

Contents

Articles

Wei Lu, Xin Li, Zhifeng Liu and Qikai Cheng.
How do Author-Selected Keywords Function
Semantically in Scientific Manuscripts? 403

Feng-Tyan Lin.
Drawing a Knowledge Map of Smart City
Knowledge in Academia..... 419

Reviews of Concepts in Knowledge Organization

Stella G. Dextre Clarke.
The Information Retrieval Thesaurus 439

Steven J. Dick.
Astronomy's Three Kingdom System:
A Comprehensive Classification System of
Celestial Objects460

Aleksandra A. Nikiforova.
Soil Classification467

Books Recently Published489

KNOWLEDGE ORGANIZATION

KO

Official Journal of the International Society for Knowledge Organization

ISSN 0943 – 7444

International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation

KNOWLEDGE ORGANIZATION

This journal is the organ of the INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION (General Secretariat: Amos DAVID, Université de Lorraine, 3 place Godefroy de Bouillon, BP 3397, 54015 Nancy Cedex, France. E-mail: secr@isko.org).

Editors

Richard P. SMIRAGLIA (Editor-in-Chief), Institute for Knowledge Organization and Structure, Shorewood WI 53211 USA.
E-mail: KOeditor-in-chief@knoworg.org

Joshua HENRY, Institute for Knowledge Organization and Structure, Shorewood WI 53211 USA.

Peter TURNER, Institute for Knowledge Organization and Culture, Shorewood WI 53211 USA.

J. Bradford YOUNG (Bibliographic Consultant), Institute for Knowledge Organization and Structure, Shorewood WI 53211, USA.

Editor Emerita

Hope A. OLSON, School of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, Northwest Quad Building B, 2025 E Newport St., Milwaukee, WI 53211 USA. E-mail: holson@uwm.edu

Series Editors

Birger HJØRLAND (Reviews of Concepts in Knowledge Organization), Department of Information Studies, University of Copenhagen. E-Mail: birger.hjorland@hum.ku.dk

María J. LÓPEZ-HUERTAS (Research Trajectories in Knowledge Organization), Universidad de Granada, Facultad de Biblioteconomía y Documentación, Campus Universitario de Cartuja, Biblioteca del Colegio Máximo de Cartuja, 18071 Granada, Spain. E-mail: mjlopez@ugr.es

Editorial Board

Thomas DOUSA, The University of Chicago Libraries, 1100 E 57th St, Chicago, IL 60637 USA. E-mail: tmdousa@uchicago.edu

Melodie J. FOX, Institute for Knowledge Organization and Structure, Shorewood WI 53211 USA. E-mail: melodie.j.fox@gmail.com

Jonathan FURNER, Graduate School of Education & Information Studies, University of California, Los Angeles, 300 Young Dr. N, Mailbox 951520, Los Angeles, CA 90095-1520, USA.
E-mail: furner@gseis.ucla.edu

Claudio GNOLI, University of Pavia, Science and Technology Library, via Ferrata 1, I-27100 Pavia, Italy. E-mail: claudio.gnoli@unipv.it

Ann M. GRAF, School of Library and Information Science, Simmons University, 300 The Fenway, Boston, MA 02115 USA.
E-mail: ann.graf@simmons.edu

Jane GREENBERG, College of Computing & Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 USA, E-mail: jg3423@drexel.edu

José Augusto Chaves GUIMARÃES, Departamento de Ciência da Informação, Universidade Estadual Paulista–UNESP, Av. Hygino Muzzi Filho 737, 17525-900 Marília SP Brazil. E-mail: guima@marilia.unesp.br

Michael KLEINEBERG, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin. E-mail: michael.kleineberg@ub.hu-berlin.de

Kathryn LA BARRE, School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, MC-493, Champaign, IL 61820-6211 USA. E-mail: klabarre@illinois.edu

Devika P. MADALLI, Documentation Research and Training Centre (DRTC) Indian Statistical Institute (ISI), Bangalore 560 059, India. E-mail: devika@drtc.isibang.ac.in

Daniel MARTÍNEZ-ÁVILA, Departamento de Ciência da Informação, Universidade Estadual Paulista–UNESP, Av. Hygino Muzzi Filho 737, 17525-900 Marília SP Brazil. E-mail: dmartinezavila@marilia.unesp.br

Widad MUSTAFA el HADI, Université Charles de Gaulle Lille 3, URF IDIST, Domaine du Pont de Bois, Villeneuve d'Ascq 59653, France. E-mail: widad.mustafa@free.fr

H. Peter OHLY, Prinzenstr. 179, D-53175 Bonn, Germany.
E-mail: Peter.Ohly@gmx.de

M. Cristina PATTUELLI, School of Information, Pratt Institute, 144 W. 14th Street, New York, New York 10011, USA.
E-mail: mpattuel@pratt.edu

K. S. RAGHAVAN, Member-Secretary, Sarada Ranganathan Endowment for Library Science, PES Institute of Technology, 100 Feet Ring Road, BSK 3rd Stage, Bangalore 560085, India. E-mail: ksragav@hotmail.com

Heather Moulaison SANDY, The iSchool at the University of Missouri, 303 Townsend Hall, Columbia, MO 65211, USA.
E-mail: moulaisonhe@missouri.edu

M. P. SATIJA, Guru Nanak Dev University, School of Library and Information Science, Amritsar-143 005, India.
E-mail: satija_mp@yahoo.com

Aida SLAVIC, UDC Consortium, PO Box 90407, 2509 LK The Hague, The Netherlands. E-mail: aida.slavic@udcc.org

Renato R. SOUZA, Applied Mathematics School, Getulio Vargas Foundation, Praia de Botafogo, 190, 3o andar, Rio de Janeiro, RJ, 22250-900, Brazil. E-mail: renato.souza@fgv.br

Rick SZOSTAK, University of Alberta, Department of Economics, 4 Edmonton, Alberta, Canada, T6G 2H4. E-mail: rszostak@ualberta.ca

Joseph T. TENNIS, The Information School of the University of Washington, Box 352840, Mary Gates Hall Ste 370, Seattle WA 98195-2840 USA. E-mail: jtennis@u.washington.edu

Maja ŽUMER, Faculty of Arts, University of Ljubljana, Askerceva 2, Ljubljana 1000 Slovenia. E-mail: maja.zumer@ff.uni-lj.si

How do Author-Selected Keywords Function Semantically in Scientific Manuscripts?†

Wei Lu*, Xin Li**, Zhifeng Liu***, Qikai Cheng****

Wuhan University, School of Information Management,
Information Retrieval and Knowledge Mining Laboratory, Wuhan 430072, China,

*<weilu@whu.edu.cn>, **<lucian@whu.edu.cn>, ***<zfliu17@163.com>, ****<chengqikai@whu.edu.cn>

Wei Lu is Professor of information science, Director of the Information Retrieval and Knowledge Mining Center, and Vice Dean of the School of Information Management, Wuhan University. He was also a visiting scholar at the City University of London, UK (2005-2006) and The Royal School of Library and Information Science, Denmark (2011-2012). His research interests are information retrieval, knowledge mining and visualization, knowledge organization and knowledge management. He is also an editorial advisory board member of the *Journal of Data and Information Science* and the *Journal of the China Society for Scientific and Technical Information*.



Xin Li is a doctoral student at the School of Information Management, Wuhan University. He is in the last year of a masters-doctorate combined program now. He received his bachelor's degree in health information management from Tongji Medical College, Huazhong University of Science and Technology (HUST) and in English literature from the School of Foreign Language, HUST. His research interests include health knowledge organization, bioinformatics, data mining and science of science.



Zhifeng Liu is a graduate student at the School of Information Management, Wuhan University. He received his bachelor's degree in information management and systems from Huazhong University of Science and Technology. His research interests include informetrics, text mining and medical informatics.



Qikai Cheng is a lecturer at the School of Information Management, Wuhan University. He received his doctoral degree in information science at Wuhan University in 2015. He is also a visiting scholar at the University of Pittsburgh and a researcher at the Information Retrieval and Knowledge Mining Center of Wuhan University. His research interests are natural language processing, information retrieval and text mining.



Lu, Wei, Xin Li, Zhifeng Liu and Qikai Cheng. 2019. "How do Author-Selected Keywords Function Semantically in Scientific Manuscripts?" *Knowledge Organization* 46(6): 403-418. 57 references. DOI:10.5771/0943-7444-2019-6-403.

Abstract: Author-selected keywords have been widely utilized for indexing, information retrieval, bibliometrics and knowledge organization in previous studies. However, few studies exist concerning how author-selected keywords function semantically in scientific manuscripts. In this paper, we investigated this problem from the perspective of term function (TF) by devising indicators of the diversity and symmetry of keyword term functions in papers, as well as the intensity of individual term functions in papers. The data obtained from the whole *Journal of Informetrics (JOI)* were manually processed by an annotation scheme of keyword term functions, including "research topic," "research method," "research object," "research area," "data" and "others," based on empirical work in content analysis. The results show, quantitatively, that the diversity of keyword term function decreases, and the irregularity increases with the number of author-selected keywords in a paper. Moreover, the distribution of the intensity of individual keyword term function indicated that no significant difference exists between the ranking of the five term functions with the increase of the number of author-selected keywords (i.e., "research topic" > "research method" > "research object" > "research area" > "data"). The findings indicate that precise keyword related research must take into account the distinct types of author-selected keywords.

Received: 11 April 2019; Revised: 21 June 2019; Accepted: 27 June 2019

Keywords: term functions, author-selected keywords, research topic

† This study was supported by the Major Project of the National Social Science Foundation of China (17&ZDA292) and the National Natural Science Foundation of China (71473183). The support provided by China Scholarship Council (CSC) during a visit of Xin Li to Indiana University Bloomington is also acknowledged. The authors would like to express special gratitude to Yi Bu and Yong Huang for their valuable comments and editorial assistance. The authors are also grateful to the anonymous referees and editors for their invaluable and insightful comments.

1.0 Introduction

Author-selected keywords are considered as a significant conduit of scientific concepts, ideas and knowledge (Cobo, López-Herrera, et al. 2011; Ding, Chowdhury, and Foo 2001; Névóol, Doğan, and Lu 2010; Van Raan and Tijssen 1993) and have been widely utilized in indexing, knowledge management, bibliometrics and information retrieval. For instance, a keyword co-occurrence network was constructed to map the knowledge structure of technology foresight research by (Su and Lee 2010). Khan and Wood (2015) conducted a co-keywords clustering to detect emerging themes in the information technology management domain. More recently, Wu (2016) adopted a keyword-based patent network approach to identify technological trends and evolution in the field of green energy. All of these studies can be summarized as “keyword analysis,” whose general workflow entails data retrieval and collection, keywords identification and preprocessing, frequency counting, network generation, analysis and visualization, and interpretation and conclusion.

However, the indiscriminating use of keyword analysis remains controversial given the existence of certain problems such as the lack of an authoritative criterion for the selection of keywords (e.g., Chen and Xiao 2016; Milojević et al. 2011; Smiraglia 2013), the presence of possible bias due to the “indexer effect” (Michel Callon, Rip, and Law 1986; He 1999), ignoring semantic roles and their relationships between keywords (Wang et al. 2012) and the discipline attributes of keywords (Chen and Xiao 2016; J. Choi, Yi, and Lee 2011).

Actually, each author-selected keyword plays a specific semantic role or function, which can be called a “term function” (TF) in a scientific paper. Specifically, a keyword could be the topic discussed or the method adopted or it also could play another semantic role in a scientific paper. In most extant studies of keyword analysis, keywords that play different semantic roles that should have been weighted differently are treated as equally important by simple counting and aggregation for different tasks (Ferrara and Salini 2012). However, “topic,” “domain,” “method” and “application” keywords should have been assigned unequal weights for generating accurate research topic networks. Hence, to overcome these problems, the semantic function of author-selected keywords played in scientific manuscripts should be elucidated. In addition, understanding how the author-keywords function semantically in scientific manuscripts is also beneficial to the organization and indexing of scientific papers in databases and to determine papers’ accessibility and citations in scientific communities.

The overall aim of this paper is to reveal the patterns of author-selected keywords in scientific papers from the perspective of term function, whose results will substantially contribute to the improvement of keyword indexing and

keyword analysis. To realize this goal, the following research questions are posed:

- 1) What is the distribution of author-selected keyword term functions in scientific papers?
- 2) What is the regularity of the diversity and symmetry of author-selected keyword term functions in scientific papers?
- 3) What is the distribution of the intensity of individual keyword term functions in scientific papers?
- 4) What is the relationship between the author-selected keyword ranking and its term functions in scientific papers?

In this study, we first annotated term functions for all author-selected keywords in our dataset, for which an annotation scheme based on empirical work in content analysis is presented. Then, we introduced a framework to compute the diversity and symmetry of keyword term functions in a single paper, as well as the distribution of the intensity of individual keyword term functions, using concepts from network science and “true diversity,” which can be understood as a normalization for the Shannon entropy. We also analyze the relationships between keyword rankings and keyword term functions.

The remainder of this paper is organized as follows. Section 2.0 reviews studies regarding author-selected keywords and term function (TF). Section 3.0 presents the dataset and the annotation scheme for keyword term function, as well as the framework to represent and evaluate the diversity, intensity and symmetry of author keyword term functions in papers. In Section 4.0, the main results of this study are described in detail. Finally, in Section 5.0, conclusions and directions for future work are presented.

2.0 Literature review

2.1 Author-selected keywords

Author keywords have been generally regarded as one of the most important forms of bibliographic metadata in bibliometrics and scientometrics, as well as being a significant conduit of scientific concepts, ideas and knowledge (Cobo, López-Herrera, et al. 2011; Ding, Chowdhury and Foo 2001; Névóol, Doğan and Lu 2010; Van Raan and Tijssen 1993). Therefore, author-selected keyword analysis has a long tradition of widespread application in hotspot detection, trend analysis and mapping the knowledge structures in both natural and social sciences, e.g., in environmental acidification (Law et al. 1988), polymer chemistry (M. Callon, Courtial and Laville 1991), chemical engineering (Peters and van Raan 1993), software engineering (Coulter, Monarch and Konda 1998), knowledge discovery (He 1999), in-

formation retrieval (Ding, Chowdhury and Foo 2001), ethics and dementia (Baldwin et al. 2003), geographic information system (GIS) (Tian, Wen and Hong 2008), biomedical science (Névéal, Doğan and Lu 2010), technology foresight (Su and Lee 2010), fuzzy sets theory (Cobo, López-Herrera et al. 2011b), tourism (B. Wu et al. 2012), strategic management (Keupp, Palmié and Gassmann 2012), information technology management (Khan and Wood 2015) and biofuels (Wu 2016). However, with the wide-ranging applications of author-selected keyword analysis, problems with the method have become increasingly evident and have begun to be actively discussed by researchers. For example, Callon, Rip and Law (1986) and He (1999) pointed out the “indexer effect” of author-selected keywords at a theoretical and technical level. More recently, Wang et al. (2012) suggested that experts’ knowledge be integrated into the process of co-word analysis to improve precision; Chen and Xiao (2016) put forward methods for keyword selection that take keyword discrimination into account by considering their frequency both in and out of the domain. In this paper, we will analyze author-selected keywords of different term functions, which should have been weighted unequally in different bibliometric tasks.

Additionally, author-selected keywords have also been widely utilized for the classification and clustering of scientific documents (Jones and Mahoui 2000), the “gold-standard” for automatic keyword indexing and extraction (Matsuo and Ishizuka 2004; Ren 2014; Gil-Leiva 2017), automatic thesaurus development (Gil-Leiva and Alonso-Arroyo 2007; Tseng 2002; J. Wang 2006), the retrieval and recommendation of scientific papers in digital libraries (Lu and Kipp 2014; Schaffner 2009), citation counts prediction (Sohrabi and Iraj 2017; Uddin and Khan 2016) and the comparison with social tags (Y. Choi and Syn 2016; Lu and Kipp 2014).

2.2 Term function in scientific texts

Term function (TF) refers to the specific semantic role that a word, a term or a phrase plays in scientific texts (Xin, Qikai and Wei 2017), including “topic,” “method,” “technology,” etc. For instance, in the paper entitled “Knowledge discovery through co-word analysis” (He 1999), the TF of the term “knowledge discovery” is a “topic”; whereas, for the term “co-word analysis,” it is a “method.” Notably, the TF of the same term can differ in different contexts, for example, the TF of the term “knowledge discovery” is a “method” in the article entitled “Intelligent query answering by knowledge discovery techniques” (Han et al. 1996). In addition, academic terms have numerous other functions according to different classifications, such as “goal,” data and “application,” which are also quite common in scientific contexts.

With the dramatic growth in the number of scientific publications, it has become a challenge to understand a scientific community by identifying important topics, methods, applications and the relations between them. In the extant literature, this question has been mainly addressed using bibliometric methods, for example, considering citation networks and topic models (Ding 2011; Song et al. 2014) and generating crude topic clustering based on contextual cues. However, several researchers concluded that these methods could not answer certain key questions, such as “what methods were used for a particular topic?” and pointed out that the need to identify the semantic roles of scientific terms by analyzing the text itself, i.e., the identification of the term function (TF) in scientific texts (Kondo et al. 2009; Tsai, Kundu and Roth 2013).

Identification of term functions has received increasing interest with the rapid development of natural language processing and machine learning. Key terms that play different semantic roles have been identified, such as the identification of “head,” “goal” and “method” in research papers’ titles based on a rule extracted from the structure of titles (Kondo et al. 2009), the recognition of “technology” and “effect” from research papers and patents based on machine learning (Nanba, Kondo and Takezawa 2010), the identification of “focus,” “techniques” and “domain” from article abstracts by using semantic extraction patterns (Gupta and Manning 2011), the recognition of “techniques” and “application” from scientific literature using an unsupervised bootstrapping algorithm (Tsai, Kundu and Roth 2013) and the identification of “method” and “task” from scientific papers based on the Markov Logic Network (Huang and Wan 2013).

More recently, a comprehensive framework for term function in academic texts was presented by Xin, Qikai and Wei (2017). In his study, Cheng categorized term functions into “domain-independent term function” (including “topic” and “method” in three levels) and “domain-related term function” (different sub categories in different domains). Based on this classification, approaches have been used, including conditional random fields with word2vec and machine learning to rank, for automatic recognition of domain-independent term functions in scientific papers in computer science. In addition, Heffernan and Teufel (2018) presented an automatic classifier for identifying problems and solutions in scientific texts. It remains unknown, however, precisely how author-selected keywords function semantically in scientific manuscripts. Understanding qualitatively and quantitatively the patterns of author-selected keywords from term function perspectives, in our view, is of great benefit for improving keyword indexing and keyword analysis in bibliometric tasks.

3.0 Methodology

This section examines the overall process of the methodology, as illustrated in Figure 1. To investigate the author-selected keyword patterns, the approach is designed to be executed in four discrete steps: 1) data collection and pre-processing; 2) term function annotation; 3) indicator computing; and, 4) patterns analysis.

3.1 Step 1: data collection and processing

In this step, we collected the publication records from the *Journal of Informetrics (JOI)*. To probe the author-selected keyword patterns from the term function (TF) perspective in scientific manuscripts, all 842 articles published between

2007 to 2017 from *JOI* were manually collected from the Web of Science. A total of 149 articles were excluded, because they were not articles but, for example, brief communications, book reviews, editorial statements, errata or critical remarks. Finally, 693 articles were selected as the dataset in this study. For each of these articles, we have not only obtained the author-selected keyword lists, but have also extracted the title, abstract and the hyperlink to its detailed information page for term function annotation in the subsequent step. To investigate the relationship between term functions and the ranking of keywords, we also recovered the position of each author-selected keyword in the keyword lists.

The distribution of the number of author-selected keywords per paper is shown in Figure 2. There are a total of

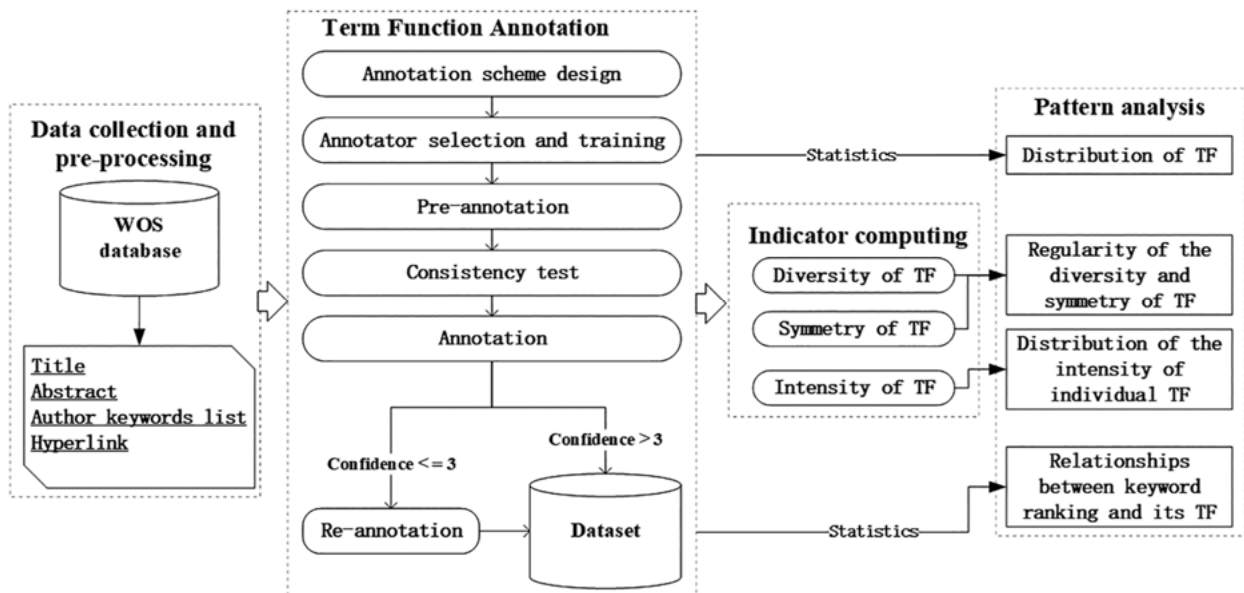


Figure 1. Framework of author keyword pattern analysis from the term function (TF) perspective.

