subject Indexing" and wrote:

> Four primary types of actor perform subject indexing work: 1) professional indexers, 2) domain experts [authors in the listing above], 3) casual indexers [readers in the listing above], and 4) machines [i.e., the programmers]. These subject indexing actors all have agency and all act on information objects.

Holstrom also discusses how these kinds of actors may be combined. In special libraries (e.g., the National Library of Medicine) the professional indexers often have an advanced degree in a special domain in addition to their training as information specialists or indexers. Other kinds of actors might be added, for example, fandoms, which are often extremely knowledgeable about the documents being indexed (e.g., in certain genres of fiction and computer games).

An example of an empirical investigation of the different actors' indexing is Lee et al. (2015), which is a study of full-text documents on Alzheimer's disease together with their indexing with MeSH terms in PubMed (MEDLINE) and citation patterns. This article found that different actors in this domain represented different views: MEDLINE indexers emphasize amyloid-related entities, including methodological terms, while authors focus on specific biomedical terms, including clinical syndromes. The complex networks of citing relationships to a certain extent reflected the impact of basic science discoveries in research on Alzheimer's disease.

A debate in library and information science is thus about the effectiveness of keywords assigned by different actors such as authors, indexers, taggers or cited papers (see Section 2.1 about Keywords Plus).

The following kinds of actors are discussed in this section: 5.1 Professional indexers; 5.2 Domain experts / authors; 5.3 Casual indexers (readers, users); 5.4 Machines.

5.1 Professional indexers

Professional indexers are persons who are employed for the task of indexing documents as one of their main tasks. They may be trained primarily in information science and knowledge organization, or they may be subject specialists with training in indexing (such a training may be a general indexing training from a LIS department, or it may be in-house training focusing on the specific system, e.g., as in the MEDLINE database[33]). (Indexing by subject specialists without employment or training in indexing is discussed in Section 5.2). Professional indexers mostly make use of a large, controlled vocabulary for assigning keywords to documents. It is extremely difficult to identify studies about the quality of indexing made by professional indexers, as Lancaster (2003, 88) wrote:

> Regrettably, not much research has been performed on the factors that are most likely to affect the quality of indexing. An attempt has been made to identify such factors in Figure 35 but this is based more on common sense or intuition than on hard evidence.

Lancaster's Figure 35 (89) classified factors affecting indexing quality in five groups:

- Indexer factors (e.g., subject knowledge and experience)
- Vocabulary factors (e.g., specificity/syntax and ambiguity or impression)
- Document factors (e.g., subject matter, complexity, and summarization)
- "Process factors" (e.g., type of indexing, rules and instructions)
- Environmental factors (e.g., heating/cooling, lighting, and noise)

Here we are discussing the professional indexer. One of the things we know for sure is that inter-indexer consistency is low, i.e., indexers perform rather differently (see, e.g., Lancaster 2003, chapter 5: "Consistency of Indexing"). It is remarkable that under "indexer factors" Lancaster did not mention anything about the indexers' theories. It seems important, for example, whether indexers have a document-oriented view or a request-oriented view on indexing (see Hjørland 2018, p. 612-3, §3.1 (https://www.isko.org/cyclo/indexing#3.1 and p. 619, §3.8.2 / https://www.isko.org/cyclo/indexing#3.8). From the domain-analytic perspective, the most important determinant of people's behavior is the "theories", that govern them ("theory" to be understood in the perspective of the so-called "theory theory"[34]).

Therefore, instead of supposing that "professional indexers" perform in one specific way, it is more fruitful to suggest that they perform differently based on, among other issues, how they have been trained and their specific theories and views on both indexing and the domain they are indexing, that have influenced them. Although Lancaster (2003) and others found that theories do not exist in indexing, this is, as discussed by Hjørland (2018), not true (and the relative absence of theory discussions in KO may cause lack of progress in this field).

It has often been assumed that human indexing in general, and professional indexing in particular, is the "gold standard" against which computer-based indexing should be measured. According to what has already been written, this is a problematic assumption because indexers perform very differently. Another issue is that it is complicated to determine which set of index terms is the best for a given purpose: This is not something that human indexers just know, but is, as emphasized by Swanson (1986)[35], a scientific hypothesis in need for examination.

5.2 Domain experts / authors

Authors of scientific/scholarly papers sometimes assign keywords to their own articles. In this case, we have highly competent domain experts, who are not professional indexers. Author keywords are, like the title and the abstract, a part of the paratext (Skare 2020), more precisely the peritext of journal articles. Author assigned keywords have, according to Hartley and Kostoff (2003, 434), been employed by many journals since at least 1975, in order to improve their indexing and to increase the probability of the article being retrieved and potentially cited. This employment seems to be increasing at the least in the biomedical domain (cf., Névéol, Doğan and Lu 2010). There is, however, a great variation between

journals in the use of such keywords. Some journals do not apply author keywords (but may instead provide keywords by editors or no keyword), some journals demand that such "keywords" should consist of index terms from a controlled vocabulary (*Journal of the Association for Information Science and Technology*, for example, requires authors to apply terms from the *ASIS&T Thesaurus of Information Science, Technology, and Librarianship* (Redmond-Neal and Hlava 2005). See further Hartley and Kostoff (2003) about different uses of author keywords among journals.

The number of terms that may be applied also varies much. Today, many scientific journals ask authors to provide a certain number of keywords when submitting their articles. Databases such as *Web of Science* and *Scopus* have a specific field for author keywords allowing studies of these.[36] Uddin and Khan (2016) used Scopus to investigate author keywords in the field of obesity and found that in this domain there was a significant relation between the number of author keywords assigned and the citation counts that the article obtained. The analysis of databases has shown that author keywords may take up different functions by describing different aspects of the document, such as "research topic", "research method", "research area", "data" or other (see e.g., Lu et al. 2019), with the indication of research topic as the most common function. Caragea et al. (2014, 1443) wrote:

> Hence, while we believe that authors are the best keyphrase annotators for their own work, there are cases when important keyphrases are overlooked or expressed in different ways, possibly due to the human subjective nature in choosing important keyphrases that describe a document.

Such subjectivity is, however, unavoidable, and it should be a trivial conclusion about all kinds of keyword annotations, whether they are done by human beings or by computers (programmed by human beings) that they represent different kinds of subjectivity. What seems to be important is to be able to characterize different kinds of subjectivity and utilize such knowledge for optimizing keyword assignment.

Névéol, Doğan and Lu (2010) is an important study of author-keywords in biomedical journal articles. It did not (as it is often done) consider MEDLINE indexing as a gold standard that author keywords are compared to. In contrast, their aim was to access whether topics covered by author keywords are also covered by MeSH indexing terms. In their sample of 14,398 articles, which included both author keywords and MEDLINE indexing terms, authors provided on average 5.3 (± 1.9) keywords while indexers assigned on average 13.0 (± 11.9) indexing terms. The article used Surgenor, Corwin and Clerico (2001[37]) as a running example in their article (see Table 3).

Névéol, Doğan and Lu (2010) found that in a sample of 300 pairs of author keywords and indexing terms:

- 36% of the author keywords were covered by MeSH but not selected by the MEDLINE indexers
- 16% were covered by MeSH but not yet linked
- 15% were covered by multiple MeSH headings
- 33% were not covered by MeSH

The authors wrote that from the MEDLINE indexing perspective, their results show that the majority (62%) of the author keywords are already covered by an exact or closely related MeSH indexing term in MEDLINE indexing. They also pointed out that because 49% of the author keywords are either not covered in MeSH (33%) or not linked to their equivalent in MeSH (16%), this suggests that author keywords may be helpful for MeSH terminology development.

The question is, of course, whether keywords suggested by authors, but missing in MeSH, or available in MeSH but not used in the indexing of the documents in the sample, are fruitful keywords, that should be included in MeSH and should have been used by MEDLINE indexers.