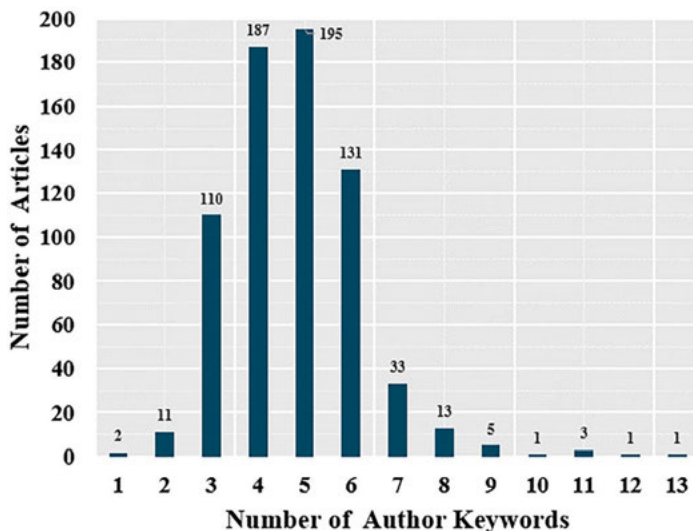


Figure 2. Histogram of the number of keywords in the *Journal of Informetrics (JOI)*. An irregular distribution is found, in which most of the papers include three to six keywords.

3,311 author-selected keywords in all 693 articles, and the average number of author-selected keywords per article is found to be 4.78. It is also found that the range of author-selected keywords for each paper varied from one to thirteen. A few papers contained fewer than two keywords or more than eight keywords (approximately 1.9%), while most papers contained three to six keywords (approximately 89.9%).

3.2 Step 2: term function annotation

3.2.1 Annotation scheme design

In prior studies regarding term function recognition (TFR), words in scientific papers that have been recognized include “topic,” “method,” “problem,” “solution,” “goal,” “technology,” “focus,” “domain,” etc. (Heffernan and Teufel 2018; Xin, Qikai and Wei 2017; Tsai, Kundu and Roth 2013; Kondo et al. 2009; Gupta and Manning 2011; Huang and Wan 2013). Concerning the term function of each author-selected keyword in each article, we present an annotation scheme for author-selected keywords, based on empirical work in content analysis. In the first place, we captured all possible term functions of author-selected keywords. Then, to simplify our analysis, these term functions were integrated and reduced to a smaller set comprising only the most frequent term functions. This set, i.e., the annotation scheme for term functions of author-selected keywords includes the following categories: 1) research topic; 2) research method; 3) research object; 4) research area; 5) data; and, 6) others. The detailed description and source for each category of term function is shown in Table 1.

In order to guarantee the precision of term function annotation, the method of human annotation is selected. The term function of author-selected keywords is difficult to annotate, because, in principle, it requires interpretation of the author’s intentions and the content of the entire paper. Consequently, in most cases, it is impossible to know exact term function without understanding academic context, because the same keyword can have a totally different term function in different conditions.

3.2.2 Annotators selection and training

Before term function annotating, four PhD students were selected from the School of Information management, Wuhan University. Four criteria were used in the selection of annotators. Specifically, the annotators had to: 1) be very familiar with informetrics and bibliometrics; 2) have good English reading and writing skills; 3) have published more than two academic articles in peer-reviewed journals in the field of informetrics; and, 4) be in or beyond their second year in the PhD program. Then, the selected annotators were trained and asked to point to textual evidence for assigning a particular term function.

3.2.3 Pre-annotation and consistency test

To guarantee annotation consistency, prior to starting the annotating, we randomly chose sixty-nine articles (9.96%) comprising of 337 author-selected keywords from the JOI dataset and arranged for four annotators to annotate term functions in two parallel groups. Then, the kappa coefficient (Carletta 1996), which is a statistic measuring pairwise

No.	Categories	Description	Source
1	Research Topic (T)	Problems or topics discussed in research articles.	Hoey 2013; Kondo et al. 2009; Heffernan and Teufel 2018; Xin, Qikai and Wei 2017)
2	Research Method (M)	Methods or solutions used in research articles, including theories, bibliometric indicators, algorithms, math formulas, models, etc. For examples, “Bradford’s law,” “h-index,” “PageRank algorithm,” “Hall’s model.”	Augenstein et al. 2017; Heffernan and Teufel 2018; Xin, Qikai and Wei 2017; Mesbah et al. 2017; Tsai, Kundu and Roth 2013; Sahragard and Meihami 2016
3	Research Object (O)	The object that the research studied, including people, group, organization, materials or objects.	Xin, Qikai and Wei 2017; Tsai, Kundu, and Roth 2013
4	Research Area (A)	The academic area or background of the article, for instance, “bibliometrics,” “physics,” “science of science,” and “library and information science (LIS).”	Hoey 2013; Carletta 1996; Sahragard and Meihami 2016
5	Data (D)	The dataset used in the study or the data created by the study, for examples, “APS dataset,” “X corpus,” or “Web of Science,” etc.	Kondo et al. 2009; Mesbah et al. 2017; Sahragard and Meihami 2016
6	Others (OT)	Cannot be included in the former categories.	Kondo et al. 2009; Xin, Qikai and Wei 2017

Table 1. The detailed description for each category of term function of author-selected keywords.

agreements among a set of coders' category judgements, was used for quantifying the consistency. Finally, the coefficients were 0.843 and 0.817 respectively (average 0.830 > 0.75), which was considered sufficiently high for annotating to proceed separately, particular given the conservative nature of the kappa coefficient.

3.2.4 Annotation

In the process of annotating, annotators were asked to carefully read the title and abstract for a comprehensive understanding of the academic context of each keyword in the original dataset and were encouraged to click the hyperlink for its full text to make a further confirmation. Moreover, annotators were asked to record the **Annotation Confidence (ac)** of each article. The value of $ac \in [1,2,3,4,5]$, in which a higher value of ac represents that the annotator is more confident in his or her work. If an article's value of ac is below four, the article will be annotated again by all annotators together.

3.3 Step 3: indicator computing

To quantify the intensity of individual term functions in a paper, as well as the diversity and symmetry of term functions of author-selected keywords in each article, the information provided in each article of our dataset is treated as a bipartite network (Newman 2010), which is a network with links established only among nodes and belonging to distinct groups. As shown in Figure 3, the bipartite network derived from each paper establishes links between

author-selected keywords and their possible term functions. As can be seen from Figure 3, each author-selected keyword is annotated to one term function, while one term function can have multiple author-selected keywords assigned, which can represent the regularity of term functions of author-selected keywords in a paper.

3.3.1 Term function intensity

The term function intensity measure was used to calculate the strength of an individual term function in a scientific paper's author-selected keyword list. In this paper, we first define f as the matrix storing the relationship between author-selected keywords and their term functions in the bipartite network. The following equation was used:

$$f_{ij} = \begin{cases} 1, & \text{if the term function of author keyword } j \text{ is } i \\ 0, & \text{otherwise} \end{cases}$$

In the example provided in Figure 3, $f_{1j} = 1$ only for $j =$ "1st keyword" and $j =$ "4th keyword." Then the intensity of a given term function is given by the following equation:

$$TF \text{ Intensity } (I_i) = \frac{\sum_j \omega_j f_{ij}}{\sum_i \sum_j \omega_j f_{ij}}$$

where ω_j is the weight associated to the j -th author-selected keyword. Differently from Edilson et al. (2017), we weighted the importance of each author-selected keyword to the research according to its rank in the keyword list, as defined by the following equation:

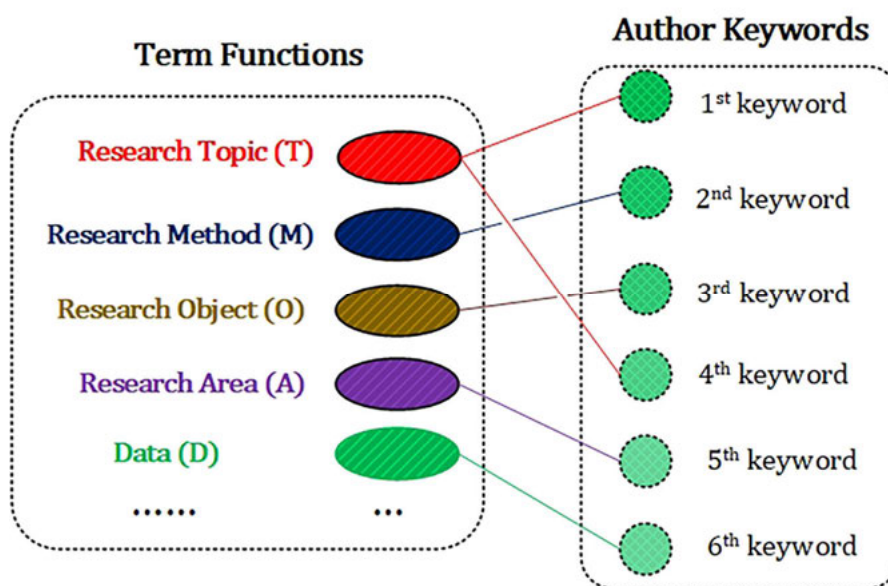


Figure 3. Example of a bipartite network representing the relationship between author-selected keywords and their term functions. Note that the total amount of keywords and particular term functions vary according to article.

$$\omega_j = \begin{cases} -0.1R + 0.6, & R < 5 \\ 0.1, & R \geq 5 \end{cases}$$

3.3.2 Term function diversity

For an article, the “term function diversity” measure calculates the level of variety of the term functions for author keyword lists. Drawing on the accessibility concept, a centrality measurement that can be understood as a normalization for the Shannon entropy was employed in this study. This measurement was originally proposed by Travençolo and Costa (2008) to compute the effective number of access nodes when an agent walks randomly on a network from a starting node. Compared to the traditional measurements, network features are used that go beyond the simple static network topology and can be utilized to quantify the effective number of neighbors (Amancio, Oliveira jr, and da F. Costa 2015). In this paper, a simple interpretation of the diversity measure in terms of network quantities was used to compute term function diversity, which has been extensively done in several studies (Corrêa Jr et al. 2017; Silva et al. 2016; Travençolo and Costa 2008). Notably, the “term function intensity” of each term function ranges in the interval [0,1], and thus we can measure its distribution of it using the entropy concept for all elements in the set of term functions. The following equation was then used to calculate the “term function diversity” of an article:

$$TF \text{ Diversity } (\varphi) = \exp(-\sum_{i \in I} I_i \log I_i)$$

3.3.3 Term function symmetry

The measure of “term function symmetry” examines the distributions of the “term function intensity” of each term function in a scientific paper. Thus, this measure represents how intensity varies across different term functions in a paper using a normalization of “term function diversity.” The normalized TF diversity, referred to as a symmetry of the intensity of individual term function in a paper, takes a range of values restricted in the interval [0,1]. Therefore, the term function symmetry was represented by the following equation:

$$TF \text{ Symmetry } (\sigma) = \frac{\varphi}{n_t}$$

where $n_t \in [1,6]$ is the total number of term functions in the paper. Note that σ is a symmetry measure, because it reaches its maximum value ($\sigma = 1$) when all term functions are assigned equally to the paper.

3.4 Step 4: patterns analysis

In this paper, we reveal the patterns of author-selected keywords from four aspects. First, we described the distribution of author-selected keyword term functions using a statistical method. Second, the results of indicators including “term function diversity” and “term function symmetry” were employed to represent the regularity of author-selected keyword term functions in a scientific manuscript. Third, we also used the indicator “term function intensity” to depict the distribution of the strength of individual term functions in the dataset. Finally, the relationships between author-selected keyword ranking in the article’s keyword list and their term functions were identified to analyze the author’s potential indexing patterns.

Term Function	Percentage
Research Topic (T)	40.75%
Research Method (M)	37.79%
Research Object (O)	7.66%
Research Area (A)	9.55%
Data (D)	1.05%
Others (OT)	3.19%

Table 2. Frequency of appearance for each type of keyword term function, considering all of the papers in the dataset. Each author-selected keyword was counted as a distinct occurrence, even if it appeared in more than one paper in the dataset.

4.0 Results

4.1 The distribution of author-selected keyword term functions

The overall count for the author-selected keyword term functions in the dataset are shown in Table 2. The most common was “research topic,” accounting for 40.75% of the total. “Research method” was a clear second, comprising more than a third of the total (37.79%). The other term functions scored between 7% and 10%, except for “data,” which had very low frequency. In addition, the average number of “research topic[s]” per paper was 2.19, which is the highest among the five term functions. The average number of “research method[s]” per paper is 1.90, ranking second. The other term functions’ average number per paper scored around 0.50, except for “data,” whose average number was very low (0.18).

The distribution of the article numbers of different term functions in the dataset are presented in Figure 4 from which it can also be seen that “research topic” and “research method” are the top two term functions. We also find that the range of the number of “research topic” or “research method” for a paper varies from one to eight. A

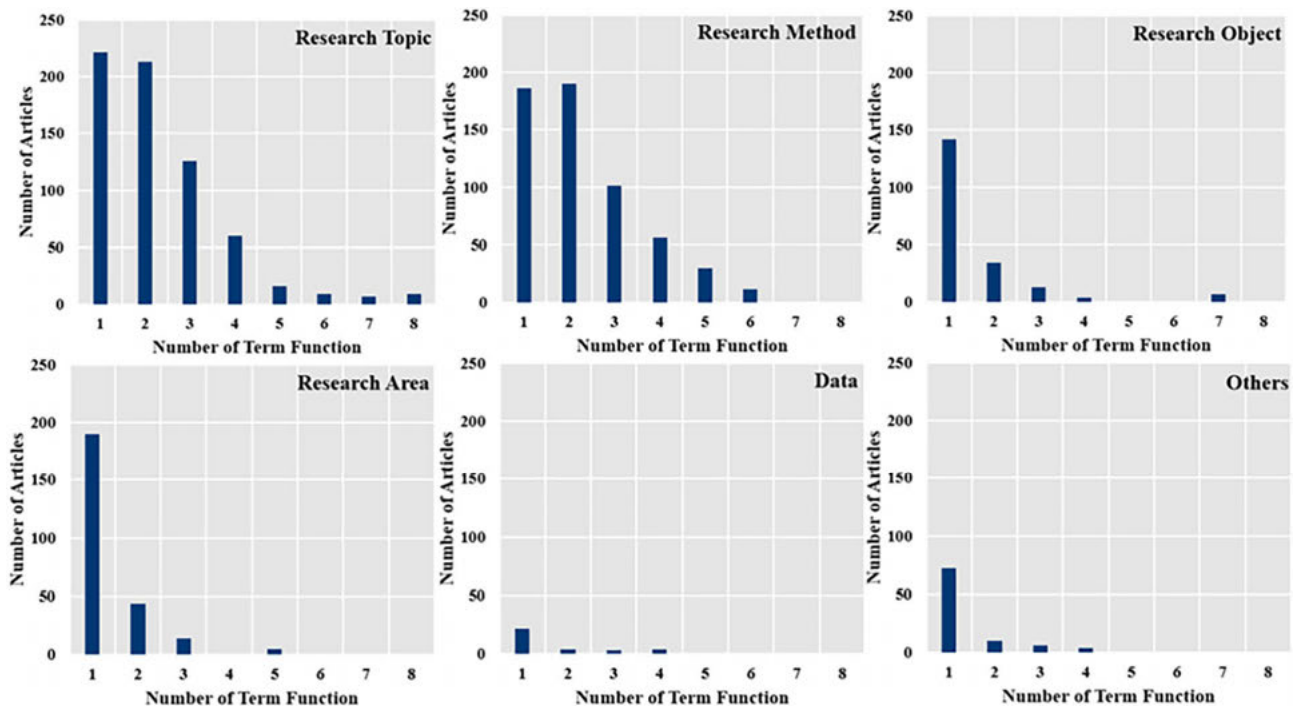


Figure 4. The distribution of the article numbers of different term functions in the dataset.

few papers (less than 15%) contain more than three “research topic” or “research method” keywords, while the most common scenario covers papers that contain one or two individual term functions (more than 50%). Moreover, the range of the number of the other three term functions for a paper is between one and four, while very few papers contain five “research area” or seven “research object” term functions. Most individual term functions have only occurred once in a paper, 70.6% for “research object,” 75.1% for “research area” and 66.7% for “data.”

4.2 The regularity of author-selected keyword term functions in papers

4.2.1 The diversity of author-selected keyword term functions

To investigate how keyword term functions vary in scientific papers, we used the “diversity of term functions in a paper” (φ) as a measure of the variability, as defined in Section 3.3.2. Considering that the value of $\omega_j f_{ij}$ varies according to the number of author-selected keywords (n_K) and the ranking of author-selected keywords (R), we decided to separately compute the values of φ for each n_K . As shown in Figure 5, the red line is the reference curve when the number of keywords assigned to each term function is equal; and if the keyword ranking (i.e. $\omega_j = 1$) is ignored, the reference will fit to the curve $\varphi = n_K$. The other curve denotes the points observed in our *JOI* dataset,

from which one can find that, when n_K increases, the diversity of term functions of author-selected keywords in a paper also increases, thus confirming a relatively strong correlation between these quantities. Moreover, it reaches its highest point (φ is approximately 2.5) when the number of author-selected keywords is six. When $n_K = 1$, $\varphi = 1$, as one should anticipate from the equation above. One can also find that the largest deviations between these quantities (i.e., $n_K - \varphi$) were found for the papers tagged by many author-selected keywords. Note that, in general, the number of “research topic” keywords in papers tagged by more than eight author-selected keywords was usually more than five, which makes the diversity of term functions quite irregular. Considering that this set of articles has eight author-selected keywords, the paper with the most irregular distribution of keyword term functions has a total diversity of term functions of only approximately two. Despite these discrepancies, we can conclude that, in a typical paper tagged by three to six author-selected keywords, the diversity of term functions is relatively high and the difference between n_K and φ is relatively small, as the differences in the number of keyword term functions tagged in these studies is insignificant.

4.2.2 The symmetry of author-selected keyword term functions

The irregularity of author-selected keyword term functions was also investigated in terms of “symmetry of term

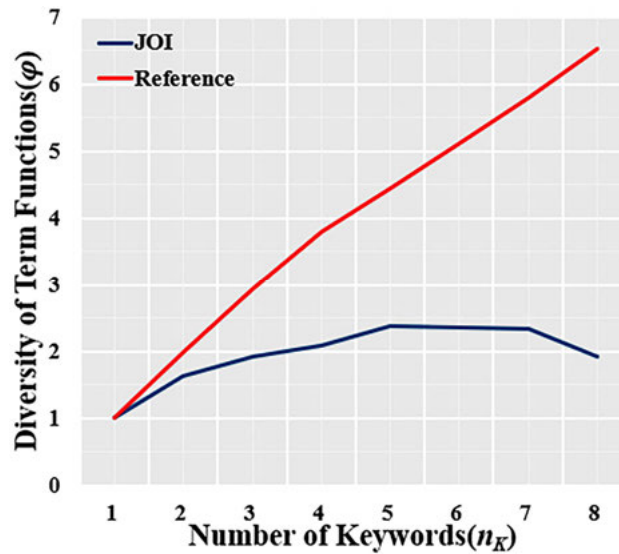


Figure 5. The diversity of author-selected keyword term functions (φ) as a function of the number of author-selected keywords (n_K). Because, in some cases, some term functions are tagged by more than others, their diversity of them is lower than the reference when each term function is tagged equally. The largest deviations occur for the papers tagged by many author-selected keywords.

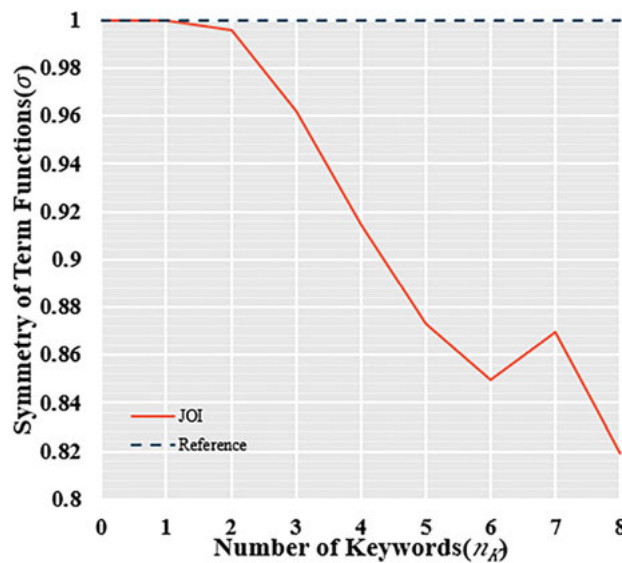


Figure 6. Symmetry of author-selected keyword term functions in papers (σ) as a function of the number of author-selected keywords (n_K).

functions” (σ), as defined in Section 3.3.3. As illustrated in Figure 6, for each value of n_K , we can obtain the corresponding value of TF symmetry. The blue dotted line is the reference line $\sigma_{\text{maximum}} = 1$, and the other line represents the curve obtained by linking the points representing the average symmetry obtained for each n_K , when $n_K = 1, \sigma = 1$. Overall, one can find that the average symmetry of author-selected keyword term functions

monotonically decreases when the number of author-selected keywords increases from $n_K = 1$ to $n_K = 6$. However, when the number of author-selected keywords is more than five, the falling rate of the symmetry decreases significantly. This indicates that the distribution of keyword term functions becomes more irregular when the number of author-selected keywords increases. However, the average value of symmetry is always above 0.80. So, we

further count the number of papers whose symmetry is below 0.8 and find that most of them are in the $n_K = 4$ or $n_K = 5$ group. The reason for this phenomenon might be that due to the large number of papers tagged by four to five keywords, outliers are more common in this subset of papers. In addition, it is evident that values of $\sigma < 0.8$ are not frequent in the dataset with more than six or fewer than four keywords.