Névéol, Doğan and Lu (2010, 537) wrote:

> Although the assignment of keywords to an article by the author and the assignment of MeSH indexing terms by the indexer may seem like two very similar activities, there are significant differences in terms of form and perspective. Typically, authors are asked to choose a small number of keywords, without reference to a controlled vocabulary; whereas indexers are trained to select indexing terms from MeSH according to a specific protocol. Moreover, in addition to the subjectivity inherent in an indexing task [references omitted], authors are focused on selecting keywords representing what they consider as important to describe the content of their own article. In contrast, indexers consider the article in the larger scope of the collection.

But is this a true dilemma? Is the description of the content on an article in contrast to considering the article in the larger scope of the whole of the medical literature? We should have the working hypothesis that indexing is an activity governed by knowledge and that the nature and content of this knowledge is subject to research in information science and knowledge organization. The article can be read as if the authors find that author keywords often indicate failures in either MeSH or the indexing. But how do we find out when this is the case? We need to know about the indexing training and instructions of MEDLINE indexers. Unfortunately,

MEDLINE assigned terms (20 terms)	Author assigned keywords (4 terms)
Adult	
Aged	
Cohort Studies	
Decision Making	decision-making (semantic distance to MeSH measured to 0)
Diagnosis-Related Groups	
Female	
Health Services Accessibility	
Hospital Mortality	
Hospitals, Community / organization & administration*	
Hospitals, Rural / organization & administration*	rural health services (semantic distance to MeSH measured to 0.254)
Humans	
Intensive Care Units / statistics & numerical data*	
Length of Stay	
Male	
Middle Aged	
New Hampshire / epidemiology	
Outcome Assessment, Health Care	
Patient Transfer / statistics & numerical data*	interhospital transport (semantic distance to MeSH measured to 0.361)
Prospective Studies	
Survival Analysis*	survival analysis (semantic distance to MeSH = exact match)

Table 3. MeSH terms and author keywords assigned to Surgenor, Corwin and Clerico (2001).

their indexing manual is not available for the research community ("NLM Staff access only")[38]. Looking at Surgenor, Corwin and Clerico (2001) and Table 3 above might perhaps indicate that MEDLINE indexing is too mechanical and very week in the conceptualization of the article.[39]

There have also been critique of author keywords. Hahn et al. (2007) proposed the use of automatic procedures in order to avoid kinds of subjectivity in author-assigned keywords. Among their arguments was that "human indexers, even professional ones, are liable to error as well as to the possibility of intrinsic subjective bias [...] This is not to say that authors of a structured abstract would consciously cheat, but rather there is a grey area of overstatement and overestimation of one's own results in a highly competitive scientific environment". Yes, indeed, but the authors did not consider the problems inherent in the automatic procedures and their conclusion was based on a priori thinking rather than on empirical studies. Lok (2010, 418) also addresses the issue of author's role in providing more machine-readable information about their articles. They reported an experiment that they found was not encouraging. Authors used simple vocabulary that the curators of a protein database didn't accept and claimed: "Authors are not the right people to validate their own claims. The community – referees, editors, curators, readers at large – is still needed".

It is also relevant to notice that Peset (2020) found that most new keywords introduced by authors in research papers are never used a second time and only 11% of them "survive" more than 3 years after their first appearance.

Some further studies of author keywords are reported in the endnote 40[40].

5.3 Casual indexers (readers, users)

Reader-assigned keywords are provided by the readers of documents, who can have various ages and backgrounds. Some have more general knowledge about the domain of the studied document, whereas others may not be very familiar with the domain. As stated in Section 5.0 fandoms are often extremely knowledgeable about the documents being indexed, but such indexing is mostly limited to popular genres of, for example, fiction and computer games.

Readers are often given some guidelines for annotating documents, but the reader-assigned keywords may nevertheless express a special kind of subjectivity, e.g., make comments of a private character of little usefulness for other readers.

5.4 Machines

We have already presented kinds of automatically produced indexes (Section 2.1) and automatic generation of keywords (Section 4). The general view among computer scientists is, implicitly as well as explicitly, that computer approaches are both more efficient and more effective than human indexing/keyword assignment. Gerard Salton, for example, wrote (1996, 333):

> ... forgetting all the accumulated evidence and test data, and acting as if we were stuck in the nineteenth century with controlled vocabularies, thesaurus control, and all the attendant miseries, will surely not contribute to a proper understanding and appreciation of the modern information science field.

Robertson (2008) said: "statistical approaches won, simply. They were overwhelmingly more successful [compared to other approaches such as thesauri]."

There are voices in library and information science claiming superiority of human indexers (e.g., Day 2014[41] and Warner 2019). Their arguments are not convincing, however, because they just consider humans in possession of a fundamental ability to index documents without considering the nature of knowledge about both indexing and the domain of the documents being indexed.

Hjørland (2011) criticized the dichotomy between human-based indexing versus computer-based indexing because both sides of the dichotomy may be governed by different "theories" of indexing. Humans may, for example, perform in a very computer-like manner if they follow a strict set of learned indexing rules or mechanically highlight terms in a text. What seems important is therefore to examine the theoretical foundations on which humans as well as computers work (e.g., if they are governed by document-oriented or policy-oriented indexing, cf., Hjørland 2017, §2.4). Another issue is that computers often depend on input which is created by humans. In bibliometric techniques, algorithms work on citation patterns, but such patterns consist of human decisions on which papers to cite. This also indicates that a strict distinction between human-based approaches and computer-based approaches is too simplified.

6.0 Conclusion

Two main theories on keywords and keyphrases are:

(1) Keywords are meant to extract "the essence" of the documents they characterize (see, e.g., Caragea et al. 2014, 1435). This can be called *the essentialist theory*.
(2) Keywords are assigned to or derived from documents to facilitate the findability of certain information according to the purpose of the indexing institution. This can be called *the policy-based theory*. This theory implies that things and documents have no "essence", and by implication there is not one correct way in which a set of keywords can characterize a document.

Other theories exist, for example user-oriented theories, but these theories are considered relatively unimportant and are ignored here.

The choice between the essentialist theory and the policy-based theory is seldom openly presented or discussed. Nonetheless, such theories have great importance for both manual and automated indexing as well as for the evaluating of indexing. From the viewpoint of essentialist indexing (extracting keyphrases), Caragea et al. (2014, 1435) wrote:

> The reason for not considering the entire text of a paper is that scientific papers contain details, e.g., discussion of results, experimental design, notation, that do not provide additional benefits for extracting keyphrases. Hence, [...] we did not use the entire text of a paper. However, extracting keyphrases from sections such as "introduction" or "conclusion" needs further attention.

This quote can be interpreted as indicating that the essence of a document is best expressed in the introduction and conclusion. Against this view it can be said that from, for example, the perspective of so-called "evidence-based practice", exactly the methodology section in documents is of utmost importance. We saw, for example, in Table 3 (Section 5.2) how MEDLINE indexing made high priority to methodological issues in indexing.

Closely related to the issue of content-oriented versus policy-based indexing is the issue of what kind of information should be considered in the derivation or assignment of keywords to documents. The ISO 5963:1985 standard[42] is typically document oriented as it only includes examining title, table of contents, introduction and so on – and even fails to mention the list of references. It is also implicitly assumed by this standard that the same set of indexing terms are optimal for all kinds or queries.

Conflicting with the view that a document itself is all that needs to be considered is, for example, the view repre-

sented in the information retrieval tradition, where terms in a document are related to all terms in a collection. This is, however, only a first step. A second step is taken by Morris (2010, 148), who suggested:

> The fact that between approximately 30%–40% of the interpretation of lexical cohesion in the texts were individually different provides compelling evidence in support of returning to a computational view of text meaning that includes text, reader, and writer.