4.3 The distribution of the individual term function's intensity

In this section, we will investigate which term functions tend to be tagged more frequently by an author when indexing keywords for a scientific paper. Although no straightforward studies currently exist regarding this issue, the consensus among scientists is that the nature of a research process can be viewed as a problem-solving activity (Heffernan and Teufel 2018; Jordan 1980). When indexing keywords for a paper, authors are asked to use phrases that constitute an adequate description of the paper's content (Ding, Chowdhury and Foo 2001; Gil-Leiva 2017). A pertinent question is then which keywords are indexed more by authors, "research topic" or "research method"? "Data" is also of major significance to scientific research, especially in the field of information science, in which data constitute the essential materials. In addition, "research object" and

"research area" are also essential for a rigorous design of scientific activity. This analysis illustrates the frequent occurrence of all five of these term functions of author-selected keywords. However, what are the differences among the five individual term functions according to the indexing behavior of authors?

To answer the question above, we described the distribution of the "intensity of individual term function" (I_i). We also analyzed the term function as a function of rankings to identify whether there is an implicit factor leading to the organization of rankings according to term functions.

In Figure 7, the distribution of the intensity of individual term functions of the *JOI* dataset is shown. The results are organized by the total number of author-selected keywords considered in Figure 7, with papers tagged by: a) 2; b) 3; c) 4; d) 5; e) 6; and, (f) all author-selected keywords in the *JOI* dataset. In Figure 7, as expected (Heffernan and Teufel 2018; Ding, Chowdhury and Foo 2001; Jordan 1980), it is evident that "research topic" and "research method," in general, obtain higher intensity than the others. Nonetheless, the values of TF intensity are not very different, since, on average, "research topic" and "research method" comprise approximately 40% and 30% of the intensity in paper level, respectively. When more author-selected keywords are included, one can observe a very similar pattern: while "research topic" obtain most of the in-

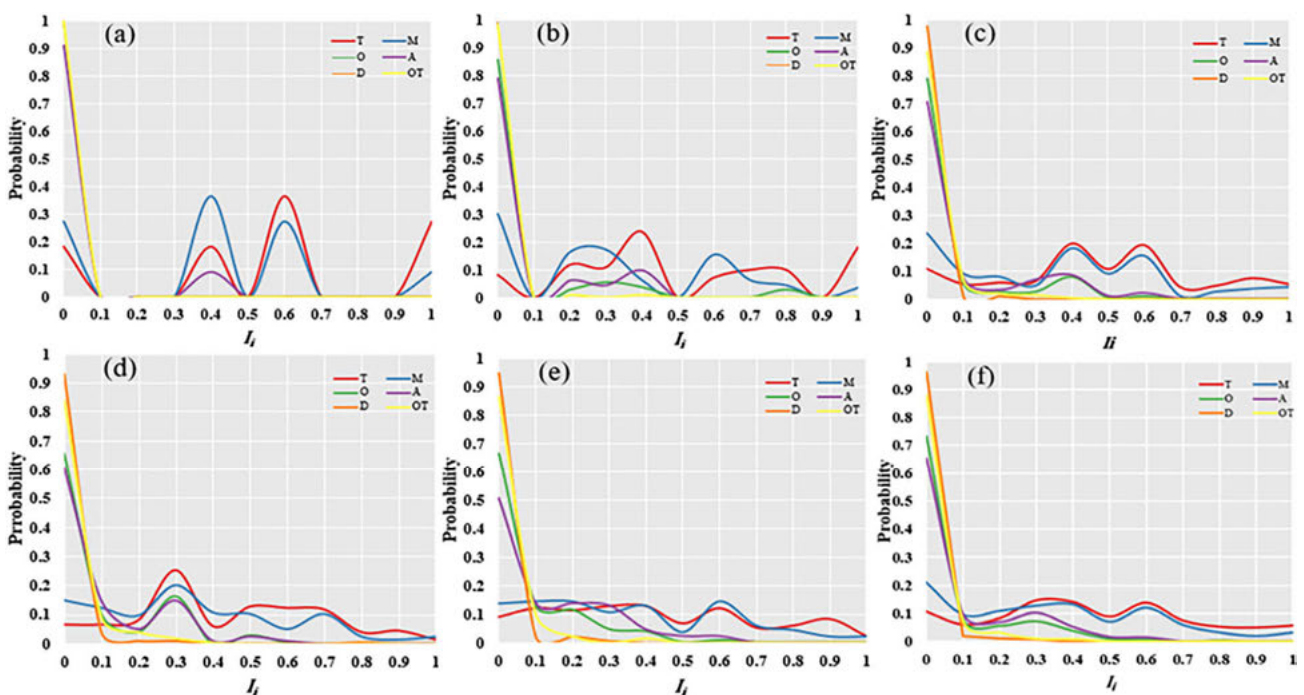


Figure 7. The distribution of individual term functions' intensity in the dataset. The results are shown considering the following number of author-selected keywords: a) 2; b) 3; c) 4; d) 5; e) 6; and, f) all. "Research topic" and "research method" are the first and second term functions, respectively, with a relative larger intensity value.

tensity, “research method” is usually ranked as the second most common term function; and “research object” (about 15%) and “research area” (10%) are third and fourth, respectively. “Data” has the least value of TF intensity in all conditions (less than 5%).

These patterns can also be observed in in Figure 8, which summarizes the average intensity of individual term function (I_i) in terms of the number of keywords. “Research topic” (upper red curve) always obtains most of the intensity, while “research method” usually appears in the second position in the ranking of average intensity. As the number of author-selected keywords increases, however, there is not a larger difference between the ranking of the five term functions on the value of TF intensity (i.e., “research topic” > “research method” > “research object” > “research area” > “data”).

4.4 The relationship between the keyword’s rank and its term function

It is conjectured that, in general, the first keywords are more frequently tagged as “research topic” or “research method,” which are considered as the core part of a paper, while the last keywords have the least significance, such as “others.” However, guidelines for ranking author-selected keywords are not always strictly followed, and thus there is no widespread evidence that exists relating ranking of author-selected keywords and specific term functions. To highlight the potential patterns in ranking keywords according to the type of their term functions, Figure 9 and Table 3 show the total amount of keywords in a particular ranking that made specific term functions. In Figure 9(a), it can be seen that, in papers tagged by only two author-selected keywords, both

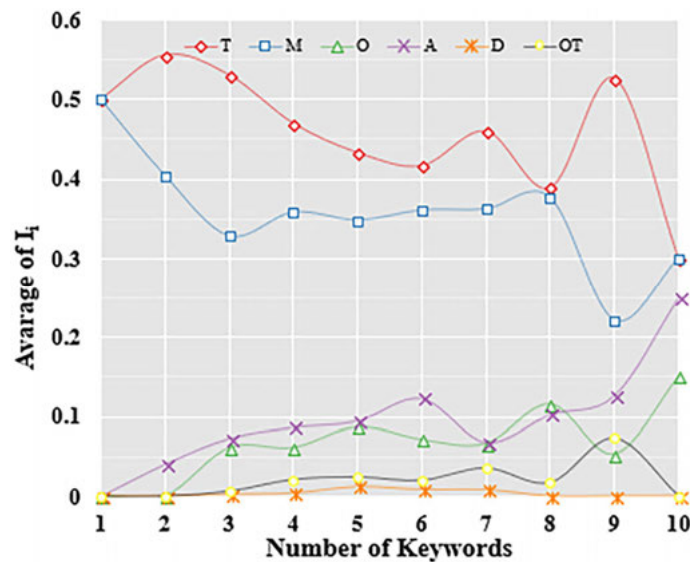


Figure 8. Average intensity of individual term function (I_i) as a function of the number of author-selected keywords (n_K) in the dataset. In general, “research topic” > “research method” > “research area” > “research object” > “data.”

n_K	Term Function (TF)				
	Research Topic (T)	Research Method (M)	Research Object (O)	Research Area (A)	Data (D)
$n_K = 2$	1 st >2 nd	1 st >2 nd	1 st >2 nd	1 st >2 nd	1 st >2 nd
$n_K = 3$	3 rd >2 nd >1 st	3 rd >2 nd >1 st	1 st >2 nd >3 rd	1 st >2 nd >3 rd	1 st >3 rd >2 nd
$n_K = 4$	1 st >2 nd >3 rd >4 th	3 rd >4 th >1 st >2 nd	1 st >2 nd >3 rd >4 th	1 st >2 nd >4 th >3 rd	2 nd >4 th >3 rd >1 st
$n_K = 5$	2 nd >1 st >3 rd >4 th >5 th	4 th >5 th >3 rd >2 nd >1 st	1 st >2 nd >3 rd >4 th >5 th	1 st >5 th >2 nd >4 th >3 rd	5 th >4 th >2 nd >3 rd >1 st
$n_K = 6$	1 st >2 nd >3 rd >4 th >6 th >5 th	5 th >6 th >4 th >3 rd >2 nd >1 st	1 st >2 nd >3 rd >4 th >5 th >6 th	1 st >2 nd >6 th >3 rd >5 th >4 th	3 rd >6 th >1 st >2 nd >5 th >4 th

Table 3. The relationship between the number of author-selected keywords tagged as specific term functions and their rankings in author-selected keyword lists.

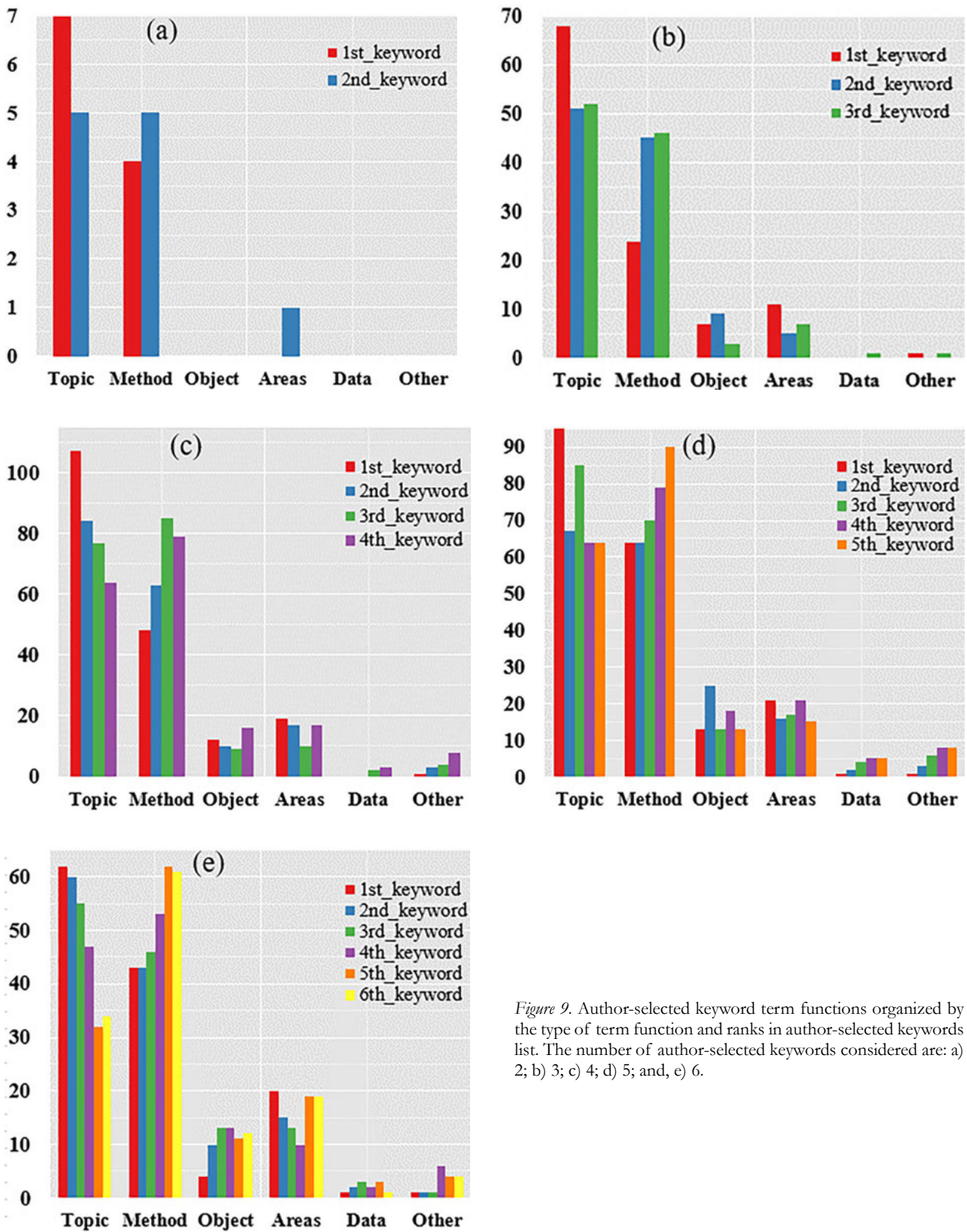


Figure 9. Author-selected keyword term functions organized by the type of term function and ranks in author-selected keywords list. The number of author-selected keywords considered are: a) 2; b) 3; c) 4; d) 5; and, e) 6.

keywords are usually tagged as “research topic,” “research method,” and “research area.” However, in most cases, the first keywords are tagged as “research topic,” as could be anticipated. Moreover, all of the “research areas” are tagged by the second author-selected keywords.

Specific term functions tagged by author-selected keywords in papers with three keywords are shown in Figure 9(b). Note that, when comparing the number of “research topic” and “research method,” the proportions are very similar. However, when considering the number according to the ranking of author-selected keywords, the first keywords obtain the largest number (1st keywords > 3rd keywords > 2nd keywords in “research topic,” 1st keywords > 2nd keywords > 3rd keywords in “research method,” which is the same as “research area”), which is different with “research object” (2nd keywords > 1st keywords > 3rd keywords) and “data” (3rd keywords > 1st keywords = 2nd keywords).

Regarding papers tagged by four author-selected keywords, as shown in Figure 9(c) it can be observed that, the least number of “research method” are tagged by first keywords. Interestingly, the second-to-last keywords take the largest number of “research method” (i.e., 3rd keywords > 4th keywords > 2nd keywords > 1st keywords). Similar patterns of contributions have also been found for papers tagged by five keywords (see Figure 9(d)) and six keywords (see Figure 9(e)). However, the first keywords are always the keywords that take the largest number of “research topic.”

According to Figure 9 and Table 3, we can summarize the several patterns relating to author-selected keyword rankings and their term functions as follows:

- 1) Pattern I: Overall, the total amount of “research topic” and “research method” keywords possesses an absolute advantage over keywords of other term functions. More specifically, when the number of author-selected keywords is less than four, the total amount of “research topic” is predominant. Meanwhile, the total amount of “research method” increases rapidly from four to more keywords, and “research topic” and “research method” are almost equal. This pattern reveals the significance of topics and methods to a scientific research in the author’s cognition, which is also in accordance with previous studies that interpret scientific research as a problem-solving activity (Heffernan and Teufel 2018; Jordan 1980). Interestingly, several studies maintain that the semantic role of all domain-independent terms in a scientific paper can be divided into topics or methods (Xin, Qikai and Wei 2017).
- 2) Pattern II: Different keyword term functions have their own preferential positions in author-selected keyword lists, although all of these keyword term functions can appear at every position. Specifically, “research topic” tends more to be tagged by keywords at the first three

positions (i.e., 1st, 2nd, and 3rd keyword in the list, see Figure 9). Conversely, “research method” keywords are more likely to appear at the last two keywords in the list. Moreover, the first two and the last two positions are where “research area” keywords always occur, which exhibits a symmetric behavior as a function of keyword ranking.

- 3) Pattern III: The number of “research topic” keywords approximately decreases with keyword ranking, while the number of “research method” keywords increases with keyword ranking. This indicates that it is easier for authors to think of the topic of the research than the methods used in the study when they index keywords.

On the whole, it can be concluded that the keyword ranking, and its term function are strongly related by evidence of the aforementioned patterns. These patterns also confirm that there is no obvious relationship between the intensity and ranking of keyword term functions, although the rank of keywords is weighted in this study, as shown in Section 3.3.1. For example, “research topic” ranked in the first positions and has the maximum intensity, on average; whereas, “research method” obtains the second largest intensity and is always tagged by last two keywords in the list. Meanwhile, from pattern I, one can find that the key factor that affects the value of intensity of individual term functions is the number of specific term function keywords in author-selected keywords lists. In addition, we note here that, since the scale of “data” keywords is very small, no obvious regularity is found.

5.0 Conclusion and future work

Although author-selected keywords have long been utilized in knowledge organization, information retrieval, social tags, keyword extraction, indexing and thesaurus development, few studies have investigated the patterns of author-selected keywords in scientific papers. However, for a more fine-grained indexing and retrieval of scientific papers, for example, retrieving studies in which co-word analysis comprises the “research topic” but not “research method,” it is necessary to identify the term functions of keywords in scientific papers. Additionally, analyzing the patterns of author-selected keywords from the term function perspective also constitutes the basis for the construction of a semantic network of keywords, which will be of great significance for knowledge organization and traditional bibliometric tasks, such as hot spot identification, trends analysis and mapping the knowledge structure of hard sciences and social sciences. Therefore, in this paper, we have mainly analyzed the potential patterns of author-selected keywords from the perspective of term function (TF).

The main contributions of this study are threefold. First, in order to investigate the patterns of author-selected keywords in scientific manuscripts, this paper, by treating the relationship between author-selected keywords and term functions as a bipartite network, proposes a new method based on the concept of accessibility and true diversity to quantify the diversity and symmetry of keyword term functions (φ and σ) at the paper level and the intensity of individual term function (I_i) at the function level. These measures can effectively describe the irregularity of author-selected keywords from the term function perspective. Second, this study also found that a strong relationship exists between a keyword's ranking and its term function. We confirmed that "research topic" and "research method" keywords are the most frequent in scientific papers. Despite this well-known pattern, three patterns of author-selected keywords are also found, depending on the relationship between the amount of specific term function keywords and their rankings. For instance, "research topic" tended to be tagged more by keywords at the first three positions. Interestingly, "research method" keywords were more likely to appear at the last two keywords in the list, which indicates that there is no obvious relationship between the intensity and ranking of keyword term functions. Third, we also designed an annotation scheme for author-selected keyword term functions, with which a corpus comprising 3,311 author-selected keywords from 693 scientific papers (all original research papers published between 2007 and 2017 in the *Journal of Informetrics*) are obtained with rigorous human annotation. Great care was taken in constructing this corpus by professionals to ensure the quality. Hence, this corpus could be valuable for the tasks of term function recognition, keyword extraction and more fine-grained co-word network analysis in the further study.

The results of this study should be interpreted in the context of its limitations. The main defect is that we analyzed the author-selected keywords only from the field of informetrics and bibliometrics. The reason for this is that the annotation of term functions manually for keywords is difficult due to its huge workload to interpret author intentions and the content of the whole article. In the future, we will perform studies that analyze and compare patterns of author-selected keywords among different natural sciences and social sciences. Furthermore, we will also investigate the patterns of other kinds of keywords from the perspective of term function, for example, KeyWords Plus in the Web of Science or MeSH (*Medical Subject Headings*) terms in PubMed. Finally, we raise an open-ended question of whether the diversity of keyword term functions (φ), the symmetry of keyword term functions (σ) and the intensity of individual term function (I_i) can affect scientific papers' citations. We believe that much room still exists for further research, and we anticipate interesting results in consequent work.