We saw another example in Section 5.2 where Lok (2010, 418) suggested: "Authors are not the right people to validate their own claims. The community – referees, editors, curators, readers at large – is still needed". This quote also indicates the need for a broader interpretation of documents.

A third step can be illuminated by the philosophy of Hegel (here quoted from Mácha 2015, 19):

> Hegel was indeed an adherent of the doctrine of internal relations. He writes in his *Logic*: 'Everything that exists stands in correlation, and this correlation is the veritable nature of every existence.' [Hegel 1968, p. 235.] To *adequately* understand the veritable nature (i.e., the essence) of every single thing, one has to understand its relations to every other thing and, in the end, to the whole, to the Absolute. To put the doctrine in negative terms: we cannot isolate or abstract one single thing out of the whole and understand it *adequately* in isolation. [Cf. Kain 2005, pp. 4–6].

Translated to keyword assignment this means that keywords cannot be assigned by considering a document in isolation. It must be assigned by considering the tradition or paradigm in which the document is written, as well as the interests that the indexing is meant to fulfill.

Acknowledgements

The authors thank two referees for their valuable feedback to a former version of this article and Claudio Gnoli, who served as editor as Birger Hjørland was included as a co-author, for great work improving the manuscript.

Notes

1. *Wikipedia* (2020) wrote: "An index term, subject term, subject heading, or descriptor, in information retrieval, is a term that captures the essence of the topic of a document. Index terms make up a controlled vocabulary for use in bibliographic records. They are an integral part of bibliographic control, which is the function by which libraries collect, organize and disseminate documents. They are used as keywords to retrieve documents in an information system, for instance, a catalog or a search engine. A popular form of keywords on the web are tags which are directly visible and can be assigned by non-experts. Index terms can consist of a word, phrase, or alphanumerical term. They are created by analyzing the document either manually with subject indexing or automatically with automatic indexing or more sophisticated methods of keyword extraction. Index terms can either come from a controlled vocabulary or be freely assigned."
2. In the former DIALOG system, for example, the names of journals were phrase indexed, and could not be searched by single words in the name, but titles of articles were both 'word indexed' and 'phrase indexed'.
3. Mooers and Mooers (1993) claimed that 'key words' are the direct descendants of Mortimer Taube's Uniterms, but according to the *Oxford English Dictionary* (2020a) it goes back to 1827. However, Mooers and Mooers may have had a more specific use in mind.
4. Concerning the concept "subject access point" see Hjørland and Kyllesbech Nielsen (2001).
5. Larson (1991, 211): "Experience in catalog use may not necessarily imply that users have been "conditioned" to avoid subject searches [searches using Library of Congress Subject Headings, not searches using title keywords], though such conditioning appears to be a likely result of gaining experience in catalog use, whether card or online catalog. We would suggest, as a hypothesis for further study, that individual users' experiences of subject search failure and information overload lead them to reduce their use of the subject index and to increase their use of alternate means of subject access, such as title keyword searching and shelf browsing following a known item search." See Gross and Taylor (2005) for a response to Larson (1991) defending subject headings: "It was found that more than one-third of records retrieved by successful keyword searches would be lost if subject headings were not present, and many individual cases exist in which 80, 90, and even 100 percent of the retrieved records would not be retrieved in the absence of subject headings."
6. Alternatively Soergel (1974, 31) defined: "A 'subject heading' is a specific type of descriptor, namely, a descriptor used in an alphabetical subject catalog or printed index". Soergel (1974, 126): In subject headings, "[t]he main headings and the subheadings are independent elements designed by terms and arranged alphabetically (with minor deviations in some catalogs). Many subject headings are created by combining a main heading with a subheading, and for those headings the place in the arrangement is determined by its components. The citation order is fixed: Main heading—sub-

heading. A closer look reveals a more complex situation".

7. ISO 5127:2017: "Preferred term, descriptor. *Term* (3.1.5.25) used to represent a *concept* (3.1.1.02) when *indexing* (3.8.2.01)".
8. ISO 5127:2017: "3.7.3.04 subject heading: heading (2) <*access point*>" (3.7.3.01) expressing an aspect of the contents of all or part of a *document* (3.1.1.38) and also used to collocate *entries* (3.2.1.32) for documents having the same or similar content. Note: See also *keyword* (3.8.1.07); *content descriptor* (3.8.3.19)".
9. The date 1950 is debatable. Lancaster (1968) said that in 1947 Calvin Mooers began using the term *descriptor* on a system called Zato 1 for documents subject classification as of words extracted from their own texts. Roberts (1984) also suggests that the term descriptor was coined in 1947 and explains the difficulties in confirming this information.
10. Modern thesauri tend to include many more descriptors. The Thesaurus of Psychological Index Terms, for example, contains about 8,000 terms, including preferred and entry terms. Source: https://www.apa.org/pubs/databases/training/thesaurus
11. The PsycINFO database, for example, wrote: "What Are Index Terms? Index terms are controlled vocabulary terms used in database records to make searching easier and more successful. By standardizing the words or phrases used to represent concepts, you don't need to try and figure out all the ways different authors could refer to the same concept. Each record in APA's databases contains controlled vocabulary terms from the Thesaurus of Psychological Index Terms." https://www.apa.org/pubs/databases/training/thesaurus
12. Soergel (1974, 31-4) defined descriptors differently: "A 'subject heading' is a specific type of descriptor, namely, a descriptor used in an alphabetical subject catalog or printed index. Some people use 'descriptor' only in connection with ISAR [information storage and retrieval] systems using combinational indexing (usually implemented through edge-notched cards, peek-a-boo cards or computerized methods). Descriptors in this usage represent predominantly elemental concepts, whereas subject headings represent predominantly compound concepts. While this usage corresponds to the original definition of 'descriptor' it has not gained currency in the profession. A 'class number' is another specific type of descriptor, namely the notation from a classification system such as the Dewey Decimal Classification or the Library of Congress Classification."
13. A phrase is sometimes called "a multiword", "a syntagmatic term" or "a complex term" (Lauriston 1994, 149).
14. Anderson and Pérez-Carballo (2005, 1.61, 19) wrote: "'Keyword' is often used to indicate the more important free-text terms". However, on p. 222 (§12.75) another meaning is suggested, "keyword searches using *Library of Congress subject headings*: A third way of searching is the 'keyword' approach. It should be invoked by the intelligent OPAC in several instances. In this first example, a searcher has used the 'exact' approach to 'jazz music' but wants more options. The next step would be to apply the 'keyword-in-main-heading approach'" (retrieving all subject headings which include 'jazz'. The authors add: "in this case, 'music' and 'jazz' are treated as independent keywords.
15. Wellisch (1995, 248) distinguishes two senses of "keyword": (1) the first word of a heading, often called 'lead term'; (2) "a significant term word taken from the title or the text of a document and used in a heading (but not necessarily the first word in it)". He writes (2501) that in the KWOC (Key Word Out of Context) index, the two senses of *keyword* were here conflated so that a keyword serving as an index term was at the same time also used as a lead term. On p. 252-3 Wellisch introduces a third meaning: Author assigned keywords. Wellisch claims that the KWIC was not first proposed by Hans Peter Luhn, but by Andrea Crestadoro in 1858. However, Wellisch has no reference to Crestadoro in the bibliography, and he probably meant Crestadoro 1856 (as cited by Olson and Boll 2001, 112). But Olson and Boll wrote the opposite of Wellisch: "Andrea Crestadoro invented the basic concept that topics should be described by a standardized vocabulary". Therefore, these two sources are in conflict. It will not be examined here which is right.
16. Broughton (2011, 262) did not define "keyword", but implicitly understood it in the same broad sense by defining "keyword list: a list of words, or descriptors, to be used for indexing documents. The keyword list is usually characterized by a lack of structure in its composition, and does not have cross-references between terms. It may be no more than an alphabetical list of terms taken from documents or previously used in indexing".
17. East Carolina University Libraries. *Search Basics for the Health Sciences: Keywords vs. Subject Headings*. Retrieved from: https://web.archive.org/web/20201020200459/https://libguides.ecu.edu/c.php?g=89808&p=579367