References

- Amancio, Diego R., Osvaldo N. Oliveira jr and Luciano da F. Costa. 2015. "Topological-Collaborative Approach for Disambiguating Authors' Names in Collaborative Networks." *Scientometrics* 102: 465-85. doi:10.1007/s11192-014-1381-9
- Augenstein, Isabelle, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman and Andrew McCallum. 2017. In *Proceedings of the 11th International Workshop on Semantic Evaluation SemEval-2017*, ed. Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer and David Jurgen. Vancouver, BC: Association for Computational Linguistics, 546-55. doi:10.18653/v1/S17-2091
- Baldwin, Clive, Julian Hughes, Tony Hope, Robin Jacoby and Sue Ziebland. 2003. "Ethics and Dementia: Mapping the Literature by Bibliometric Analysis." *International Journal of Geriatric Psychiatry* 18: 41-54. doi:10.1002/gps.770
- Callon, M., J. P. Courtial and F. Laville. 1991. "Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry." *Scientometrics* 22: 155-205. doi:10.1007/BF02019280
- Callon, Michel, Arie Rip and John Law. 1986. *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Cham: Springer.
- Carletta, Jean. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic." *Computational Linguistics* 22: 249-54.
- Chen, Guo and Lu Xiao. 2016. "Selecting Publication Keywords for Domain Analysis in Bibliometrics: A Comparison of Three Methods." *Journal of Informetrics* 10: 212-23.
- Choi, Jinho, Sangyoon Yi and Kun Chang Lee. 2011. "Analysis of Keyword Networks in MIS Research and Implications for Predicting Knowledge Evolution." *Information & Management* 48: 371-81.
- Choi Youngok and Sue Yeon Syn. 2016. "Characteristics of Tagging Behavior in Digitized Humanities Online Collections." *Journal of the Association for Information Science and Technology* 67: 1089-104.
- Cobo, Manolo J, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma and Francisco Herrera. 2011. "An Approach for Detecting, Quantifying and Visualizing the Evolution of a Research Field: A Practical Application to the Fuzzy Sets Theory Field." *Journal of Informetrics* 5: 146-66.
- Cobo, Manolo J, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma and Francisco Herrera. 2011. "Science Mapping Software Tools: Review, Analysis and Cooperative Study among Tools." *Journal of the American Society for Information Science and Technology* 62: 1382-402.
- Corrêa Jr, Edilson A, Filipi N Silva, Luciano da F Costa and Diego R Amancio. 2017. "Patterns of Authors

- Contribution in Scientific Manuscripts.” *Journal of Informetrics* 11: 498-510.
- Coulter, Neal, Ira Monarch and Suresh Konda. 1998. “Software Engineering as Seen through Its Research Literature: A Study in Co-word Analysis.” *Journal of the American Society for Information Science* 49: 1206-23.
- Ding, Ying. 2011. “Topic-based PageRank on Author Cociation Networks.” *Journal of the American Society for Information Science and Technology* 62: 449-66.
- Ding, Ying, Gobinda G Chowdhury and Schubert Foo. 2001. “Bibliometric Cartography of Information Retrieval Research by Using Co-Word Analysis.” *Information Processing & Management* 37: 817-42.
- Ferrara, Alfio and Silvia Salini. 2012. “Ten Challenges in Modeling Bibliographic Data for Bibliometric Analysis.” *Scientometrics* 93: 765-85.
- 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: 1175-87.
- Gil-Leiva, Isidoro. 2017. “SISA-Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules.” *Knowledge Organization* 44: 139-62.
- Gupta, Sonal and Christopher Manning. 2011. “Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers.” In *Proceedings of Fifth International Joint Conference on Natural Language Processing 8-13 November 2011 Chiang Mai, Thailand*. Asian Federation of Natural Language Processing, 1-9. <https://www.aclweb.org/anthology/I11-1>
- Han, Jiawei, Yue Huang, Nick Cercone and Yongjian Fu. 1996. “Intelligent Query Answering by Knowledge Discovery Techniques.” *IEEE Transactions on Knowledge and Data Engineering* 8: 373-90.
- He, Qin. 1999. “Knowledge Discovery through Co-Word Analysis.” *Library Trends* 48, no. 1: 133-59.
- Heffernan, Kevin and Simone Teufel. 2018. “Identifying Problems and Solutions in Scientific Text.” *Scientometrics* 116: 1367-82. doi:10.1007/s11192-018-2718-6
- Hoey, Michael. 2013. *Textual Interaction: An Introduction to Written Discourse Analysis*. Abingdon: Routledge.
- Huang, Shanshan and Xiaojun Wan. 2013. “AKMiner: Domain-Specific Knowledge Graph Mining from Academic Literatures.” In *Proceedings of 14th International Conference on Web Information Systems Engineering October 2013, Nanjing China*. Cham: Springer, 241-55.
- Jones, Steve and Malika Mahoui. 2000. “Hierarchical Document Clustering Using Automatically Extracted Keyphrases.” In *Proceedings of the Third International Asian Conference on Digital Libraries, Seoul, Korea*. Berkeley, CA: ACM Press, 113-20.
- Jordan, Michael P. 1980. “Short Texts to Explain Problem-Solution Structures and Vice Versa.” *Instructional Science* 9: 221-52.
- Keupp, Marcus Matthias, Maximilian Palmié and Oliver Gassmann. 2012. “The Strategic Management of Innovation: A Systematic Review and Paths for Future Research.” *International Journal of Management Reviews* 14: 367-90
- Khan, Gohar Feroz and Jacob Wood. 2015. “Information Technology Management Domain: Emerging Themes and Keyword Analysis.” *Scientometrics* 105: 959-72.
- Kondo, Tomoki, Hidetsugu Nanba, Toshiyuki Takezawa and Manabu Okumura. 2009. “Technical Trend Analysis by Analyzing Research Papers’ Titles.” In *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznan, Poland, November 6-8, 2009, Revised Selected Papers*. Lecture Notes in Computer Science 6562. Lecture Notes in Artificial Intelligence 6562. Berlin: Springer, 512-21. doi:10.1007/978-3-642-20095-3_47
- Law, John, Serge Bauin, J Courtial and John Whittaker. 1988. “Policy and the Mapping of Scientific Change: A Co-Word Analysis of Research into Environmental Acidification.” *Scientometrics* 14: 251-64.
- Lu, Kun and Margaret E. I. Kipp. 2014. “Understanding the Retrieval Effectiveness of Collaborative Tags and Author Keywords in Different Retrieval Environments: An Experimental Study on Medical Collections.” *Journal of the Association for Information Science and Technology* 65: 483-500.
- Matsuo, Yutaka and Mitsuru Ishizuka. 2004. “Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information.” *International Journal on Artificial Intelligence Tools* 13: 157-69.
- Mesbah, Sepideh, Kyriakos Fragkeskos, Christoph Lofi, Alessandro Bozzon and Geert-Jan Houben. 2017. “Facet Embeddings for Explorative Analytics in Digital Libraries.” In *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*, ed. Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis and Ioannis Karydis. Lecture Notes in Computer Science 10450. Information Systems and Applications 10450. Cham: Springer, 86-99.
- Milojević, Staša, Cassidy R Sugimoto, Erjia Yan and Ying Ding. 2011. “The Cognitive Structure of Library and Information Science: Analysis of Article Title Words.” *Journal of the American Society for Information Science and Technology* 62: 1933-53.
- Nanba, Hidetsugu, Tomoki Kondo and Toshiyuki Takezawa. 2010. “Automatic Creation of a Technical Trend Map from Research Papers and Patents.” In *Proceedings of the 3rd International Workshop on Patent Information Retrieval*

- 26 October 2010 Toronto, Ontario, Canada. New York: ACM, 11-16. doi:10.1145/1871888.1871891
- Névél, Aurélie, Rezarta Islamaj Doğan and Zhiyong Lu. 2010. "Author Keywords in Biomedical Journal Articles." In *AMIA Annual Symposium Proceedings 2010*. Bethesda, MD: AMIA, 537-41.
- Newman, Mark. 2010. *Networks: An Introduction*. Oxford: Oxford University Press.
- Peters, H. P. F. and Anthony F. J. van Raan. 1993. "Co-Word-Based Science Maps of Chemical Engineering. Part I: Representations by Direct Multidimensional Scaling." *Research Policy* 22: 23-45.
- Raan, Anthony F. J. van and Robert J. W. Tijssen. 1993. "The Neural Net of Neural Network Research: An Exercise in Bibliometric Mapping." *Scientometrics* 26: 169-92. doi:10.1007/bf02016799
- Ren, Feiliang. 2014. "An Unsupervised Cascade Learning Scheme for 'Cluster-Theme Keywords' Structure Extraction from Scientific Papers." *Journal of Information Science* 40: 167-79.
- Sahragard, Rahman and Hussein Meihami. 2016. "A Diachronic Study on the Information Provided by the Research Titles of Applied Linguistics Journals." *Scientometrics* 108: 1315-31.
- Schaffner, Jennifer. 2009. *The Metadata Is the Interface: Better Description for Better Discovery of Archives and Special Collections: Synthesized from User Studies*. Dublin, OH: OCLC Programs and Research.
- Silva, Filipi N., Diego R. Amancio, Maria Bardosova, Luciano da F. Costa and Osvaldo N. Oliveira Jr. 2016. "Using Network Science and Text Analytics to Produce Surveys in a Scientific Topic." *Journal of Informetrics* 10: 487-502.
- Smiraglia, Richard P. 2013. "Keywords, Indexing, Text Analysis: An Editorial." *Knowledge Organization* 40: 155-9.
- Sohrabi, Babak and Hamideh Iraj. 2017. "The Effect of Keyword Repetition in Abstract and Keyword Frequency per Journal in Predicting Citation Counts." *Scientometrics* 110: 243-2.51.
- Song, Min, SuYeon Kim, Guo Zhang, Ying Ding and Tamy Chambers. 2014. "Productivity and Influence in Bioinformatics: A Bibliometric Analysis Using PubMed Central." *Journal of the Association for Information Science and Technology* 65: 352-71.
- Su, Hsin-Ning and Pei-Chun Lee. 2010. "Mapping Knowledge Structure by Keyword Co-Occurrence: A First Look at Journal Papers in Technology Foresight." *Scientometrics* 85: 65-79.
- Tian, Yangge, Cheng Wen and Song Hong. 2008. "Global Scientific Production on GIS Research by Bibliometric Analysis from 1997 to 2006." *Journal of Informetrics* 2: 65-74.
- Travençolo, Bruno Augusto Nassif and L da F Costa. 2008. "Accessibility in Complex Networks." *Physica Letters A* 373: 89-95.
- Tsai, Chen-Tse, Gourab Kundu and Dan Roth. 2013. "Concept-Based Analysis of Scientific Literature." In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. New York: ACM, 1733-8. doi:10.1145/2505515.2505613
- Tseng, Yuen-Hsien. 2002. "Automatic Thesaurus Generation for Chinese Documents." *Journal of the American Society for Information Science and Technology* 53: 1130-8.
- Uddin, Shahadat and Arif Khan. 2016. "The Impact of Author-Selected Keywords on Citation Counts." *Journal of Informetrics* 10: 1166-77.
- Wang, Jun. 2006. "Automatic Thesaurus Development: Term Extraction from Title Metadata." *Journal of the American Society for Information Science and Technology* 57: 907-20.
- Wang, Zhong-Yi, Gang Li, Chun-Ya Li and Ang Li. 2012. "Research on the Semantic-Based Co-Word Analysis." *Scientometrics* 90: 855-75.
- Wu, Bihu, Honggen Xiao, Xiaoli Dong, Mu Wang and Lan Xue. 2012. "Tourism Knowledge Domains: A Keyword Analysis." *Asia Pacific Journal of Tourism Research* 17: 355-80.
- Wu, Chao-Chan. 2016. "Constructing a Weighted Keyword-Based Patent Network Approach to Identify Technological Trends and Evolution in a Field of Green Energy: A Case of Biofuels." *Quality & Quantity* 50: 213-35.
- Xin, L., Qikai C. and Wei L. 2017. "CS-LAS: A Scientific Literature Retrieval and Analysis System Based on Term Function Recognition (TFR)." In *Proceedings of 16th International Conference of the International Society for Scientometrics and Informetrics 16-17 March 2017, Wuhan, China*. Wuhan: Wuhan University Press, 1346-56.

Drawing a Knowledge Map of Smart City Knowledge in Academia

Feng-Tyan Lin

1239 Siping Road, Shanghai, P.R. China,
<ftlin@tongji.edu.cn>

Dr. Feng-Tyan Lin obtained his bachelor's degree in engineering (urban planning) at National Cheng Kung University, Tainan, Taiwan, in 1977. He obtained his master's and PhD degrees in computer science at Northwestern University, USA, in 1986 and 1989, respectively. Professor Lin's research has focused on information cities and computer theories which can be applied to urban planning and architectural design. He was a professor at National Taiwan University and National Cheng Kung University, Taiwan, and retired in 2015. He currently serves as an overseas co-PI in Tongji University, China.



Lin, Fen-Tyan. 2019. "Drawing a Knowledge Map of Smart City Knowledge in Academia." *Knowledge Organization* 46(6): 419-438. 27 references. DOI:10.5771/0943-7444-2019-6-419.

Abstract: This research takes the academic articles in the Web of Science's core collection database as a corpus to draw a series of knowledge maps, to explore the relationships, connectivity, distribution, and evolution among their keywords with respect to smart cities in the last decade. Beyond just drawing a text cloud or measuring their sizes, we further explore their texture by identifying the hottest keywords in academic articles, construct links between and among them that share common keywords, identify islands, rocks, reefs that are formed by connected articles—a metaphor inspired by Ong et al. (2005)—and analyze trends in their evolution. We found the following phenomena: 1) "Internet of Things" is the most frequently mentioned keyword in recent research articles; 2) the numbers of islands and reefs are increasing; 3) the evolutions of the numbers of weighted links have fractal-like structure; and, 4) the coverage of the largest rock, formed by articles that share a common keyword, in the largest island is converging into around 10% to 20%. These phenomena imply that a common interest in the technology of smart cities has been emerging among researchers. However, the administrative, social, economic, and cultural issues need more attention in academia in the future.

Received: 7 May 2019; Revised: 18 August 2019; Accepted: 23 August 2019

Keywords: articles, keywords, knowledge maps, smart cities

1.0 Motivation

In the era of the information explosion, processing information into knowledge for better management and decision making has become necessary. Many ongoing efforts explore the issues of knowledge in various fields, such as library and information science (LIS), business administration, industrial production, public health, public policy, and smart cities. Being roused by this wave, many research interests have also emerged in knowledge management (KM) as a collectively scientific discipline.

The study of KM has three interrelated aspects: methodology, ontology, and sociology. Methodologies of KM include codification, classification, tag clouds, knowledge map construction, visualization, text mining, and topological analysis. Some will be discussed further and employed later in this paper. In addition to methodology, the ontology of KM consists of organized knowledge of specific knowledge domains explored by various methods, while the sociology of KM, combining with epistemology and axiology, concerns social, cultural, organizational, and po-

litical factors associated with successful implementation of knowledge management.

Although the scope of KM ontology has been increasing, it is far from complete. As Hjørland (2008) suggests, knowledge organization (KO) should not be limited to a narrow meaning restricted to document description, indexing, classification, and organization. Rather, KO has a broader meaning related to how knowledge is socially organized and how individual sciences are organized. He claims that KO in the narrow sense cannot develop a fruitful body of knowledge without considering KO in the broader perspective. The claim also holds true in the territory of KM.

Among the various booming subjects in KM, we found that there are relatively few articles discussing the knowledge management of smart cities. Being a prevailing topic, smart cities, including smart transportation, smart public health and safety, smart education, and smart governance, etc., has attracted interest from many cities, researchers, scholars, engineers, industries, and businesses. Most of those works focus on developing frameworks, strategies, innovative tech-

nologies and devices, and application systems for constructing smart cities. Far fewer articles study the KM of smart cities, which propose conceptual visions, suggest frameworks and models, and identify key factors for building up smart city knowledge bases (e.g., Boyer 2016; Biloslavo and Zornada 2004; Brachos et al. 2007). Only a handful of articles discuss the sociological aspects of KM in smart cities (e.g., Meijer and Bolívar 2016; Jennex and Zakharova 2006). As far as we know, no research exists on keyword distribution in academic research articles.

To remedy this, we will use the academic articles in the Web of Science core collection database as a testbed to explore the relationships, connectivity, distribution, and evolution among their keywords associated with smart cities, published in the last decade. The remainder of this paper is organized as follows: In section 2 we briefly review some related work; in section 3 we introduce the methods used in this research, including definitions, the analysis process, and the database used; and in section 4 we present the results. Finally, in section 5 we draw conclusions.

2.0 Related work

2.1 The character of knowledge

Knowledge has many properties, including dispersion, evolution, reusability, and guidance. Knowledge to accomplish a job is dispersed through various organizations and staff in different disciplines. Therefore, it needs cooperation among persons from different departments or even outside experts. Knowledge sharing and communication is a key factor for completing a task successfully (e.g., Liu et al. 2019, Ahmed et al. 2019). Knowledge also evolves. In other words, knowledge has a dynamic nature and cannot be static. It continually changes with human experiences, technology advancement, knowledge explication, researchers' perspectives, and social interactions (McInerney 2002). Moreover, many tasks are repeated with minor variations in different contexts. The knowledge from previous tasks can be reused and adapted for new cases. New employees can learn from the experiences of similar cases completed previously by other colleagues so that the task at hand can be carried out correctly, efficiently, and effectively. If knowledge of typical experiences is recorded in an understandable format for transferal to new staff, it will largely improve the work quality of an organization. The study of knowledge has also attracted researchers' attention so that the current situation and front edge of research can be identified for guiding future work. For example, Scharnhorst et al. (2016) captured how knowledge and knowledge systems of UDC changed over time and raised some further questions for future work.

The knowledge of smart cities has all the properties mentioned above. Articles bearing knowledge are published in academic journals, conference proceedings, and magazines. Keywords provided by authors emerge, are repeated and reused, change, and evolve. Thus, studying the knowledge concerning smart city keywords should reveal the corresponding phenomena of these properties through the construction of knowledge maps with capabilities of visualization and text analytics.

2.2 Visualization

Visualization, one of the most popular approaches, acts as a collaboration catalyst to capture the big picture of dispersed knowledge for sense making and knowledge sharing (Eppler 2013). While many word cloud visualization tools deal with individual words, Heimerl et al. (2014) took it a step further to develop a prototypical system, called Word Cloud Explore, that employs linguistic knowledge about the words and their relationships for text analysis, such as multiword expression identification, term statistics, co-occurrence highlighting, and provision of linguistic information.

Many processes for classifying raw text-based materials and interpreting the visualized result are still manual. For example, Toronto 311, a non-emergency service in a smart city, maintained an online knowledge base composed of 21,000 web pages. However, these web pages were unstructured texts, and thus, not machine-readable and difficult to reuse. To recognize the knowledge requirements of the city government, Allahyari et al. (2014) manually analyzed and identified ten knowledge patterns extracting from more than 500 Toronto 311 web pages according to their importance and frequency. In another example unrelated to smart cities, Scharnhorst et al. (2016) employed a color-coding scheme to visualize complex networks of category systems of Wikipedia and Universal Decimal Classification (UDC) so their differences could be compared using human eyes.

"Map" is a geographical term, which we borrow to describe objects in knowledge maps. Ong et al. (2005) mentioned that a knowledge map had an ocean-and-island metaphor, and the size of an island provided an estimate of the number of articles contained in a category. However, they did not explore this issue further. In this article, we take a similar analogy from geographers and geologists. Not only is the metaphor of islands referenced and their sizes are measured, but also the texture of islands is further explored by studying the evolution trend of the hot keywords, the strength of connectivity among articles, and the coverage of the biggest rock formed by completely interconnected articles over the islands.

2.3 Knowledge mining and mapping

Knowledge mining is a family of methods used to reveal the structure of knowledge embedded in a mass of unorganized documents by constructing relationships of the co-citation or co-occurrence of tags, which can be words appearing in titles, abstracts, keyword lists, or full texts.

Many knowledge-mining techniques exist. Medelyan (2018) illustrates five common approaches to disclose the internal structure of unorganized materials, namely, word spotting, rules for pattern matching, text categorization, topic modelling, and thematic analysis. Cheng et al. (2018) summarize knowledge-mining techniques into two categories: statistical analysis-oriented, such as k-means and k-nn (Chemchem and Drias 2015), and knowledge discovery-oriented, such as machine learning (Ong et al. 2005).

A knowledge map results from knowledge mapping. The layout of a knowledge map may be arranged in a sequential line-by-line form, a tree, a circle, or a complex network. Concepts or words of knowledge may be organized by alphabetical order, occurrence frequency, or semantic proximity in different fonts, sizes, weights, colors, and places for readers to easily capture the whole structure and perform tasks, such as searching, browsing, impression formation, recognition, and matching (Bateman et al. 2008; Gambette and Véronis 2010; Rivadeneira et al. 2007; Heimerl et al. 2014).

Knowledge mapping is an essential subfield of knowledge management and has been applied to many fields. It assists public and private organizations and academic and research communities understand the whole picture of scattered knowledge retained in different departments or places with the purpose of making strategic plans, transferring knowledge and learning experiences, inspiring brainstorming, and stimulating new knowledge. For example, by analyzing the number of papers downloaded from the arXiv in the “artificial intelligence” (AI) section through 18 November 2018, Hao (2019) classified the research history of AI into three major trends. To cope with rapid growth and ever-changing knowledge in the field of smart production, Cheng et al. (2018) discussed and suggested the application of knowledge mapping in production management, while Su and Jiang (2007) applied it to assisting fuel pump design. Su and Jiang (2007) suggested a product design task-oriented knowledge organizing method. Liu, et al. (2009) developed a virtual collaboration platform for enterprise knowledge construction by allowing members to tag their documents, and then asked a domain expert to draw a domain knowledge map based on tags collected from members’ contributions.

The research on knowledge maps of smart cities is very rare and needs to be a dedicated topic. Balaid et al. (2016) systematically reviewed the development status of know-

ledge mapping. He concluded that the study of knowledge mapping was still in an early stage, and a large portion of existing research only covered very limited disciplines. In the field of smart cities, this observation is also true, where Mora, Deakin and Reid (2018) remains a singular work. They mapped a network structure of publications in the field of smart cities in the period from 1992 to 2012 by combining co-citation clustering and text-based analysis. They identified five major thematic tracks in the publications concerning smart cities, namely experimental, ubiquitous, corporate, single, and holistic. Their work is closely aligned with ours, in that we also are interested in drawing a knowledge map of smart city research, analyzing its evolution in the last decade and identifying hot topics. However, some major departures differentiate the two studies, including timespan, inclusion of tags, article selection criteria, research methods and findings. We will further compare their work with ours in Section 5.1.

3.0 Method

3.1 Definition

A knowledge map is an application of graph theory that studies the topology of nodes and links. Knowledge mapping has two kinds of nodes: articles and keywords. Thus, there are also two kinds of links: single and composite links. A single link is established between two articles based on a common keyword. A composite link is composed of one or more simple links between two articles. In other words, while many simple links may exist between two articles due to sharing many common keywords, there is at most one composite link between two articles.

Articles are loosely connected, like an island, if they are directly or indirectly connected by composite links. The size of an island is the number of articles of it. In an island, one may find rocks, where articles share a common keyword. At the same time, there may be a lot of singular articles without any common keywords with other articles. The singular articles are called reefs and are not considered islands. Formal definitions of these concepts are given as follows.

Let D be a set of articles p_1, p_2, \dots , denoted as $D = \{p_1, p_2, \dots\}$, the frequency of a keyword k in D , $freq(k, D)$, is the number of occurrences of k in P . It is noted that there are no duplicated keywords in an article. In other words, $freq(k, D)$ also indicates the number of articles in D that share a common keyword k . Furthermore, let article p_i have keywords $K_i = \{k_{i1}, k_{i2}, \dots\}$, and p_j have $K_j = \{k_{j1}, k_{j2}, \dots\}$. If there is a keyword k , where $k \in K_i$ and $k \in K_j$, ie., p_i and p_j share a common keyword k , p_i and p_j are linked with respect to k , and a simple link $L(p_i, p_j, k)$ is established. A composite link $CL(p_i, p_j)$ between articles p_i and p_j is a composition of

all the simple links between them. The weight of a composite link $CL(p_i, p_j)$ is denoted as $WL(p_i, p_j)$; that is the number of simple links between p_i and p_j . $CL(p_i, p_j)$ is also called the “strength” between p_i and p_j , since it is the number of common keywords of p_i and p_j . The stronger the strength $CL(p_i, p_j)$ is, the more the common keywords. Two articles p_i and p_j are “loosely connected” if: 1) there exists a simple link $L(p_i, p_j, k_w)$ directly connecting them; or, 2) there exists simple links $L(p_i, p_i, k_w)$, and $L(p_j, p_j, k_w)$, where p_i and p_j are indirectly connected via p_i through possibly different keywords k_w and k_w . However, when $k_w = k_w$, it is said that p_i and p_j are “strongly connected.” The connectivity of an article p_i , $con(p_i)$, is the number of composite links to other articles, while the weighted connectivity of an article p_i , $wcon(p_i)$, is the total number of links to other articles. An “island” is a set of loosely connected articles, while a “rock” is a set of strongly connected articles. It is noted that a rock is a complete graph, where all the elements of the rock are linked to each other. The size of an island or a rock is the number of articles belonging to it. The “coverage” of a rock on an island is defined as the size of the rock divided by that of the island. A reef is a singular article that has no link, or common keyword, to any other articles. The size of a reef is always 1. In Figure 1, article p_1 has keywords $K_1 = \{a, b, c, d, e, f, g\}$, article p_2 has keywords $K_2 = \{a, g, w\}$. Thus, there are two links $L(p_1, p_2, a)$ and $L(p_1, p_2, g)$ between p_1 and p_2 . Furthermore, the composite link $CL(p_1, p_2)$ is composed of $L(p_1, p_2, a)$ and $L(p_1, p_2, g)$ with weight $WL(p_1, p_2) = 2$. Meanwhile, the composite link $CL(p_1, p_3)$ has weight $WL(p_1, p_3) = 1$, since there is only one link between them. As a result, the connectivity of p_1 is

$con(p_1) = 5$ with weight $wcon(p_1) = 7$. In Figure 1, $p_2, p_7, p_8,$ and p_9 are strongly connected as a rock with a size of four, since they share a common keyword w and form a complete graph. Meanwhile, p_2 and p_3 are loosely connected, although they do not have any common keywords, but they share different keywords, say a and b , with p_1 . In this way, $p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8,$ and p_9 are loosely connected as an island with a size of nine. The coverage of rock $p_2, p_7, p_8,$ and p_9 on the island is 0.44%. There is another island formed by p_{10} and p_{11} with a size of two. There is a reef p_{12} that does not share any keywords with any other articles. It is noted that a reef is not taken as an island.

3.2 Analysis process

To study the evolution of knowledge associated with a topic of interest, which is represented by an exact keyword or keywords, there are three stages: construction and enumeration of knowledge maps, analyses of temporal knowledge maps, and interpretation.

3.2.1 Construction and enumeration of knowledge maps

In this stage, articles and their keywords are retrieved, and maps are constructed and enumerated based on a given keyword KW, which is the core concept on which the study focuses. We take KW as an initial keyword to retrieve all the articles D_i whose titles or keywords contain KW from a journal database in a certain time interval t , say one year.

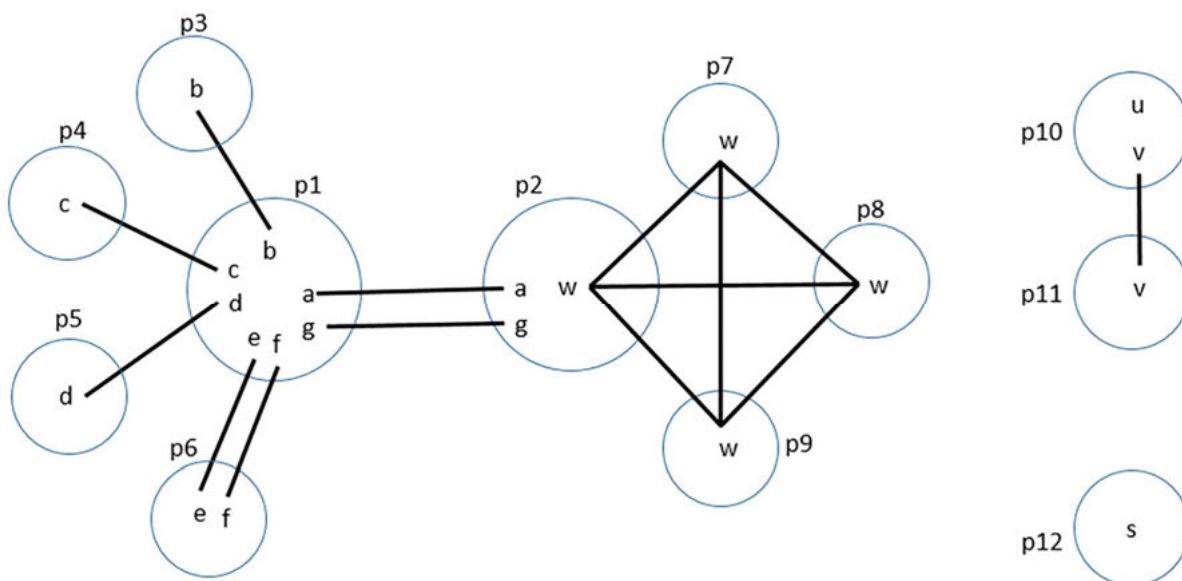


Figure 1. Explanatory diagram of a knowledge map

The journal database may cover several consecutive time intervals. Knowledge maps are constructed for each time interval, such that a temporal evolution can be analyzed.

Since KW will appear in every article under this situation, all the articles will be connected and form a single big stone island due to the common keyword KW. It turns out to be a trivial problem. Thus, the given keyword KW should be removed from data set D_i , and it can be thought of as the scope of the study. Then, the frequency for each remaining keyword, the simple and composite links between pairs of articles, the degree and their weights can be constructed and enumerated. Finally, islands, rocks, and reefs are identified, and their numbers are enumerated. The detailed procedure is illustrated below.

1. For a given keyword KW, retrieve all the articles whose titles or keywords contain KW from a journal database during a certain time period, which can be divided into several time intervals. Let D_i be the set of articles retrieved from time interval t .
2. For each time interval t
 - 2.1 Extract the set of keywords K_i for each article p_i from D_i .
 - 2.2 Transform synonyms, the original compound nouns of abbreviations, acronyms, and initials into standard keywords in lowercase letters.
 - 2.3 Remove the given keyword KW from all the K_i .
 - 2.4 Let K be the union of all the keywords in D_i ; i.e., $K = K_i \cup K_j \cup \dots = \{k_{i1}, k_{i2}, \dots\} \cup \{k_{j1}, k_{j2}, \dots\} \cup \dots$
 - 2.5 For every keyword $k \in K$ associated with D_i , count its frequency $freq(k, D_i)$.
 - 2.6 For every pair of articles p_1 and p_2 , make a simple link $L(p_i, p_j, k)$ between them if they share a common keyword k .
 - 2.7 For every pair of articles p_1 and p_2 , make a composite link $CL(p_i, p_j)$ between them if any simple link $L(p_i, p_j, k)$ exists.
 - 2.8 For every article $p \in D_i$, count its degree $deg(p)$ and weighted degree $wdeg(p)$.
 - 2.9 Identify islands, count the number of islands and the size of each island of D_i .
 - 2.10 Identify number of reefs of D_i .
 - 2.11 For every island, identify internal rocks with respect to different keywords; count the number of rocks and the size of each rock.

3.2.2 Analysis of temporal knowledge maps

After constructing and enumerating a temporal series of knowledge maps within the scope of a given keyword KW, several analyses can be performed:

- What keywords have the highest frequency? Do they change over the course of time?
- How many islands and reefs do the articles form?
- What are the sizes of the islands from the largest to the smallest? Are any trends evident?
- What is the relationship between the largest island and rock? Can one find the largest rock in the largest island?
- What is the highest strength (weighted link) between two articles?

3.2.3 Interpretation

Finally, one may interpret the results of the analysis in terms of domain knowledge. For example, if the scope of interest is “smart city” (the given keywords), some interpretations and queries may be made as follows.

- What are the hottest terms? Do they change over the course of time?
- What terms are emerging? What terms are fading out?
- Are there competing groups within the interested topic?

3.3 Datasets and software

In present research practices, researchers consult academic databases and use various tools for their research work. Many databases, such as Web of Science, Scopus, Crossref, ArXiv, etc., collect academic articles. Additionally, many tools, such as VOSViewer, CiteSpace, HistCite, SciMAT, Sci2, etc., visualize and analyze the relations among articles in databases (Chen 2017). Although they possess friendly user interfaces, convenient analysis functions, and colorful visual windows for dynamic layouts, they are general-purpose software tools, insufficient to support analyses where particular characteristics of specific disciplines need customized considerations. For instance, in this paper, instead of using a fixed selection criterion, we must choose different percentages of keywords as the hottest keywords in different time periods due to different total numbers of articles and keywords in different years.

We used the Web of Science core collection database (<https://clarivate.com/products/web-of-science/web-science-form/web-science-core-collection/>) as a testbed, which collects articles mainly from academic journals and conferences. Web of Science provides two methods to access their databases. The authorized users either visit their web pages or gains access through API to download retrieved articles with their titles, authors, publication names, year of published, organizations, and other auxiliary information after specifying an interested database, search words, timespan, and citation indices. In our research, articles with the keywords “smart city” and published in the period from 2009 to 2018 were selected. Keywords in aca-

demographic articles are semi-structured, freely provided by the authors, and composed of an indefinite number of single or compound nouns. Individual articles and keywords are two study units, as shown in Figure 1, for further analysis using the process described in Section 3.2.

4.0 Results

4.1 Graphical knowledge maps of islands

As mentioned above, knowledge maps can be presented in graphical figures or textual lists. In this subsection, we present graphical knowledge maps of the years 2009 and 2018 in Figures 2 (a) and (b), respectively. Each dot represents an article. Each link denotes that there are common keywords between the two linked articles, while its thickness represents its strength or weight, i.e., the number of common

keywords. Note that singular nodes, having no common keyword with any other articles, are not shown in the figures. Figure 2(a) is relatively simple to read and understand, as there are only thirteen articles in three islands. In this case, all the weights of the links are equal to one. However, Figure 2(b) is quite messy. There is a big black “rock” and several smaller black “rocks” in the biggest island, while some much smaller islands line the lower left. Although some thicker links can be seen in the figure, it is almost impossible to identify and count them by visualization only.

4.2 Frequency and coverage

In early years, few articles mentioned smart cities (Figure 3(a)); however, the numbers have increased dramatically in the last decade. Figure 3(b) shows that the number of articles with keywords quickly increased in the last decade

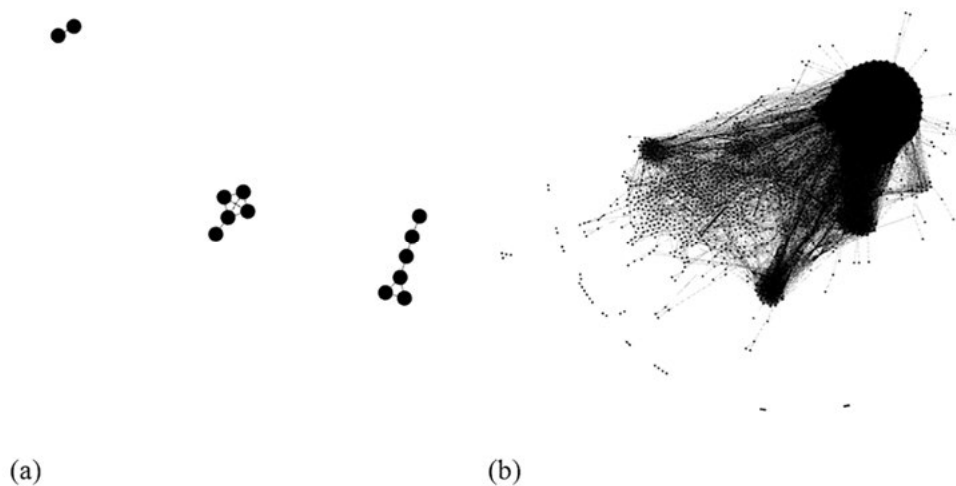


Figure 2. Knowledge maps of year 2009 (a) and 2018 (b).

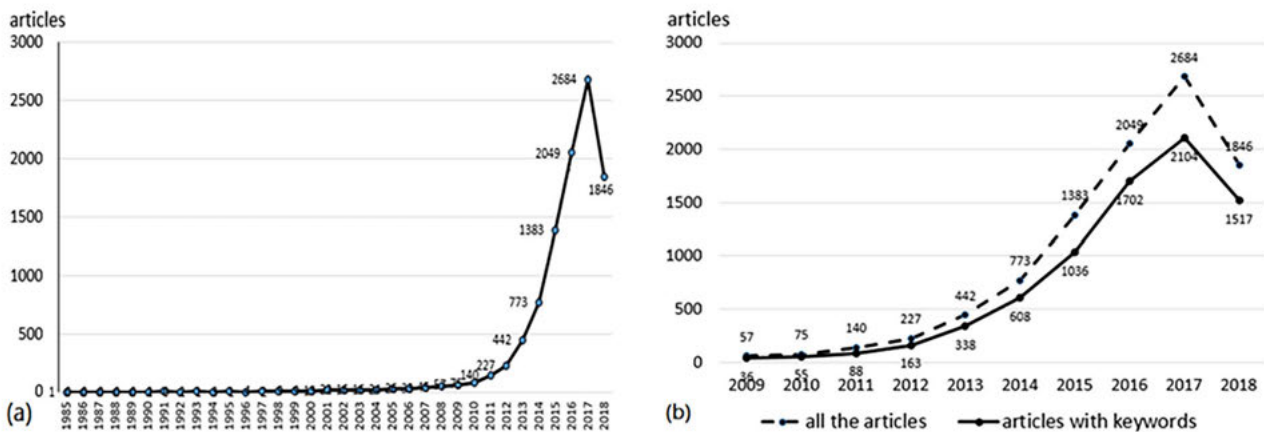


Figure 3. Number of articles since 1985 (a), and since 2009 (b).

except for a slight decrease in 2018. We will further analyze the evolution in the last decade in the remainder of this paper. It is noted that not all the articles provide their keywords. Also, the number of keywords, where duplicate keywords in different articles are counted only once, follow a similar trend; however, it drops earlier in the year 2016 (Figure 4). Although the number of both articles and keywords being used by the authors dropped in the last few years, at this moment it is hard to say whether they will continue to decrease in the near future.

It will be very interesting to know which keywords are most commonly mentioned and examine the evolution of them. For being manageable, the number of hot keywords should be limited. In this research, due to different amounts of articles in different periods, different criteria are needed to select keywords for a meaningful compari-

son. In this research, the number of articles with keywords in the first half of the decade is relatively smaller than that of the second half. It calls for different criteria to choose keywords from the first and last halves of the decade for identifying “hot” keywords and their associated trends. As a rule of thumb, the keywords that occurred more than once each year in the first half are chosen, while more than five times in the second half of the decade. As a result, the number and percentage of hot keywords increased in the first half of the decade from seven (4.00%) to 122 (9.8%), while the number of hot keywords in the second half varied from twenty to ninety-five, and their percentages are kept in the range between 1% to 2%. The hot keywords in the year 2013 were chosen by both of criteria of the first and second half of the decade. The gap between these two criteria is 8.27% (Figure 5).

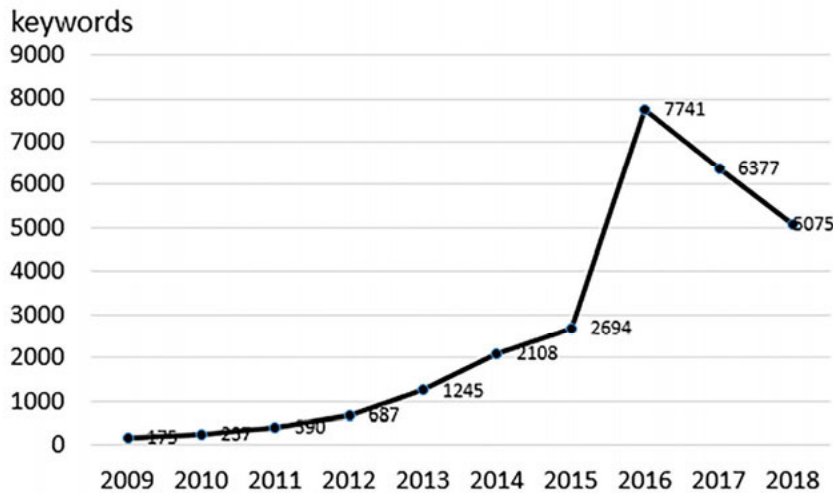


Figure 4. Numbers of keywords.

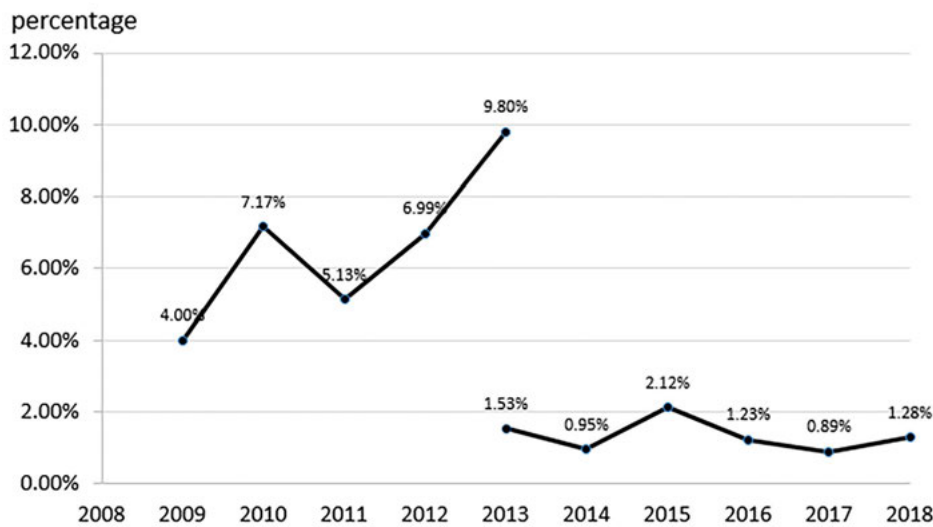


Figure 5. Percentage of hot keywords against all keywords.

As an illustration, Table 1 shows the evolution in major hot keywords, whose frequencies are in the top three highest in any year of the last decade. The percentages below the frequencies are quotients of frequencies divided by the number of articles with keywords of their corresponding years. Thus, we call the percentages “coverages.” In this way, we identify the thirteen major hottest keywords, namely: IoT, big data, cloud computing, sustainability, smart grid, ICT, urban development, smart growth, GIS, tourism, ubiquitous computing, smart planet, and u-city.

The second column of Table 1 indicates the properties of these keywords, where “T” means technology that smart cities employ, while “V” means values that smart cities pursue. There are seven keywords with “T” and six with “V.” Although they seem roughly equal, but as a matter of fact, keywords concerning value is overwhelmed by those

concerning technology in terms of frequency. For example, in 2018, keywords with “T” cover 32.9%, while keywords with “V” only cover 4.48% of the articles with keywords.

The hottest keywords shifted yearly. In the first two years, they were “smart growth,” while in the second two years they were “smart grid.” The frequencies of the hottest keywords in the first four years were relatively small. During the years 2013 to 2018, “IoT” (Internet of Things) held the position of the hottest keyword with a trend of increasing frequencies and percentages against the numbers of articles with keywords.

Figures 6 and 7 illustrates the evolving trends of the thirteen hottest keywords in terms of frequencies and their coverage in the corresponding years. Figure 6 shows that the differences between the frequencies of the top and

years		2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
number of articles with keywords		36	55	88	163	338	608	1036	1702	2104	1517
IoT	T				9	34	56	106	245	340	288
					5.52%	10.06%	9.21%	10.23%	14.39%	16.16%	19.98%
big data	T					4	16	46	83	96	70
						1.18%	2.63%	4.44%	4.88%	4.56%	4.61%
cloud computing	T				5	8	21	41	73	47	48
					3.07%	2.37%	3.45%	3.96%	4.29%	2.23%	3.16%
sustainability	V		3	3	3	9	18	19	31	47	48
			5.45%	3.41%	1.84%	2.66%	2.96%	1.83%	1.82%	2.23%	3.16%
smart grid	T			11	18	21	29	53	76	61	44
				12.50%	11.04%	6.21%	4.77%	5.12%	4.47%	2.90%	2.90%
ICT	T					10	12	14	29	8	28
						2.96%	1.97%	1.35%	1.70%	0.38%	1.85%
urban development	V			3					16		7
				3.41%					0.94%		0.46%
smart growth	V	4	9		10	5	6		8		7
		11.11%	16.36%		6.13%	1.48%	0.99%		0.47%		0.46%
GIS	T	2	4			6	9	8	15		6
		5.56%	7.27%			1.78%	1.48%	0.77%	0.88%		0.40%
tourism	V			3		3		6	9		6
				3.41%		0.89%		0.58%	0.53%		0.40%
ubiquitous computing	T	3			5	2		6			
		8.33%			3.07%	0.59%		0.58%			
smart planet	V			3							
				3.41%							
u-city	V		4								
			7.27%								

Table 1. Thirteen major hottest keywords.