    "Keywords are:
    - natural language terms that describe your topic
    - able to be combined in any number of ways
    - lacking consistency in usage, definition, and sometimes spelling (e.g., GERD vs. GORD[U.K.])
    - either single words or phrases
    - used to search for matching words or phrases anywhere in the records the database contains (such as title, abstract, journal title)

- used when no appropriate subject heading exists as an equivalent
- sometimes either too broad or too narrow, resulting in either too many or too few results
- reflective of recent phenomena in advance of when the subject headings are added

Subject Headings are:

- "controlled" vocabulary used by an organization (e.g. the National Library of Medicine) to describe the concepts in the literature collected by that organization or database (such as MEDLINE or CINAHL).
- consistent in their definition across the records in the database.
- less flexible and must be chosen from the thesaurus used by the database; if the incorrect subject heading is selected, none of the results will be relevant.
- only searched for in the subject heading field of the record.
- helpful for retrieving a set of articles with fewer irrelevant results
- slow to change--this means that the most recent changes in knowledge--on diseases, drugs, devices, procedures, concepts--may not be reflected in the controlled vocabulary".

18. Our understanding of keyword is thus different from the traditional approach, which considers a keyword an (intended) objective description of a document, cf., the definition by Feather and Sturges (1996, 341) quoted above. Also, when Firoozeh et al. (2020) claim: "The terms 'keyword' and 'keyphrase' do not refer to any theory," we are here arguing that two opposing theories are in play: (1) the traditional view of one correct set of keywords for a given document (2) the alternative theory that different perspectives and interests requires different sets of keywords.
19. A concordance consists of a list of the words in the text with a short section of the context that precedes and follows each word. Although a concordance is often defined as "an alphabetical list of the principal words used in a book or body of work, listing every instance of each word with its immediate context", where "principal words" could be understood as synonym for keywords, in reality a concordance includes many more words than key-word indexes (usually all words except words from a stop list). See however endnote 23 about KWIC and related indexes.
20. Library of Congress (2006, 14) wrote: "With post-coordinate indexing, subject concepts are entered as single terms so that users are required to coordinate them. Boolean searching and other advanced techniques are required in order to locate resources on the compound and/or complex subjects in which the searchers are interested. False associations may easily occur because relationships among terms can be unclear". Slide 13 presented the following examples of pre-coordination:

- Gold mining—United States—History—19th century
- Diamond mining-South Africa—History—20th century

"In pre-coordinated indexing, appropriate terms are chosen and coordinated into subject-subdivision combinations at the time of indexing or cataloging. On the screen you'll see a couple of examples of pre-coordinated strings.

You'll see **Gold mining—United States—History—19th century**. What we have here is a topical subject heading followed by a geographic subdivision, a topical subdivision (History), and a chronological period (19th century). In the second, it is the same structure, but with different components in that particular heading string. In each of these cases, what you get from the pre-coordinated string is a certain amount of context."

Slide 14 presented the following examples of post-coordination in indexing:

- Gold mining
- Diamond mining
- United States
- South Africa
- History
- 19th century
- 20th century"

P. 14: "With post-coordinate indexing, subject concepts are entered as single terms so that users are required to coordinate them. Boolean searching and other advanced techniques are required in order to locate resources on the compound and/or complex subjects in which the searchers are interested. False associations may easily occur because relationships among terms can be unclear. If you look at the slide, you can see an example of what happens when terms are post-coordinated instead of being used in pre-coordinated strings. In contrast to the previous slide, in which it was clear that the resource was about gold mining in the United States in the 19th century, in this example it is unclear whether it's about gold mining in the United States in the 19th century, or about diamond mining in South Africa in the 19th century, or diamond mining in the United States. And what century is it? This is how false drops happen. A post-coordinated system often causes a significant number of false drops, because the context is missing. One might never know exactly what that work is about, without the strings." Another example is that a search after "alcoholism in women", us-

ing #alcoholism and #women as keywords, may retrieve "false drops" such as "alcoholism in men and its implications for their relations to women".

21. For example, Lok (2010, 416-7) reported about Elsevier's system *Reflect*, which automatically recognizes and highlights the names of genes, proteins and small molecules in articles in the journal *Cell*.
22. Luhn is normally given the honor for the idea of the KWIC index although indexes based on the a related idea have occurred earlier.
23. According to the discussion in Section 1 and the note on concordances in Section 2, we have now the paradox, that the "keywords" in, among other, KWIC-indexes are not really "keywords" because they are not selected (but contain all terms in the title/document except those deselected by the list of stop words). "KWIC index" should therefore, strictly speaking, be termed "term index in context".
24. Lancaster (2003, 52-3; italics in original) explains the terms *cycling* and *rotation* and wrote: "*Rotation* is essentially the same as cycling except that the entry term is highlighted in some way (e.g., italicized or underlined) rather than being moved to the leftmost position [examples follow]." The term *Permuterm index* was used (and trademarked) for a kind of index used in the *Science Citation Index* and its family of related products, cf., Garfield (1976), who argued that permutations are different from cyclings and wrote about the drawbacks of KWIC indexes.
25. Some sources claim that KWAC index "provides for the enrichment of the keywords of the title with additional significant words taken either from the abstract [o]f the document or its contents." See, for example, https://www.librarianshipstudies.com/2017/02/keyword-augmented-in-context-kwac.html and http://inmyown-terms.com/kwic-kwac-kwoc-not-knock-knock-joke/
26. KeyWords Plus on *Clarivate Analytics'* home page: https://support.clarivate.com/ScientificandAcademicResearch/s/article/KeyWords-Plus-generation-creation-and-changes?language=en_US
27. In the literature it is often assumed that assigned indexing (whether human or algorithmic) is always based on a controlled vocabulary (see e.g., Beliga et al 2015, 2), but this is not the case. Humans may, for example, assign words from their own associations, and machines may, for example, chose words from the bibliographic network of the document, a la "concept symbols" of Small (1978). (Somewhat related to Keywordplus®, cf., Garfield and Sher 1993, although this particular method is derived from words in the references of the document and thus formally not a form of assigned indexing).
28. Strictly speaking not just words can be used to represent the contents of documents, but also, for example, classification codes and pictures (Jacob and Shaw 1996).
29. The following classification is not exhaustive. In addition to the main approaches, other approaches exist and new ones may be invented.
30. Beliga (2015, 2) suggested a somewhat different classification in which statistical, linguistic and other approaches were considered subcategories of unsupervised machine learning. This seems to explain that many of, for example, the statistical techniques are often used in papers about unsupervised learning.
31. The term "bag of words" was used by Harris (1954), who wrote "... language is not merely a bag of words..."
32. Siddiqi and Sharan (2015, 22; dotted listing added) presented the following approaches to keyword selection based on location in the document:
    - First N terms: Only the first N terms from the document are selected. The logic is that the important keyphrases are found in the beginning of the document as generally important information is put at the beginning.
    - Last N terms: Only the last N terms of the document are selected. The logic is that the most important keyphrases are found in the last part of the document since important keyphrases are found in their concluding parts of the document
    - At the beginning of a paragraph: It weights terms according to their relative position in a paragraph. The logic is that the important keyphrases are likely to be found near to the beginning of paragraphs.
    - At the end of its paragraph: Weights a term according to its relative position in its paragraph. The logic is that the important keyphrases are likely to be found near to the end of paragraphs.
    - Resemblance to title: Rates a term according to the similarity of its sentence with the title of the article. Phrases similar to the title will have a higher score.
    - Maximal section headline importance: Rates a term according to its most important presence in a section or headline of the article. It is known that some parts of papers are more important from the aspect of presence of keyphrases such as abstract, introduction and conclusions.
    - Accumulative section headline importance: It is very similar to the previous one but it weights a term according to all its presences in important sections or headlines of the article."
33. National Library of Medicine (2018): "Most MEDLINE indexers are either Federal employees or employees of firms that have contracts with NLM for biomedical indexing. A prospective indexer must have no less than a bachelor's degree in a biomedical science. A read-

ing knowledge of certain modern foreign languages is typically sought. An increasing number of recent recruits hold advanced degrees in biomedical sciences. Federal employees must be United States citizens, but citizenship is not mandatory for contractors. Indexers are trained in principles of MEDLINE indexing, using the Medical Subject Headings (MeSH) controlled vocabulary as part of individualized training. The initial part of the training is based on an online training module (partially available to the public at http://www.nlm.nih.gov/bsd/indexing/index.html), followed by a period of practice indexing. NLM does not accept other indexing training programs as a substitute. About 1% of MEDLINE indexing is performed by indexers at the International MEDLARS Centers in Sweden and Brazil." (Accessed October 9, 2020).

34. Concerning the concept "theory" see Hjørland (2015), who suggested the following definition (116-7): "A theory is an explicit or implicit statement or conception that might be questioned (and thus met with an alternative theory), which is more or less substantiated and dependent on other theories (including background assumptions). We use the term theory about a statement or conception when we want to emphasize that it might be wrong, biased, bad or insufficient for its intended use and therefore should be considered and perhaps replaced by another theory."
35. Swanson wrote about searching, but his argument is equally relevant for indexing. He wrote that if people working on a task search the literature, some relevant documents may not be retrieved because any search strategy is a theory that may be refused but can never be finally proved. A search strategy is refused if it is possible to discover just one relevant document which has been missed by that strategy. Swanson concluded (1986, 114): "Any search function is necessarily no more than a conjecture and must remain so forever".
36. However, terms indexed as author-keywords seem to be confused with editor-assigned keywords and terms from CVs. A search for "information" as author-keyword in WoS provided a list of journals. Some of these journals do use author-assigned keywords (e.g., *Journal of Documentation*), on the other hand *Journal of the Association for Information Science and Technology* demands that authors must select terms from *ASIS&T Thesaurus of Information Science, Technology, and Librarianship* (author-selected, but not free terms). Finally, the journal *Knowledge Organization* uses editor-selected keywords (cf., Smiraglia 2013), which the systems apparently cannot distinguish from author-assigned keywords.
37. Névéol, Doğan and Lu (2010) wrote "Surgenor et al. 2009", but the correct printing year for this reference is 2001.
38. https://web.archive.org/web/20201022102233/https://www.nlm.nih.gov/bsd/indexing/training/INT_030.html (accessed October 22, 2020): Indexing Manual (*NLM Staff access only*) The Indexing Manual provides discussion on all aspects of indexing at NLM. The Indexing Manual is provided to all new indexers and is available online to NLM staff. The online version contains a search feature. Technical Memoranda (*NLM Staff access only*) Technical Memoranda are updates to the Indexing Manual that provide further clarification of a specific issue or address an immediate need. Technical Memoranda are distributed online and on paper to NLM staff.
39. A main use of Surgenor, Corwin and Clerico (2001) is for decision making in hospital planning, more precisely regionalization of hospital services. This is far better expressed by the author keywords, although a term such as "regionalization of hospital services" (or just "regionalization") is missed. The MEDLINE indexing is extremely individual oriented and seems to fail to connect the document with purpose for which is was written (decision making concerning regionalization of hospital services).
40. A study by Strader (2009) compared the usage of author-supplied keywords to the Library of Congress Subject Headings (LCSH), using a sample of electronic theses and dissertations from the Ohio State University, each of them having both authors keywords and LCSH headings. The results show that a large part of authors assigned keywords don't match with LCSH headings, suggesting that keywords are useful as an additional source of information, even though part of them overlap with the content of the abstract. The author suggests that such a big lack of matching between keywords and LCSH can be explained by the fact that a controlled vocabulary cannot be always updated with the popular terms in current research. Therefore, it seems that authors keywords and controlled vocabularies are complementary, and both should be adopted, as a valuable point of access for users and cataloguers. Similar results have been found in a study by Schwing, McCutcheon and Maurer (2012), that replicated the analysis of Strader, on a different sample of electronic theses and dissertations. A partial matching between authors keywords and LCSH emerged, suggesting a complementarity of them, as well as an overlapping of author supplied keywords, title and abstract. Such overlapping may reduce the usefulness of them as original and unique access points. Complementarity between subject headings and keywords has also been reported by Gross and Taylor (2005), McCutcheon (2009) and Gil-Leiva and Alonso Arroyo (2007). Smiraglia (2013, 158) criticized using author keywords list to improve retrieval. He ini-

tially suggested (p. 155) that every word in an article is a keyword but realized that for some persons they mean "a list of author-contrived 'keywords' underneath the abstract" as common in many journals. He wrote: "I do not think lists of author-contrived keywords are useful." His editorial concluded (158): "The potential use of keywords for retrieval and indexing seems clear. That is, the presence of keywords, whether in a separate list or in their usual place in the text, has the potential to influence the formal indexing of research, and also to influence resource location or selection by researchers." And "What is less clear is how those keywords should be generated. Empirical extraction of the terms is most accurate and therefore most reliable for indexing, retrieval or just for text analysis. Should editorial policy change to incorporate the use of formal keywords in Knowledge Organization it would make the best sense to generate the terms empirically, using text analysis tools designed for statistical term extraction." It can be added that from the next volume (41, 2014) *Knowledge Organization* assigns keywords to all articles (but not by the author but by the editor), whereas *Journal of Documentation*, for example, uses free author-assigned keywords and *Journal of the Association for Information Science and Technology* uses author-assigned descriptors from the ASIS&T Thesaurus of Information Science, Technology, and Librarianship (see above; Redmond-Neal and Hlava 2005). Smiraglia points out that keywords are a core tool in information retrieval, but we should extract them from the actual text (including title or abstract) instead of providing a separate list. He also expresses concerns about the danger of keyword lists influencing the decision-making process of the indexers:

"The potential use of keywords for retrieval and indexing seems clear. That is, the presence of keywords, whether in a separate list or in their usual place in the text, has the potential to influence the formal indexing of research, and also to influence resource location or selection by researchers.

What is less clear is how those keywords should be generated. Empirical extraction of the terms is most accurate and therefore most reliable for indexing, retrieval or just for text analysis. Should editorial policy change to incorporate the use of formal keywords in *Knowledge Organization* it would make the best sense to generate the terms empirically, using text analysis tools designed for statistical term extraction."

41. Day (2014, 7): "Human indexes are what machine algorithms strive towards by the use of various syntactical and semantic techniques and technologies". Day's book has itself an index, but the present author has found it necessary to use Google to find content in this book that could not be found via the index.
42. https://www.iso.org/standard/12158.html wrote on October 25, 2020: "THIS STANDARD WAS LAST REVIEWED AND CONFIRMED IN 2020. THEREFORE THIS VERSION REMAINS CURRENT."

References

Adler, Melissa. 2009. "Transcending Library Catalogs: A Comparative Study of Controlled Terms in Library of Congress Subject Headings and User-generated Tags in LibraryThing for Transgender Books." *Journal of Web Librarianship*, 3(4): 309-31.