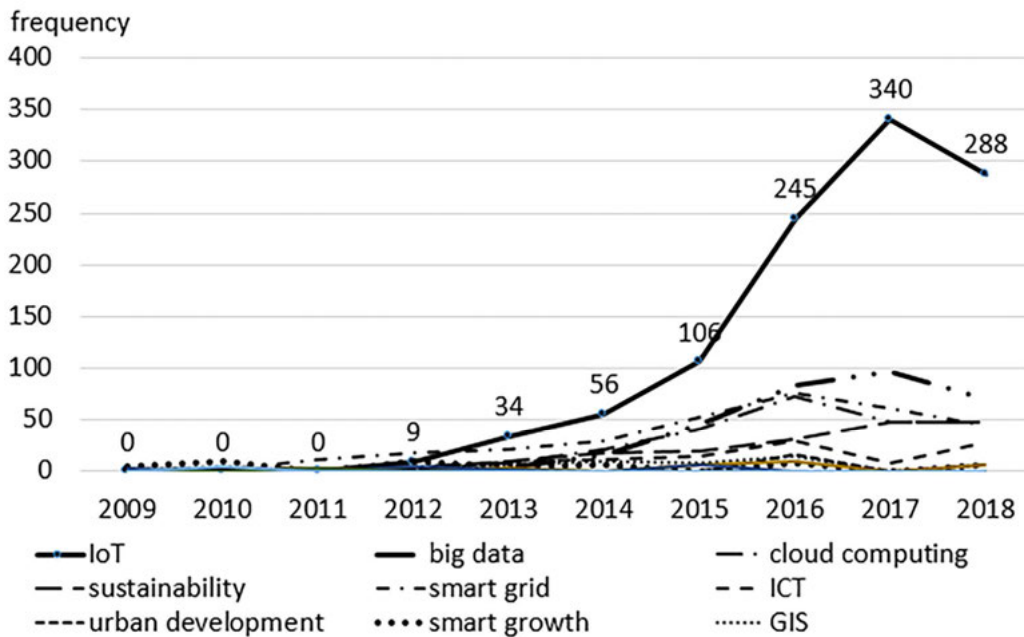


Figure 6. Frequencies of keywords.

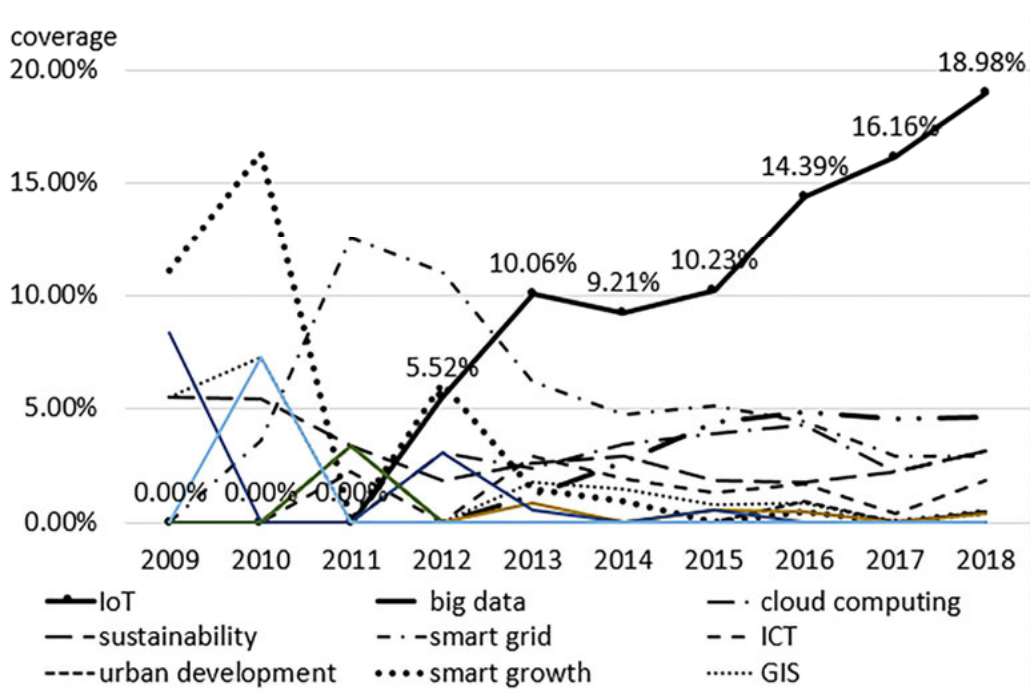


Figure 7. Coverage of keywords w.r.t all the articles.

second hottest keywords grew bigger and bigger. They can be classified into four groups. IoT, the only one member in the first group, obviously dominates the others. Some other keywords, such as “smart growth,” had high coverage in the early years, but quickly shrank later. Big data,

cloud computing, sustainability, smart grid, and ICT can be treated as the second group, which remain their most coverages between 1.5% and 5%. Among them, “sustainability” is the only value-oriented keyword and has had increasing coverage in recent years, while the others are tech-

ology-oriented and have flat or dropping tendencies. Urban development, smart growth, GIS, and tourism comprise the third group. Their frequencies and coverages are relatively smaller but still mentioned recently. The fourth group, consisting of ubiquitous computing, smart planet, and u-city seems to have vanished. It is noticeable that most of the keywords in groups three and four are value-oriented. In other words, in earlier years, researchers focused their efforts from exploring the meaning of smart cities to enriching the value of cities. Gradually, the focus shifted to technology-related issues for making smart cities a reality. Following the same idea, the evolution of any keyword can be explored. In the future work, researchers may further vary the definition of hot keywords and get differ-

ent sets of them so that their evolution patterns can be further explored.

4.3 Knowledge islands

Connected articles form knowledge islands. As mentioned earlier, if two articles share any keywords, they are linked. Either of them can further link to another article. Thus, consecutively linked articles form an island. The size of an island is at least two articles. Any article that does not link to any another article becomes a reef. In the study concerning smart cities, the numbers of islands (Figure 8), reefs (Figure 9), and articles in islands (Figure 10) simultaneously increase. That means while the number of islands

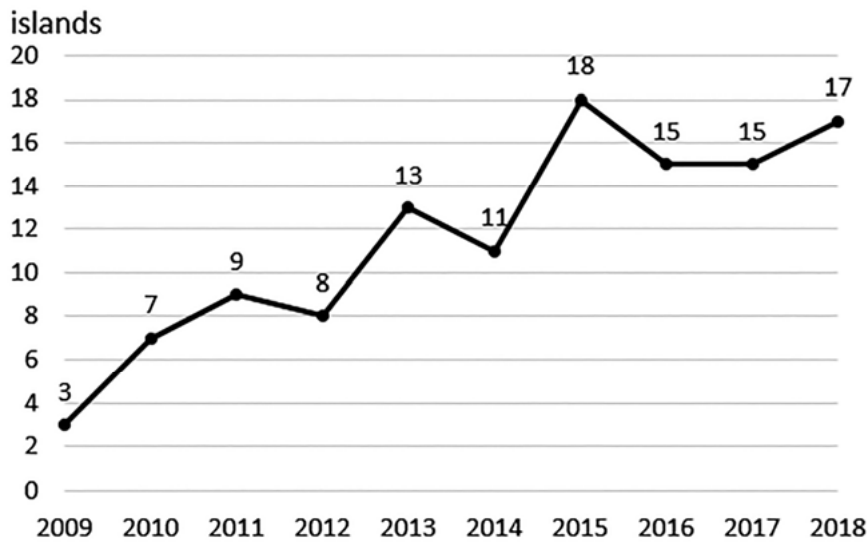


Figure 8. Numbers of islands.

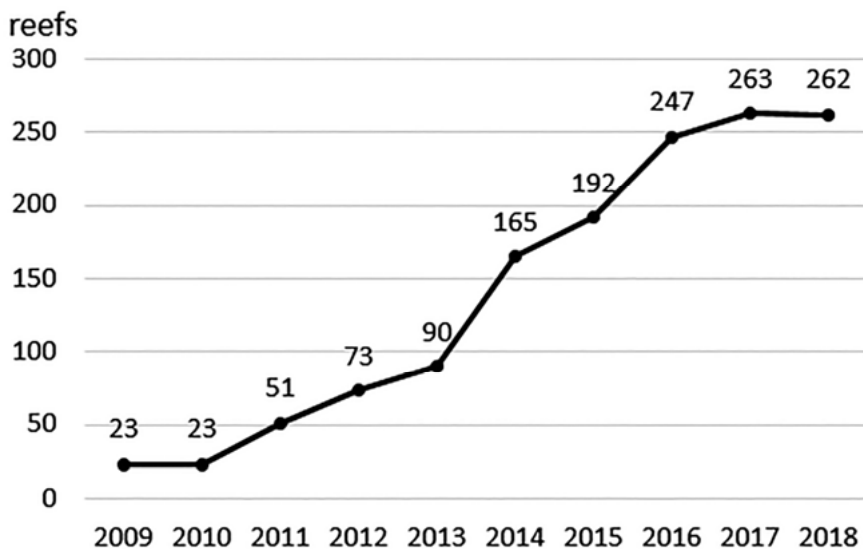


Figure 9. Numbers of reefs.

increases yearly, the largest islands are bigger and bigger, and many new standalone keywords (reefs) also emerge. Figure 10 further shows that the gap between the numbers of all the articles and articles in islands is exactly the num-

ber of reefs. Figure 11 also shows that the percentages of all the articles and articles in the largest islands against the total numbers of articles with keywords converge to around 85% gradually. In other words, there is a space of

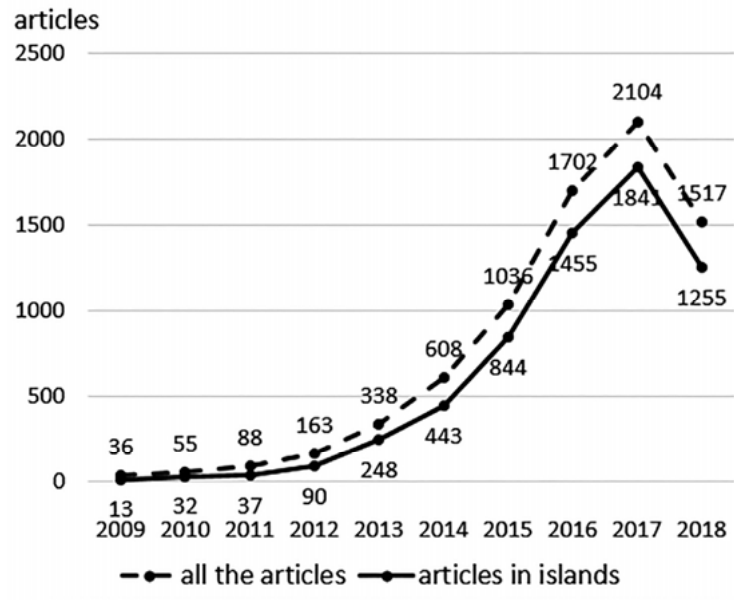


Figure 10. The numbers of all the articles and articles in islands.

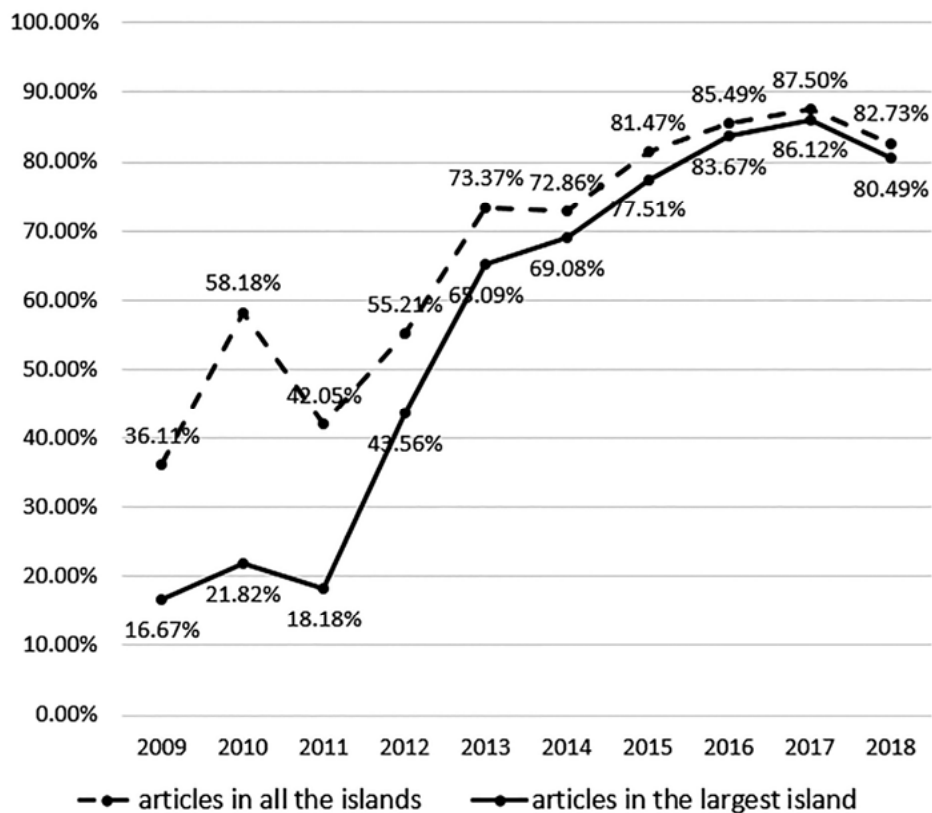


Figure 11. The percentages of all the articles and articles in the largest islands against the total numbers of articles with keywords.

around 15% for much smaller islands and reefs to keep the field of smart cities growing. Figure 12 illustrates how the percentages of number of reefs out of all the articles in corresponding years are also decreasing and converging to a range between 10% and 20%.

4.4 Island size

The size of a knowledge island is the number of articles on it. In the research field concerning smart cities, the sizes of the largest islands are much larger than those of the other islands. Thus, they are separately illustrated in two figures. In Figure 13, the sizes of the largest islands and the number of all the articles, which are shown on the polylines, dramatically increase in a similar trend except the last year. On the other hand, Figure 14 illustrates that the yearly sizes and numbers of the second-, third-, and fourth-largest islands are stable and much smaller than those of the largest islands. Furthermore, Figure 15 shows that the smaller the islands, the more numerous they are. We can imagine that the knowledge map of smart city knowledge is composed of an exceedingly large island, several much smaller islands, and a lot of reefs.

4.5 The strength of links

The weight, or strength, of a composite link is measured by the number of common keywords between two articles

that the composite link connects. Table 2 depicts that numbers of links associated with weight from one to ten during years 2009 to 2018. It shows that most of the links are of weight = 1, and in the second half of the decade the number of links with weight = 2 are less than 3% of those of weight = 1. Furthermore, Figures 16(a)-(d) illustrate the evolutions of the numbers of links with weights from one to five. We find that they have a fractal-like structure. In other words, the relative structure of evolution curves between weight = 1 and the others (Figure 16(a)) and is similar to that between weight = 2 and weights = 3,4,5 (Figure 16(b)). This phenomenon is also held between weight = 3 and weights = 4,5 (Figure 16(c)). Their correlation coefficients between consecutive weight links are calculated in Table 3, which shows that they are highly correlated.

4.6 Connectivity

The connectivity of an article measures how many other articles have common keywords with it. Figure 17 illustrates the yearly evolution of the top three articles with the highest connectivity and the reef's connectivity. The yearly top three articles with the highest connectivity are very close, and their evolutionary trend is very similar to that of the numbers of articles. Since the connectivity of reefs is zero, their evolution line lies in the x-axis. Figure 18 illustrates the number of articles with respect to their connectivity. The number of articles with connectivity less than twenty

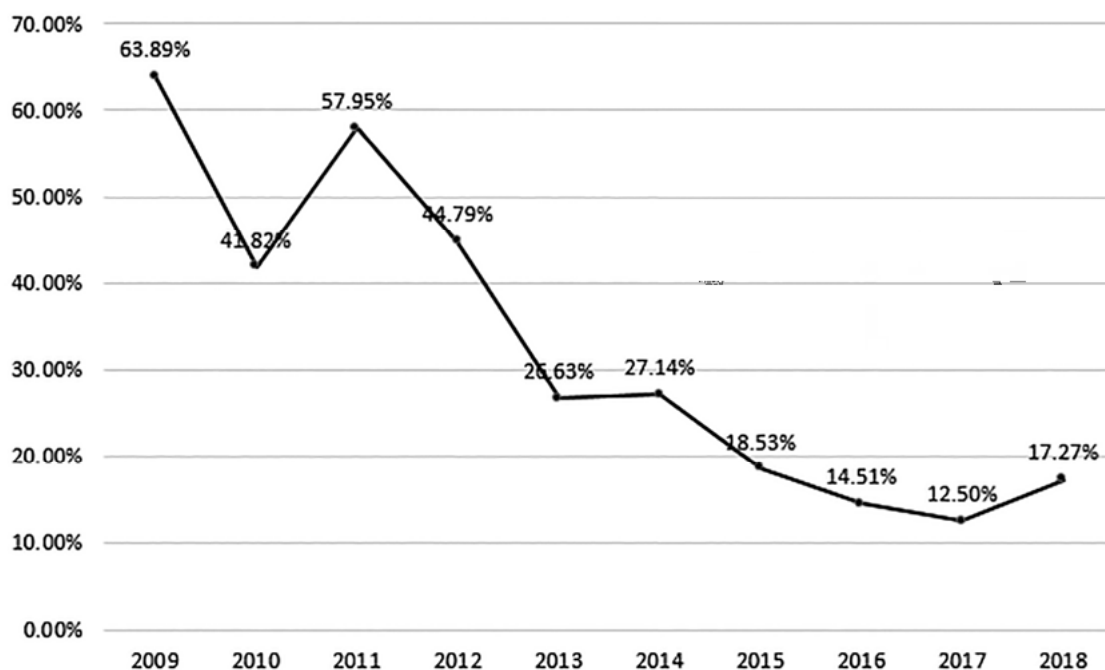


Figure 12. The percentage of number of reefs out of all the articles.

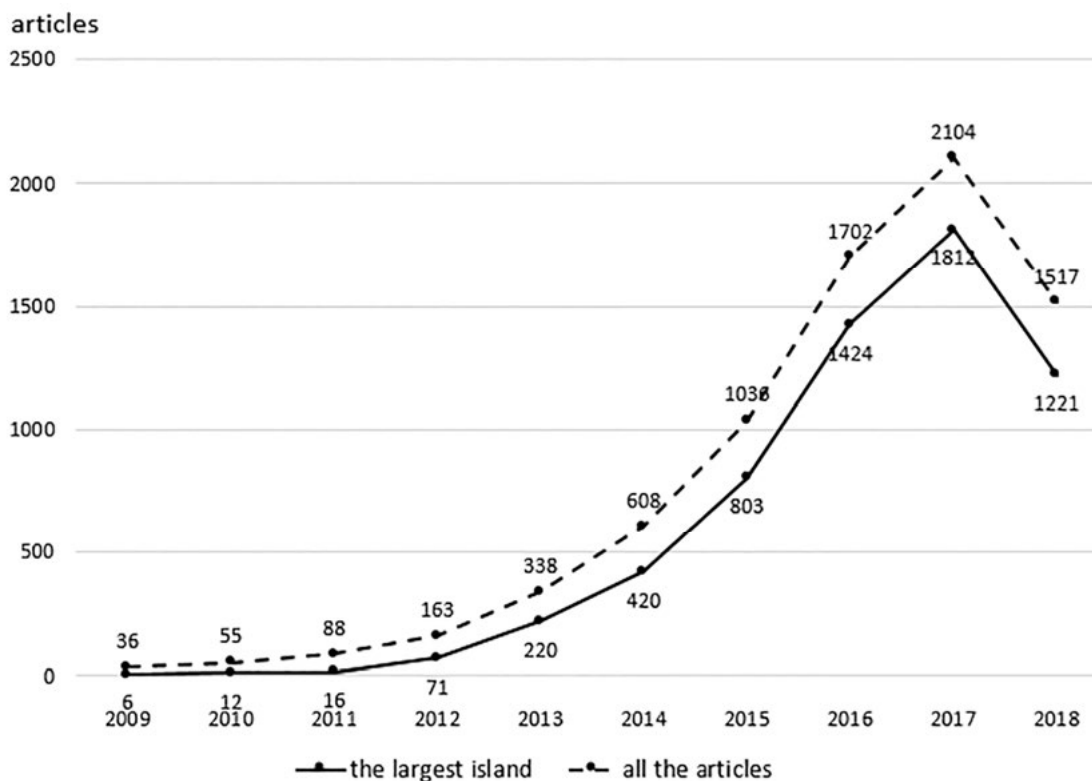


Figure 13. The yearly evolution of sizes (Y-axis) and numbers (on polylines) of the largest islands comparing to the numbers of all the articles.