Anderson, James D. and José Pérez-Carballo. 2005. *Information Retrieval Design: Principles and Options for Information Description, Organization, Display and Access in Information Retrieval Databases, Digital Libraries, Catalogs, and Indexes*. St. Petersburg, Fl: Ometeca Institute.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval. The Concepts and Technology Behind Search*, 2nd ed. New York: Addison Wesley.

Bekhuis, Tanja. 2015. "Keywords, Discoverability, and Impact." *Journal of the Medical Library Association* (JMLA) 103(3): 119-20. doi: http://dx.doi.org/10.3163/1536-5050.103.3.002

Beliga, Slobodan, Ana Meštrović and Sanda Martinčić-Ipšić. 2015. "An Overview of Graph-Based Keyword Extraction Methods and Approaches." *Journal of Information and Organizational Science* 39(1): 1-20. https://jios.foi.hr/index.php/jios/article/view/938

Bernier, Charles L. 1980. "Subject Indexes." In *Encyclopedia of Library and Information Science*, edited by Allen Kent, Harold Lancour and Jay E. Daily New York, NY: Marcel Dekker, Volume 29, 191-205.

Bharti, Santosh Kumar, Korra Sathya Babu and Sanjay Kumar Jena. 2017. "Automatic Keyword Extraction for Text Summarization: A Survey." https://arxiv.org/ftp/arxiv/papers/1704/1704.03242.pdf

Broughton, Vanda. 2011. *Essential Library of Congress Subject Headings*. London: Facet. doi:10.29085/9781783300365

Bullinaria, John A. and Joseph P. Levy. 2007. "Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study." *Behavior Research Methods* 39(3): 510-26. https://link-springer-com.ep.fjernadgang.kb.dk/content/pdf/10.3758%2FBF03193020.pdf

Caragea, Cornelia, Florin Bulgarov, Andreea Godea and Sujatha Das Gollapalli. 2014. "Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), October 25-29, 2014, Doha, Qatar, edited by Alessandro Moschitti, Bo Pang and Walter Daelemans. Stroudsburg, PA: Association for Computational Linguistics, 1435–46. DOI: 10.3115/v1/D14-1150

Chu, Heting. 2001. "Research in Image Indexing and Retrieval as Reflected in the Literature." *Journal of the American Society for Information Science and Technology* 52(12): 1011-18. https://doi.org/10.1002/asi.1153

Cleverdon Cyril W. 1962. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. *Aslib-Cranfield Research Project*. Cranfield, UK.: College of Aeronautics.

Cohen, Jonathan D. 1995. "Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting." *Journal of the American Society for Information Science* 46(3): 162-74.

Costello, John C. 1961. "Uniterm Indexing Principles, Problems and Solutions." *American Documentation* 12(1): 20-6.

Crestadoro, Andrea. 1856. *The Art of Making Catalogues of Libraries*. London: The Literary, Scientific and Artistic Reference Office.

Day, Ronald E. 2014. *Indexing It All: The Subject in the Age of Documentation, Information, and Data*. Cambridge, MA: MIT Press.

Dolamic, Ljiljana and Jacques Savoy. 2010. "When Stopword Lists Make the Difference." *Journal of the American Society for Information Science and Technology* 61(1): 200-3.

Edmundson, Harold P. and Ronald E. Wyllys. 1961. "Automatic Abstracting and Indexing-Survey and Recommendations." *Communications of the ACM* 4(5): 226-34. https://doi.org/10.1145/366532.366545

Feather, John and Paul Sturges (Eds.). 1996. *International Encyclopedia of Information and Library Science*. London: Routledge.

Feinberg, Hilda. 1973. *Title Derivative Indexing Techniques: A Comparative Study*. Metuchen, NJ: Scarecrow Press.

Firoozeh, Nazanin, Adeline Nazarenko, Fabrice Alizon and Béatrice Daille. 2020. "Keyword Extraction: Issues and Methods." *Natural Language Engineering* 26(3): 259–91. doi:10.1017/S1351324919000457

Frank, Elbe, Gordon W. Paynter, Ian H. Witten, Carl Gutwin and Craig G. Nevill-Manning. 1999. "Domain-Specific Keyphrase Extraction." *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (IJCAI-99), Stockholm, Sweden, July 31-August 6, 1999. San Francisco, CA: Morgan Kaufmann, Vol. 1: 668-73.

Furner, Jonathan. 2010. "Folksonomies". In *Encyclopedia of Library and Information Sciences*, 3rd ed., edited by Marcia J. Bates and Mary Niles Maack. Boca Raton, FL: CRC Press, Vol. III: 1858-66. doi: 10.1081/E-ELIS3-120043238.

Garfield, Eugene. 1976. "Permuterm Subject Index: Autobiographical Review." *Journal of the American Society for Information Science* 27(5-6): 288-91. Also available at: http://www.garfield.library.upenn.edu/essays/v3p070y1977-78.pdf

Garfield, Eugene 1990a. "Keywords Plus®: ISI's Breakthrough Retrieval Method. Part 1. Expanding your Searching Power on Current Contents on Diskette." *Current Contents* 1(3)2: 5–9. http://www.garfield.library.upenn.edu/essays/v13p295y1990.pdf

Garfield, Eugene. 1990b. "Keywords PlusTM Takes You Beyond Title Words. Part 2. Expanded Journal Coverage for Current Contents on Diskette, Includes Social and Behavioral Sciences." *Current Contents* 1(33): 5-9. http://www.garfield.library.upenn.edu/essays/v13p300y1990.pdf

Garfield, Eugene and Irving H. Sher. 1993. "KeyWords Plus - Algorithmic Derivative Indexing." *Journal of the American Society for Information Science* 44(5): 298-99. http://www.garfield.library.upenn.edu/papers/jasis44(5)p298y1993.html

Garg, Muskan and Mukesh Kumar. 2018. "Identifying Influential Segments from Word Co-Occurrence Networks using AHP [Analytic Hierarchy Process]." *Cognitive Systems Research* 47: 28-41.

Gazendam, Luit, Christian Wartena and Rogier Brussee. 2010. "Thesaurus Based Term Ranking for Keyword Extraction." *Workshops on Database and Expert Systems Applications*, Bilbao, 2010, 49-53. doi: 10.1109/DEXA.2010.31

Gil-Leiva, Isidoro and Adolfo Alonso-Arroyo. 2007. "Keywords Given by Authors of Scientific Articles in Database Descriptors." *Journal of the American Society for Information Science and Technology* 58(8): 1175-87.

Gross, Tina and Arlene G. Taylor. 2005. "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results." *College & Research Libraries* 66(3): 212-30.

Hahn, Udo, Joachim Wermter, Rainer Blasczyk and Peter A. Horn. 2007. "Text Mining: Powering the Database Revolution." *Nature* 448(7150): 130.

Harris, Zellig. 1954. "Distributional Structure." *Word* 10(2/3): 146–62. doi:10.1080/00437956.1954.11659520

Hartley, James and Ronald N. Kostoff. 2003. "How Useful are 'Key Words' in Scientific Journals?" *Journal of Information Science* 29(5): 433-8.

Hegel, Georg Wilhelm Friedrich. 1968. *The Logic of Hegel*. Trans. William Wallace. Oxford: Oxford University Press.

Hjørland, Birger. 2011. "The Importance of Theories of Knowledge: Indexing and Information Retrieval as an

Example." *Journal of the American Society for Information Science and Technology* 62(1): 72-7.

Hjørland, Birger. 2015. "Theories are Knowledge Organizing Systems (KOS)." *Knowledge Organization* 42(2): 113-28.

Hjørland, Birger. 2017. "Subject (of Documents)." *Knowledge Organization* 44(1): 55-64. Also available in *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. http://www.isko.org/cyclo/subject

Hjørland, Birger. 2018. "Indexing: Concepts and Theory." *Knowledge Organization* 45(7): 609-39. Also available in *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. http://www.isko.org/cyclo/indexing

Hjørland, Birger and Lykke Kyllesbech Nielsen. 2001. "Subject Access Points in Electronic Retrieval." *Annual Review of Information Science and Technology* 35: 249-98.

Holstrom, Chris. 2019. "Moving Towards an Actor-Based Model for Subject Indexing." *NASKO: North American Symposium on Knowledge Organization* 7(1): 120-28. http://dx.doi.org/10.7152/nasko.v7i1.15631

Hulme, E. Wyndham. 1911. "Principles of Book Classification". *Library Association Record*, 13:354-8, Oct. 1911; 389-94, Nov. 1911 & 444-9, Dec. 1911.

ISO 5127:2017. *Information and Documentation: Foundation and Vocabulary*. 2nd ed. Geneva, Switzerland: The International Organization for Standardization.

ISO 5963:1985. *Documentation - Methods for Examining Documents, Determining their Subjects, and Selecting Indexing Terms.* Geneva: International Organization for Standardization.

Jacob, Elin K. and Debora Shaw. 1996. "Is a Picture Worth a Thousand Words?" *Advances in Knowledge Organization* 5: 174-81.

Jacquemin, Christian and Didier Bourigault. 2003. "Term Extraction and Automatic Indexing." In *The Oxford Handbook of Computational Linguistics,* edited by Ruslan Mitkov. Oxford: Oxford University Press, 599–615. DOI: 10.1093/oxfordhb/9780199276349.013.0033

Jäschke, Robert, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme and Gerd Stumme. 2008. "Tag Recommendations in Social Bookmarking Systems." *AI Communications* 21(4): 231-47.

Juilland, Alphonse and Alexandra Roceric. 1972. *The Linguistic Concept of Word: Analytic Bibliography*. Janua Linguarum, Series Minor, 130. The Hague: Mouton.

Kain, Philip J. 2005. *Hegel and the Other: A Study of the Phenomenology of Spirit*. Albany: State University of New York Press.

Käki, Mika. 2006. "fKWIC: Frequency-Based Keyword-In-Context Index for Filtering Web Search Results." *Journal of the American Society for Information Science and Technology* 57(12): 1606-15.

Kathait, Shailendra Singh, Shubhrita Tiwari, Anubha Varshney and Ajit Sharma. 2017. "Unsupervised Keyphrase Extraction using Noun Phrases." *International Journal of Computer Applications* 162(1): 1-5. doi: 10.5120/ijca2017913171

Krámský, Jiří. 1969. *The Word as a Linguistic Unit*. Janua Linguarum, Series Minor, 75. The Hague: Mouton.

Lancaster, Frederick W. 1968. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Wiley.

Lancaster, Frederick W. 2003. *Indexing and Abstracting in Theory and Practice*. 3rd ed. London: Facet.

Lancia, Franco. 2008 "Word Co-Occurrence and Similarity in Meaning: Some Methodological Issues." In *Mind as Infinite Dimensionality*, edited by Sergio Salvatore and Jaan Valsiner. Roma: Carlo Amore, 1–39. https://mytlab.com/wcsmeaning.pdf

Larson, Ray R. 1991. "The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog." *Journal of the American Society for Information Science* 42(3): 197-215.

Lauriston, Andy. 1994. "Automatic Recognition of Complex Terms: Problems and the TERMINO Solution." *Terminology* 1(1): 149-70.

Lee, Durin, Won Chul Kim, Andreas Charidimou and Min Song. 2015. "A Bird's-Eye View of Alzheimer's Disease Research: Reflecting Different Perspectives of Indexers, Authors, or Citers in Mapping the Field." *Journal of Alzheimer's Disease* 45(4): 1207-22. doi: 10.3233/JAD-142688

Li, Decong, Sujian Li, Wenjie Li, Wei Wang and Weiguang Qu. 2010. "A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network." In *Proceedings of the ACL 2010 Conference Short Papers, Uppsala, Sweden, 11-16 July 2010.* Uppsala, Sweden: Association for Computational Linguistics, 296–300. https://www.aclweb.org/anthology/P10-2055.pdf

Library of Congress, Policy and Standards Division. 2006. *Library of Congress Subject Headings. Module 1.2: Why Do We Use Controlled Vocabulary?* PowerPoint Presentation. Retrieved from https://www.loc.gov/catworkshop/lcsh/PDF%20scripts/1-2-WhyCV.pdf

Lok, Corie. 2010. "Speed Reading." *Nature* 463(7280): 416-8.

Lu, Wei, Xin Li, Zhifeng Liu and Qikai Cheng. 2019. "How do Author-Selected Keywords Function Semantically in Scientific Manuscripts?" *Knowledge Organization* 46(6): 403-18. https://doi.org/10.5771/0943-7444-2019- 6-403

Luhn, Hans Peter. 1957. "A Statistical Approach to the Mechanized Encoding and Searching of Literary Infor-

mation." *IBM Journal of Research and Development* 1(4): 309-17.

Luhn, Hans Peter. 1960. "Keyword-In-Context Index for Technical Literature." *American Documentation* 11(4): 288–95.

Mácha, Jakub. 2015. *Wittgenstein on Internal and External Relations: Tracing All the Connections*. London: Bloomsbury Academic.

McCutcheon, Sevim. 2009. "Keyword vs Controlled Vocabulary Searching: The One with the Most Tools Wins." *The Indexer: The International Journal of Indexing* 27(2): 62-5.

McJunkin, Monica Cahill. 1994. *Precision and Recall in Title Keyword Searches*. Master's Research Paper. Kent, OH: Kent State University. https://pdfs.semanticscholar.org/14eb/fdef43d49942c71e01a637a8b730ed1068be.pdf?_ga=2.217421270.1972569115.1590375949-1712967020.1514566464

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Medelyan. Olena and Ian H. Witten. 2006. "Thesaurus Based Automatic Keyphrase Indexing." In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill NC USA June, 2006.* New York, NY: Association for Computing Machinery, 296-7.

Milstead, Jessica L. 1984. *Subject Access Systems: Alternatives in Design*. Orlando, FL: Academic Press.

Mooers, Calvin N. 1950. "Zatocoding for Punched Cards." *Zator Technical Bulletin* 30. Boston, MA: Zator Co.

Mooers, Calvin N. 1972. "Descriptors." In *Encyclopedia of Library and Information Science*, edited by Allen Kent and Harold Lancour. New York: Marcel Dekker, Vol. 7, 31-45. (Reprinted in Mooers 2003).

Mooers, Calvin N. 2003. "Descriptors." In *Encyclopedia of Library and Information Science* 2nd ed., edited by Miriam A. Drake New York: Marcel Dekker, Vol. 2: 813-21. doi: 10.1081/E-ELIS 120008981 (Reprint of Mooers 1972).

Mooers, Calvin N. and Charlotte D. Mooers. 1993. "Oral History Interview with Calvin N. Mooers and Charlotte D. Mooers". Interviewed by Kevin D. Corbitt. Minneapolis, MN: Charles Babbage Institute, 7-11. https://conservancy.umn.edu/handle/11299/107510

Morris, Jane. 2010. "Individual Differences in the Interpretation of Text: Implications for Information Science." *Journal of the American Society for Information Science and Technology* 61(1): 141-9. https://doi.org/10.1002/asi.21222

National Library of Medicine. 2018. "Frequently Asked Questions About Indexing for MEDLINE: Who are the Indexers, and What are Their Qualifications?" https://www.nlm.nih.gov/bsd/indexfaq.html#qualifications

Névéol, Aurélie, Rezarta Islamaj Doğan and Zhiyong Lu. 2010. "Author Keywords in Biomedical Journal Articles." 537-41. *AMIA [American Medical Informatics Association] Annual Symposium proceedings*. AMIA Symposium Vol. 2010 537-41.