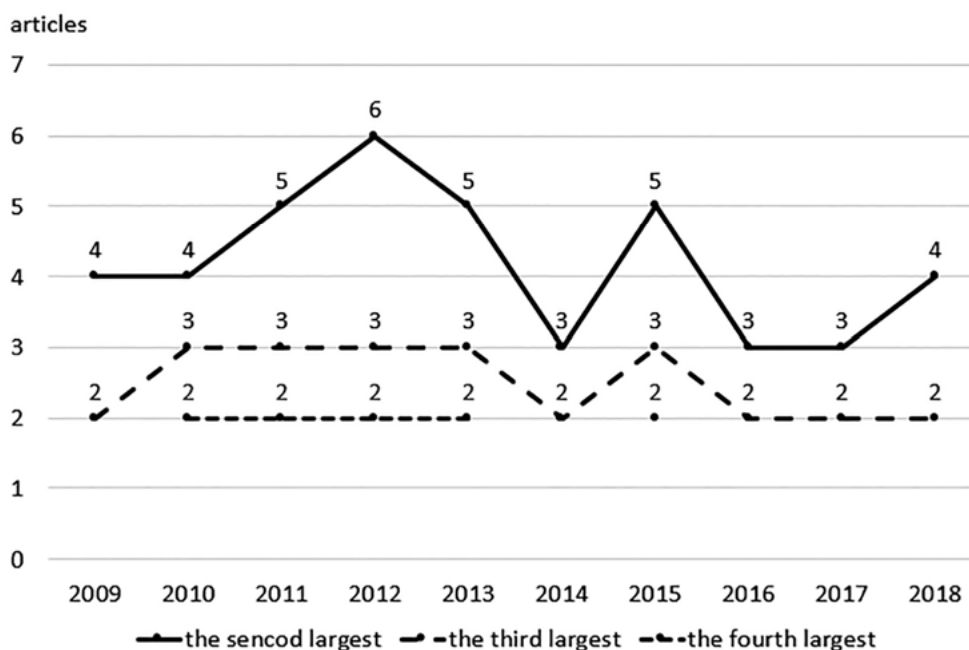


Figure 14. The yearly evolution of sizes (Y-axis) and numbers (on polylines) of the 2nd, 3rd, and 4th largest islands.

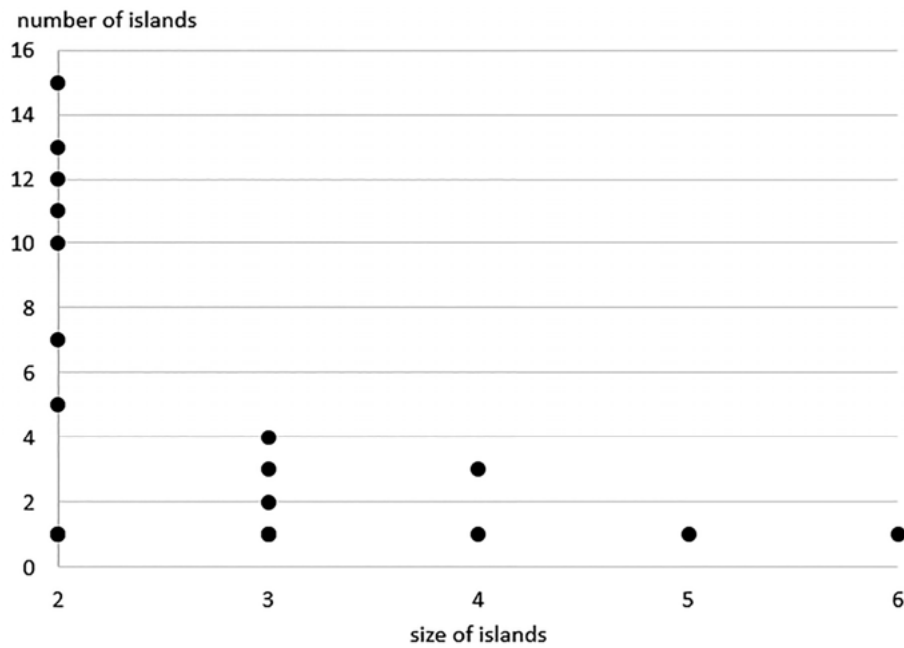


Figure 15. The relationship between numbers and sizes of islands.

	W=1	W=2	W=3	W=4	W=5	W=6	W=7	W=8	W=9	W=10
2009	14									
2010	53	5	1							
2011	43	5	1							
2012	244	4	1	1	1					
2013	911	32	1	1	3					
2014	3330	100	3	3	1					
2015	7651	199	25	13	10	3	1	3		1
2016	41056	801	27	5	3	5				
2017	63176	1470	70	7	5	4	2	1		
2018	36552	887	37	4	2					

Table 2. Numbers of links associated with different weights.

Comparison	Correlation coefficients
W=1 vs W=2	0.9953
W=2 vs W=3	0.9632
W=3 vs W=4	0.6274
W=4 vs W=5	0.9518
W=5 vs W=6	0.6307
W=6 vs W=7	0.6125
W=7 vs W=8	0.6475

Table 3. Correlation coefficients between consecutive weight links.

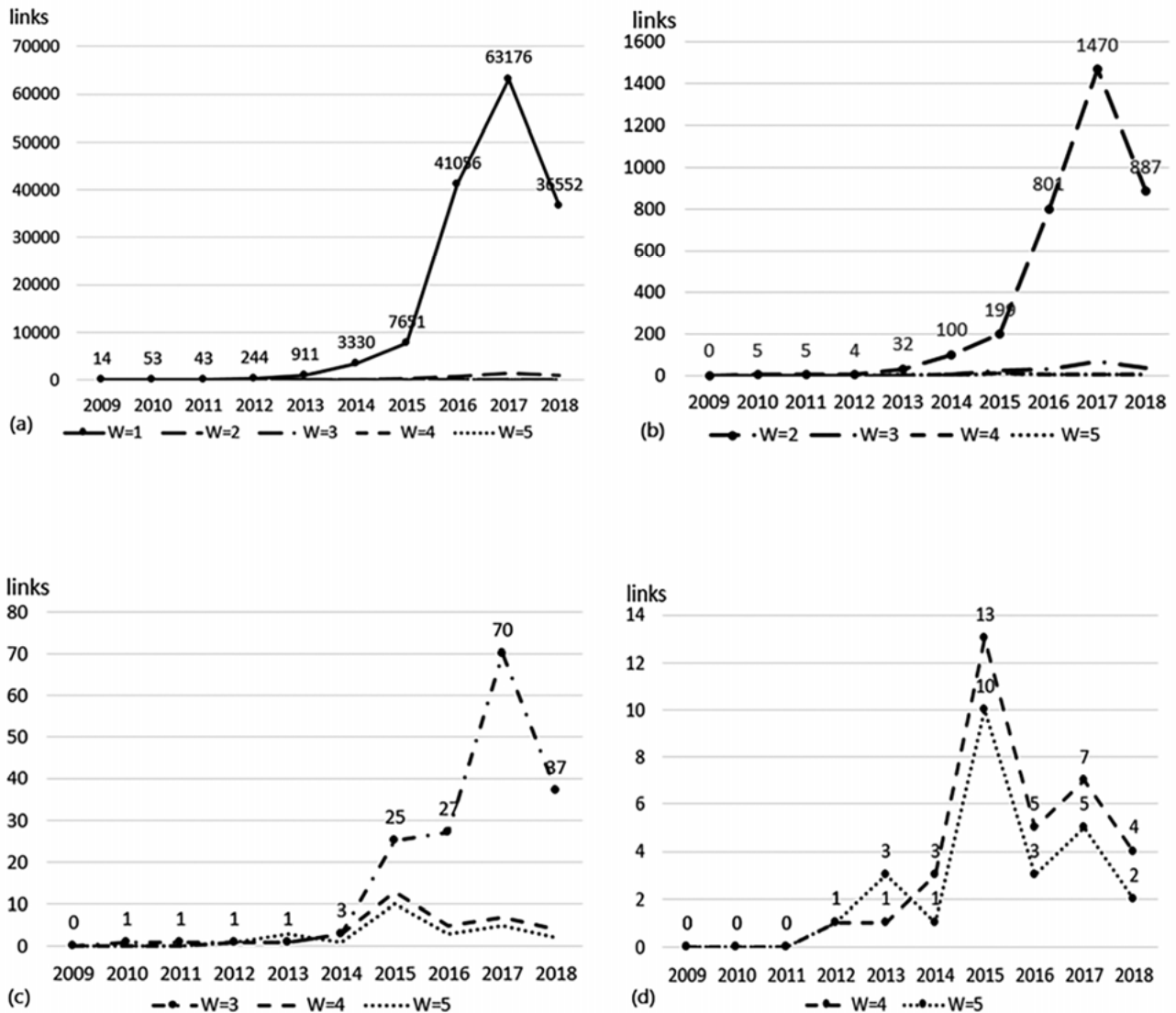


Figure 16. Numbers of links with weights 1-5 (a), 2-5 (b), 3-5(c), 4-5 (d).

is less than six. It is noted that, in this research, beyond the x-axis of Figure 18, the number of all the articles of connectivity more than twenty is one.

A rock is composed of articles sharing common keywords. These articles link to one another and form a complete graph. Different keywords will form different rocks. In other words, the number of rocks is equal to that of keywords shared by different articles. The size of a rock is measured by the number of corresponding articles. Therefore, the thirteen hottest keywords illustrated in Table 1 are the glue of the thirteen largest rocks. However, the largest rock is usually not the article with the highest connectivity. Figure 19 compares the largest rock with the highest connectivity, also shown in Figure 17. The difference between them is attributed to the other keywords that co-exist in the rock. Figure 20 shows the coverage of the largest rock

in the island where it is located. It seems that the range of the coverage may be kept between 10% and 20% in the future. Figure 19 shows that the largest rocks are not in the articles with the largest degrees. The coverages of the biggest rocks in islands where they are located converge in the 10% to 20% range (Figure 20).

4.7 Summary of findings

“Smart city” is a buzzword in recent years, and the academic community is no exception. According to Clarivate Analytics’ Web of Science database, the term “smart city” first appeared in 1985. However, only a few articles concerning smart cities existed until around 2009. After that, the number of articles has been dramatically increasing. By taking article keywords as pivotal tags to explore oceans of

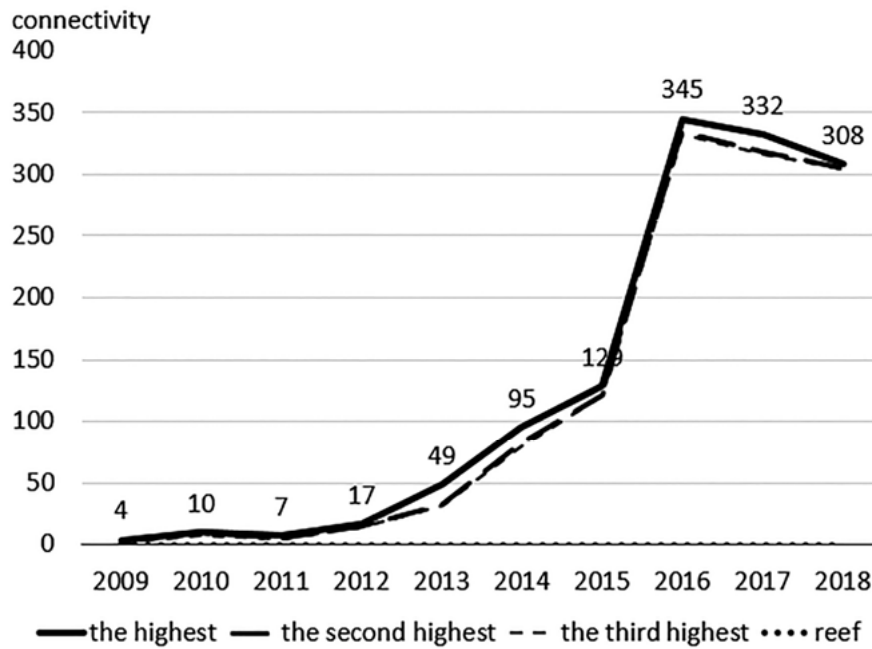


Figure 17. The yearly evolution of the top three highest and reef's connectivity.

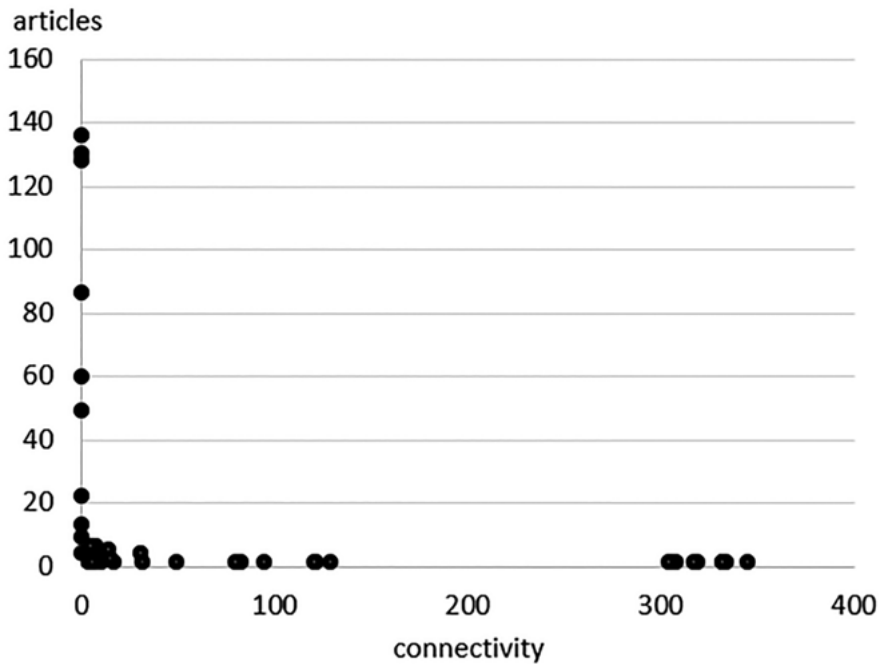


Figure 18. The relationship between the connectivity and number of articles.

academic knowledge concerning smart cities from the year 2009 to 2018, we find that the articles concerning smart cities indeed have enjoyed a booming period in the last decade, except a small drop in 2018. In other words, it is hard to judge whether the study of smart cities has matured and will decline from now on, or whether it is just a little turbulence and will keep on growing in the future.

The characteristics of hot keywords in the first and second half of the decade are quite different. They were chosen if they occurred more than two and five times in the first and second half of the decade respectively. In the first half of the decade, the percentage of the number of hot keywords against that of all the keywords increased, but it stabilized to a range of between 1% and 2%. Furthermore,

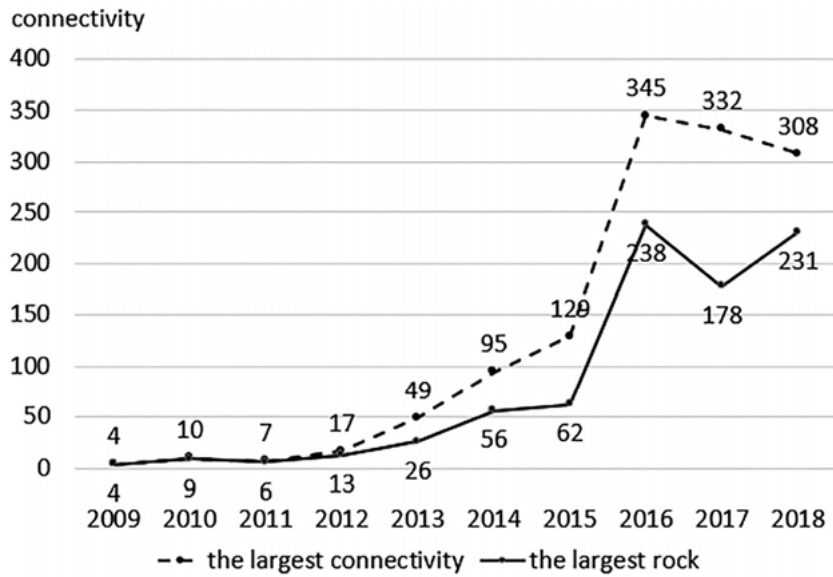


Figure 19. The comparison of connectivity between the largest rock and article.

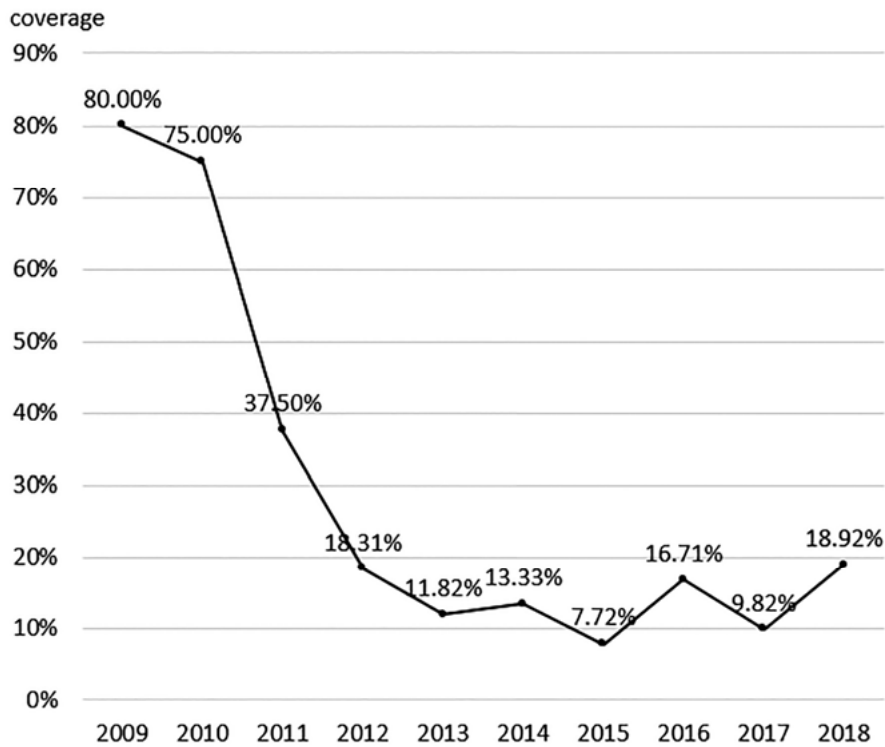


Figure 20. The coverages of the largest rocks in islands.

thirteen major hot keywords, namely IoT, big data, cloud computing, sustainability, smart grid, ICT, urban development, smart growth, GIS, tourism, ubiquitous computing, smart planet, and u-city, were selected. The numbers and percentages of keyword IoT, which was the hottest keyword in the last six years, far exceeded those of the other hot keywords.

A knowledge island contains many knowledge rocks. Since there are many keywords in an article, it can indirectly connect to other articles through different keywords, while articles sharing a common keyword are directly interconnected as a complete graph. In other words, a knowledge island is composed of loosely connected articles, while strongly connected articles form a rock. It was found

that over a decade the number of islands increased from three to more than fifteen, and their sizes increased from thirteen to more than 1,200. The sizes of the largest rock covered those of islands where they are located are around 10% to 20% for the articles about smart cities. Reefs are articles that share no keywords with other articles. As a result, an exceedingly large island, several much smaller islands, and a lot of reefs are present on the knowledge map of smart city knowledge.

5.0 Discussion

5.1 Comparative study

It is worth comparing our work with similar work done by Mora, Deakin and Reid (2018), as mentioned in section 2.3. Although the two studies used different data sources, timespans, association tags, grouping approaches, and hot keyword selections to explore how the concept of a smart city is intellectually structured, both have compatible and progressive findings. Both studies took English-language literature from scholarly databases as source articles, but Mora, Deakin and Reid searched eight databases, namely Google Scholar, Web of Science, IEEE Xplore, Scopus, SpringerLink, Engineering Village, ScienceDirect, and Taylor and Francis Online, from 1992 to 2012, while we focused on Web of Science from 2009 to 2018. Both studies involved selecting articles in which the term smart city is included in the title, abstract, keyword list, while Mora, Deakin and Reid also searched the body of the text. As a result, Mora, Deakin and Reid had 2,273 source articles, and we had 7,647 in the last decade and 6,967 in the second half. The observation that research concerning smart cities has been dramatically increasing is supported by the two independent studies where the number of articles considered in the present study from one database in the last five years is much larger than that by Mora, Deakin and Reid from various databases in twenty years. To group articles, the study authors took different tags; where Mora, Deakin and Reid used a subject-oriented co-citation approach, we used a frequency-oriented co-keyword approach. Although they used different approaches, the numbers of islands (clusters) are very close, where Mora, Deakin and Reid got eighteen clusters, we found there were seventeen or eighteen islands in the last four years.

Furthermore, both studies used different criteria to select distinct or hot keywords but still achieved some agreement and implied the trend of evolution. Mora, Deakin and Reid selected the top ten keywords in eighteen clusters and made a profile of thirty-one distinct keywords, while we selected the top three keywords from each year in the last decade and came up with a list of thirteen hot keywords. There are four keywords, namely IoT, ICT, smart

grid, and urban development, shown in both studies. On the other hand, hot keywords of 2018, namely big data, cloud computing, and sustainability, indicate the new trend of research interests. Furthermore, both studies agree that technology-oriented articles are overwhelming in the research community of the smart city.

5.2 Categorization vs. classification

It is worth mentioning that the characteristics of knowledge maps are closer to categorization than classification based on Jacob (2004), who identified classification and categorization by six systemic properties: process, boundaries, membership, criteria for assignment, typicality, and structure. Categorization processes entities using creative synthesis based on similarity and has a non-binding boundary. The criteria of category assignment can be context-dependent or context-independent; thus, the membership of an entity is flexible and can be associated with more than one category. There is no typical or representative member in a category since every member has its own different properties. The structure of a category may be flat or hierarchical. On the other hand, classification arranges entities in a systematic process based on their characteristics using predetermined assignment criteria; thus, classes are mutually-exclusive and non-overlapping, and boundaries are fixed where an entity either is or is not a member of a particular class. All members of a class are typically and equally representative. Classes can be hierarchically structured. In the case of knowledge maps, we group articles collected in WoS using a single criterion of assignment by connecting common keywords among them to form islands. If we change the relationship based on common keywords to other relationships, such as common authors, references, etc., articles will be grouped in different ways. Thus, we create a knowledge map in a categorization process. Since every article in an island (category) has different numbers of keywords associated with different frequencies, and different numbers of link strength associated with different other articles, no article can be a representative for other articles on the same island. Furthermore, the structure of a knowledge map is flat, non-hierarchical. Thus, the properties of process, criteria for assignment, typicality, and structure coincide with those of categorization. However, the other two properties, namely, membership and boundaries, behave like classification. When the criterion of assignment is given, and an article is once connected to an island, it will not change to any other island. Any two separated islands have no common keyword. In other words, any article either only belongs to an island or becomes a reef by itself. The boundaries of knowledge islands are fixed, and the islands are mutually exclusive.