Olson, Hope A. and John J. Boll. 2001. *Subject Analysis in Online Catalogs*. Englewood, CO: Libraries Unlimited.

Oxford English Dictionary. 2020a. "Keyword." Retrieved 2020-06-19 from: https://www.oed.com/.

Oxford English Dictionary. 2020b. "Terminology." Retrieved 2020-06-19 from: ttps://www.oed.com/.

Peset, Fernanda, Fernanda Garzon-Farinos, L.M. Gonzalez, X. Garcia-Masso, Antonia Ferrer-Sapena, J. L. Toca-Herrera and E. A. Sanchez-Perez. 2020. "Survival Analysis of Author Keywords: An Application to the Library and Information Sciences Area." *Journal of the Association for Information Science and Technology* 71(4): 462-73. doi: 10.1002/asi.24248

Rafferty, Pauline. 2018. "Tagging." *Knowledge Organization* 45(6): 500-16. Also available in *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. http://www.isko.org/cyclo/tagging

Read More. http://books.infotoday.com/asist/TheInfoSciLib3.shtml#ixzz6Q5qYq5Bc

Redmond-Neal, Alice and Marjorie M. K. Hlava (Eds.). 2005. *ASIS&T Thesaurus of Information Science, Technology, and Librarianship 3rd*. ed. Medford, NJ: Information Today.

Reitz, Joan M. 2004. *Online Dictionary for Library and Information Science* (ODLIS). Santa Barbara, CA: ABC-CLIO. https://products.abc-clio.com/ODLIS/odlis_p.aspx

Roberts, Norman. 1984. "The Pre-History of the Information Retrieval Thesaurus". *Journal of Documentation* 40(4): 271-85.

Robertson, Stephen. 2008. *The State of Information Retrieval: A Researcher's View*. ISKO-UK. Presentation and audio recording freely available at: http://event-archive.iskouk.org/content/state-information-retrieval-researchers-view

Salton, Gerard and Chung-Shu Yang. I 973. "On the Specification of Term Values in Automatic Indexing." *Journal of Documentation* 29(4): 351-72. https://doi.org/10.1108/eb026562

Salton, Gerard, Chung-Shu Yang and C. T. Yu. 1975. "A Theory of Term Importance in Automatic Text Analysis." *Journal of the American Society for Information Science* 26(1): 33-44. https://doi.org/10.1002/asi.4630260106

Sanderson, Mark and W. Bruce Croft. 2012. "The History of Information Retrieval Research." *Proceedings of the IEEE* 100 (Special Centennial Issue):1444-51. doi:10.1109/JPROC.2012.2189916.

Schwing, Theda, Sevim McCutcheon, and Margaret Beecher Maurer. 2012. "Uniqueness Matters: Author-Supplied Keywords and LCSH in the Library Catalog." *Cataloging & Classification Quarterly* 50(8): 903-28.

Siddiqi, Sifatullah and Aditi Sharan. 2015. "Keyword and Keyphrase Extraction Techniques: A Literature Review." *International Journal of Computer Applications* 109(2): 18-23.

Skare, Roswitha. 2020. "Paratext." In *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. https://www.isko.org/cyclo/paratext

Small, Henry G. 1978. "Cited Documents as Concept Symbols." *Social Studies of Science* 8(3): 327-40.

Smiraglia, Richard P. 2013. "Keywords, Indexing, Text Analysis: An Editorial." *Knowledge Organization* 40(3): 155-9.

Smith, Arthur. 2020. "Physics Subject Headings (PhySH)." *Knowledge Organization* 47(3): 257-66. Also available in *ISKO Encyclopedia of Knowledge Organization,* edited by Birger Hjørland and Claudio Gnoli. https://www.isko.org/cyclo/physh

Soergel, Dagobert. 1974. *Indexing Languages and Thesauri: Construction and Maintenance*. Los Angeles, CA: Melville.

Spärck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 28(1): 11-21. https://doi.org/10.1108/eb026526

Spärck Jones, Karen. 1973. "Index Term Weighting." *Information Storage and Retrieval* 9(11): 619-33. https://doi.org/10.1016/0020-0271(73)90043-0

Strader, C Rockelle. 2009. "Author-assigned Keywords versus Library of Congress Subject Headings." *Library Resources & Technical Services* 53(4): 243-50.

Subramaniyaswamy, V. and S. Chenthur Pandian. 2012. "Effective Tag Recommendation System Based on Topic Ontology Using Wikipedia and WordNet." *International Journal of Intelligent Systems* 27(12): 1034-48.

Surgenor, Stephen D., Howard L. Corwin and Terri Clerico. 2001. "Survival of Patients Transferred to Tertiary Intensive Care from Rural Community Hospitals." *Critical Care* 5(2): 100-4. doi: 10.1186/cc993.

Swanson, Don R. 1986. "Undiscovered Public Knowledge." *The Library Quarterly* 56(2): 103-18.

Syn, Sue Yeon and Michael B. Spring. 2009. "Tags as Keywords–comparison of the Relative Quality of Tags and Keywords." *Proceedings of the American Society for Information Science and Technology* 46(1): 1-19. doi: 10.1002/meet.2009.1450460247

Taube, Mortimer, C. D. Gull and Irma S. Wachtel 1952. "Unit Terms in Coordinate Indexing." *American Documentation* 3(4): 213-8.

Thomaidou, Stamatina and Michalis Vazirgiannis. 2011. "Multiword Keyword Recommendation System for Online Advertising." In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, *Kaohsiung, Taiwan, July 25-27, 2011*. Los Alamitos, CA: IEEE Computer Society, 423-7. DOI: 10.1109/ASONAM.2011.70.

Togeby, Knud. 1949. "Qu'est-ce qu'un mot?" *Travaux du Cercle Linguistique de Copenhague* (TCLC) 5: 97-111.

Uddin, Shahadat and Arif Khan. 2016. "The Impact of Author-Selected Keywords on Citation Counts." *Journal of Informetrics* 10(4): 1166-77. https://doi.org/10.1016/j.joi.2016.10.004

Uhlenbeck, Eugenius Marius. 2003. "Words." In William J. Frawley (Ed.) *International Encyclopedia of Linguistics*. 2nd ed. Oxford: Oxford University Press, Vol. 4: 377-8.

Warner, Julian. 2019. "When Might Human Indexing Be Strongly Justified." *Proceedings from the Document Academy* 5(1): Article 5. doi: https://doi.org/10.35492/docam/6/1/5

Wellisch, Hans H. 1995. "Keyword." In Hans H. Wellisch. *Indexing from A to Z*. 2nd ed. New York: H.W. Wilson, 248-53.

Wikipedia, the Free Encyclopedia. 2020. "Index term." [Redirected from "keyword"] Retrieved 2020-06-19 from: https://en.wikipedia.org/wiki/Index_term

Wiktionary, the Free Dictionary. 2020. "Term." Retrieved 2020-06-19 from: https://en.wiktionary.org/wiki/term#English

Williams, Raymond. 2015. *Keywords: A Vocabulary of Culture and Society*. New edition. New York: Oxford University Press.

Witten, Ian H., Gordon W. Paynter, Eibe Frank, Carl Gutwin and C. Craig G. Nevill-Manning. 1999. "Kea: Practical Automatic Key-Phrase Extraction." In *Proceedings of the fourth ACM conference on Digital libraries.* Berkeley, CA: ACM, 254-2. https://arxiv.org/abs/cs/9902007

Zhang, Juan, Qi Yu, Fashan Zheng, Chao Long, Zuxun Lu and Zhiguang Duan. 2016. "Comparing Keywords Plus of WOS and Author Keywords: A Case Study of Patient Adherence Research." *Journal of the Association for Information Science and Technology* 67(4): 967-72. https://doi.org/10.1002/asi.23437