5.3 Standardization of keywords

Keywords which were freely provided by the authors of articles need to be pre-processed in order to have a standardized analysis. A concept may be expressed in various terms or forms. For examples, “Internet of Things” may also be expressed as IoT, Internet-of-Things, Internet-of-Things (IoT), IoT(iot), etc. A reference table has to be built for integrating many synonyms into one. However, it is not necessary to define a limited set of control words so that innovative keywords are possible.

5.4 Future investigation

Our future work has two parts. While possible research directions for the academic community at large will be suggested in section 6.0, in this section we will discuss the future investigation of the construction and exploration of knowledge maps to explore the texture of a certain domain of knowledge, which can proceed based on the experience gained in this study. Some proposed approaches are as follows:

- measuring the distance between any two keywords: distance can be measured by the number of articles in the shortest path between two keywords; it would be very interesting to find how many years it takes for two keywords to become closer.
- measuring the density of an island: density can be measured by the ratio of the total degree of articles and/or keywords against that of a complete graph; it would be interesting to find the relationship between the evolution of densities and the cohesion of a community of interest.
- identifying patterns of life cycles of keywords: in this research, we find that some keywords in early years might be shrinking, disappearing or reviving, while others might suddenly appear in a great amount and increase dramatically; these phenomena might be affected by technology breakthroughs or socioeconomic issues.
- calculating the entropy of the distribution of islands: the entropy can be a measurement of the vitality of a community of interest a high entropy might imply a vital community in which there are many reefs or small islands with independent and creative ideas; on the other hand, a low entropy might imply that the community has focused on a set of specific topics.
- exploring fractal-like structure: we have revealed in the present study that there is a fractal-like structure embedded in the strengths of composite links in terms of their weights; this phenomenon deserves further exploration.

6.0 Conclusion

A Knowledge map is a powerful tool to capture the whole picture of a certain knowledge domain. However, one may get various pictures if different sources, timespans, association tags, grouping algorithms, and categorization processes are employed. In this article, we have explained how we explored and what we found. A comparison between this research and a similar but independent study was made, and the comparison shows that the collective findings are compatible and progressive.

The evolution shown by the knowledge maps not only illustrates the current situation of the academic community, but also indicates possible future research directions for the academic community interested in the field of smart cities. The results of this research imply that the academic community may have reached a common consensus about the issue of IoT recently. It may also signify the maturation of the topic of smart cities. Additionally, keywords concerning the value of smart cities for pursuing a better life and environment are overwhelmed by those concerning technology. Since the issues of smart cities have many facets, it is suggested that issues concerning values of smart technology, such as sustainability of urban development, social equity and justice, economic growth, adaption of climate change, etc., should be further explored in future research. Furthermore, although many scholarly databases collect published journal and conference papers, unpublished reports, and grey literature, many of them do not provide or only provide limited metadata for further academic research. If they can be downloaded more easily and made user-friendly, knowledge maps of different sources, viewpoints, tags and disciplines can be drawn more quickly. It would be very beneficial to accumulate holistic knowledge.

References

- Ahmed, Yunis Ali, Mohammad Nazir Ahmad, Norasnita Ahmad and Nor Hidayati Zakaria. 2019. “Social Media for Knowledge-sharing: A Systematic Literature Review.” *Telematics and Informatics* 37: 72-112.
- Allahyari, N., Mark S. Fox and Michael Gruninger. 2014. “City Knowledge Patterns: A Standard for Smart City Knowledge Management.” Paper presented at Semantic Cities: Beyond Open Data to Models, Standards and Reasoning Workshop at AAAI14, Quebec City July 28, 2014. <https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/viewFile/8831/8267>
- Balaida, Ali, Mohd Zaidi Abd Rozana, Syed Norris Hikmi and Jamshed Memon. 2016. “Knowledge Maps: A Systematic Literature Review and Directions for Future Research.” *International Journal of Information Management* 36: 451-65.

- Bateman, Scott, Carl Gutwin and Miguel Nacenta. 2008. "Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections." In *Proceedings of the 19th ACM Conference Hypertext and Hypermedia 16-21 June 2008 Pittsburgh, PA, USA*, ed. Brusilovsky and Hugh Davis. New York: ACM, 193-202.
- Biloslavo, Roberto and Max Zornada. 2004. "Development of a Knowledge Management Framework within the Systems Context." In *The Fifth European Conference on Organizational Knowledge, Learning and Capabilities 2-3 April 2004 University of Innsbruck, Austria*. Coventry, UK: University of Warwick. https://warwick.ac.uk/fac/soc/wbs/conf/olkc/archive/oklc5/papers/h-3_biloslavo.pdf
- Boyes, Bruce. 2016. "Smart Cities and Knowledge Management." *RealKM Magazine*. <https://realkm.com/2016/07/22/smart-cities-and-knowledge-management/>
- Brachos, Dimitris, Konstantinos Kostopoulos, Klas Eric Soderquist and Gregory Prastacos. 2007 "Knowledge Effectiveness, Social Context and Innovation." *Journal of Knowledge Management* 11, no. 5: 31-44.
- Chemchem, Amine and Habiba Drias. 2015. "From Data Mining to Knowledge Mining: Application to Intelligent Agents." *Expert Systems with Applications* 42: 1436-45.
- Chen, Chaomei. 2017. "Science Mapping: A Systematic Review of the Literature." *Journal of Data and Information Science* 2, no.2: 1-40. doi:10.1515/jdis-2017-0006
- Cheng, Ying, Ken Chen, Hemeng Sun, Yongping Zhang and Fei Tao. 2018. "Data and Knowledge Mining with Big Data towards Smart Production." *Journal on Industrial Information Integration* 9: 1-13.
- Eppler, Martin J. 2013. "What Is an Effective Knowledge Visualization? Insights from a Review of Seminal Concepts." In *Knowledge Visualization Currents*, ed. F. T. Marchese and E. Banissi. London: Springer, 3-12.
- Gambette, Philippe and Jean Véronis. 2010. "Visualizing a Text with a Tree Cloud." In *Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V. March 2009 Dresden, Germany*, ed. Hermann Locarek and Claus Weihs. Berlin: Springer, 561-9.
- Hao, Karen. 2019. "We Analyzed 16,625 Papers to Figure out Where AI is Headed Next." *MIT Technology Review* (blog), January 25. <https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/>
- Heimerl, Florian, Steffen Lohmann, Simon Lange and Thomas Ertl. 2014. "Word Cloud Explorer: Text Analytics Based on Word Cloud." In *Proceedings of the 47th Hawaii International Conference on System Sciences January 2014 Hawaii, USA*. Washington, DC: IEEE Computer Society, 1833-42.
- Hjørland, Birger. 2008. "What is Knowledge Organization (KO)?" *Knowledge Organization* 35: 86-101.
- Jacob, Elin. 2004. "Classification and Categorization: A Difference that Makes a Difference." *Library Trends* 52, no. 3: 515-40.
- Jennex, Murray E. and Iryna Zakharova. 2006. "Culture, Context, and Knowledge Management." *International Journal of Knowledge Management* 2: i-iv.
- Liu, Jun, Zhinan Zhang, Richard Evans and Youbai Xie. 2019. "Web Services-based Knowledge Sharing, Reuse and Integration in the Design Evaluation of Mechanical Systems." *Robotics and Computer Integrated Manufacturing* 57: 271-81.
- Liu, Lu, Jing Li and Chenggong Lv. 2009. "A Method for Enterprise Knowledge Map Construction based on Social Classification." *Systems Research and Behavioral Science* 26: 143-53.
- McInerney, Claire. 2002. "Knowledge Management and the Dynamic Nature of Knowledge." *Journal of the American Society for Information Science and Technology* 53: 1009-18.
- Medelyan, Alyona. 2018. "5 Text Analytics Approached - A Comprehensive Review." Thematic October 02, 2018. <https://getthematic.com/insights/5-text-analytics-approaches/>
- Meijer, Albert and Manuel Pedro Rodríguez Bolívar. 2016. "Governing the Smart City: A Review of the Literature on Smart Urban Governance." *International Review of Administrative Sciences* 82: 392-408.
- Mora, Luca, Mark Deakin and Alasdair Reid. 2018. "Combining Co-citation Clustering and Text-based Analysis to Reveal the Main Development Paths of Smart Cities." *Technological Forecasting & Social Change* 142: 56-69.
- Ong, Thian-Huat, Hsinchun Chen, Wai-ki Sung and Bin Zhu. 2005. "Newsmap: A Knowledge Map for Online News." *Decision Support Systems* 39: 583-97.
- Rivadeneira, Walkyria Goode, Daniel M. Gruen, Michael Muller and David Millen. 2007. "Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems April 28-May 03, 2007 San Jose, CA, USA*. New York: ACM, 995-8. doi:10.1145/1240624.1240775
- Scharnhorst, Andrea, Richard P. Smiraglia, Christophe Guéret and Alkim Almila Akdag Salah. 2016. "Knowledge Maps of the UDC: Uses and Use Cases." *Knowledge Organization* 43: 641-54. doi:10.5771/0943-7444-2016-8-641
- Su, Hai and Zuhua Jiang. 2007. "Construction Method of Knowledge Map based on Design Process." *Chinese Journal of Mechanical Engineering* 20: 98-104.

The Information Retrieval Thesaurus*

Stella G. Dextre Clarke

Luke House, West Hendred, Wantage, OX12 8RR, UK,
<stella@lukehouse.org>

Thesauri have been a recurrent theme in Stella Dextre Clarke's career spanning forty-five years, mostly spent on information system development and database publishing. Graduating originally in chemistry, she earned a master's degree in information science before serving in a variety of information centres, where duties typically involved vocabulary design and maintenance, then bibliographic database management and subsequently consultancy. Her work on standards and on vocabulary development was recognized in 2006 when she won the Tony Kent Strix Award for outstanding achievement in information retrieval. Now retired, she is an honorary member of ISKO and the Vice-chair of ISKO UK.



Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus." *Knowledge Organization* 46(6): 439-459. 99 references. DOI:10.5771/0943-7444-2019-6-439.

Abstract: In the post-war period before computers were readily available, urgent demand for scientific and industrial development stimulated research and development (R&D) that led to the birth of the information retrieval thesaurus. This article traces the early history, speciation and progressive improvement of the thesaurus to reach the state now conveyed by guidelines in international and national standards. Despite doubts about the effectiveness of the thesaurus throughout this period, and notwithstanding the dominance of Google and other search engines in the information retrieval (IR) scene today, the thesaurus still plays a complementary part in the organization of knowledge and information resources. Success today depends on interoperability, and is opening up opportunities in linked data applications. At the same time, the IR demand from workers in the knowledge society drives interest in hybrid forms of knowledge organization system (KOS) that may pool the genes of thesauri with those of ontologies and classification schemes.

Received: 2 April 2019; Accepted: 10 April 2019

Keywords: thesaurus, thesauri, ISO 25964, terms, information retrieval

* Derived from the article titled "Thesaurus (For Information Retrieval)" in the ISKO Encyclopedia of Knowledge Organization, Version 1.2 published 2019-03-07. Article category: KOS Kinds.

1.0 Introduction and clarification of scope

This article is about thesauri intended for use in information retrieval (IR—see note 1), rather than literary thesauri, which are generally designed for the different purpose of helping and inspiring the choice of words and phrases in normal discourse. *Roget's Thesaurus*, that very well-known literary thesaurus first published in 1852, long pre-dates the first IR thesaurus and probably inspired the invention of the latter. For this reason, there is some reference to literary thesauri in the history section of this article. In other sections, however, the term "thesaurus" invariably refers to the information retrieval thesaurus.

2.0 What is a thesaurus?

2.1 Purpose

The prime function of a thesaurus is to support information retrieval by guiding the choice of terms for indexing and searching. According to ISO 25964-1 (International Organization for Standardization 2011, Clause 4.1):

The traditional aim of a thesaurus is to guide the indexer and the searcher to choose the same term for the same concept ... a thesaurus should first list all the concepts that might be useful for retrieval purposes in a given domain. The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term ... Secondly, a thesaurus should present the preferred terms in such a way that people will easily identify the one(s) they need. This is achieved by establishing relationships between terms—and/or between concepts—and using the relationships to present the terms in a structured display.

Foskett (1980) lists seven purposes for a thesaurus, of which six could be considered subdivisions or sub-aspects of the main purpose cited above (As for his seventh purpose, a means of standardizing the use of terms in a given subject field, Foskett acknowledges that this is desirable rather than realistic). While the ISO 25964 description dates from 2011, it follows principles established long before. For example, Lancaster (1972, 25) explains:

Schultz (1967) has distinguished the functions of the information retrieval thesaurus from a thesaurus of the Roget type as follows. Roget's purpose was to give an author a choice of alternative words to express one concept; to display a set of words of similar meanings to allow an author to choose one that best suits his need. The information retrieval thesaurus tends to be more prescriptive. The thesaurus compiler chooses one term from among several possible, and directs the user to employ this one by means of references from synonyms and other alternative forms.

The use of preferred terms rather than language-independent codes or character strings is a key feature distinguishing the thesaurus from the classification schemes that were commonly used for IR before the advent of the thesaurus. Retrieval may seem simpler, to the layman, if it can be expressed in words rather than codes. But there is an ambiguity challenge to overcome—in the language of normal discourse one concept can be expressed in many different ways, and conversely one term can have many different meanings. To achieve the aim of always choosing the same term for the same concept, an artificial indexing language has to be established, in which synonyms are controlled, homographs are disentangled, and each preferred term is allowed only one meaning (although some may have very broad meanings). The thesaurus conveys that artificial language.

This *modus operandi* for the thesaurus became established in the 1960s. The computer was then in its infancy: small, primitive, and almost entirely unavailable to the communities of researchers and practitioners needing to retrieve information. Without computers to help, trained human intermediaries were needed at two critical stages of the best IR systems: to index and/or classify the source documents and (in the second stage) to perform searches of the same.

Several classic texts of this period, such as Gilchrist (1971), Lancaster (1972), Aitchison and Gilchrist (1972), and Soergel (1974), make it clear that the thesaurus is just one component in the whole IR system comprising a set of tools and procedures, all of which have to be designed in harmony. In those days, the IR system typically operated in isolation. While modern technology enables many more possibilities, needing even greater attention to compatibility among system components, today's standards still respect and support the original design principles.

2.2 Content and structure

The components of a thesaurus are most succinctly laid out in the UML (Unified Modelling Language) model shown in ISO 25964-1 and reproduced as Figure 1 below

(The model may also be seen on the official website at www.niso.org/schemas/iso25964/, and downloaded from http://www.niso.org/schemas/iso25964/Model_2011-06-02.jpg. Key features are explained in Will (2012)).

Thus, the essential core of a thesaurus is a collection of concepts represented by terms and interlinked by relationships, of which the three main types are equivalence (between terms), hierarchical (between concepts) and associative (also between concepts). By long established convention, the tags USE and UF (Use For) precede preferred and non-preferred terms respectively, and the equally characteristic tags BT, NT and RT indicate broader, narrower and associatively related concepts respectively. Figure 2 illustrates how these simple elements are traditionally displayed. A great many in-house thesauri are built in this minimal way, without calling upon the many optional extras provided for in the data model. The alphabetical display in Figure 2 may optionally be supplemented by other types of presentation, as discussed in Section 4.3 below on “systematization.”

The thesaurus can also be visualized as a complex web of interlinked concepts in which each concept is labelled by one or more terms in one or more languages. It has these main features:

- The semantic scope of a concept is indicated partly by the totality of terms labelling it, partly by the hierarchical relationships linking it to broader and/or narrower concepts, and where this is not enough, by a scope note and sometimes term definitions.
- Admissible hierarchical relationships are of three types: generic, instancial or partitive (subject to some restrictions on the eligible types of partitive link). It is optionally possible to distinguish these types, using the tags BTG/NTG, BTI/NTI, BTP/NTP respectively.
- Admissible associative relationships apply to non-hierarchical situations wherever two concepts are so associated that an indexer or a searcher should consider using one of them as well as, or instead of, the other.
- Concepts may be presented and ordered in arrays with node labels, following the principles of facet analysis
- Concepts may also be grouped in loose structures to suit particular domains or applications
- Concepts not explicitly included in the thesaurus may be represented by combinations of preferred terms (in situations known as “compound equivalence”—for example: “coal mining USE coal + mining”)
- It is also possible to add metadata to terms, to concepts, to relationships and to the thesaurus as a whole, such as dates of introduction or change, version history, house-keeping data, copyright information, etc.

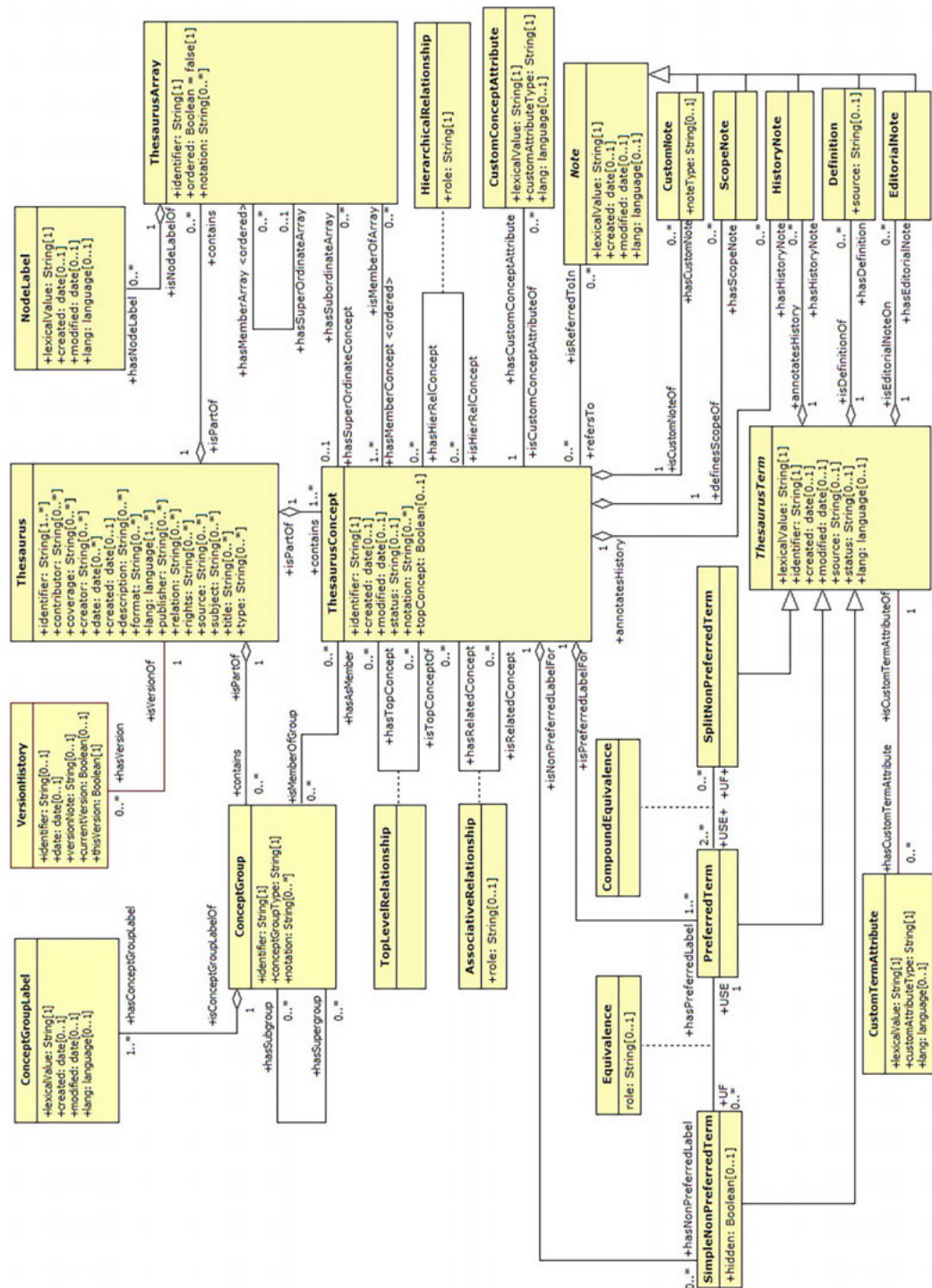


Figure 1. Thesaurus data model as recommended in ISO 25964-1.

pesticides		plant products	
UF:	<i>fumigants</i>	NT:	cereals
BT:	agrochemicals		fruits
NT:	fungicides		spices
	herbicides		vegetables
	insecticides	RT:	plants
RT:	pests		
pests		plants	
NT:	pest insects	RT:	plant products
	plant pests		
RT:	pesticides	<i>porkers</i>	
		USE:	pigs
pigs		poultry	
UF:	<i>bogs</i>	BT:	livestock
	<i>porkers</i>	NT:	chickens
	<i>sons</i>		ducks
BT:	livestock		geese
			turkeys
plant pests		RT:	eggs
BT:	pests		
		sheep	
		BT:	livestock
		RT:	wool

Figure 2. Brief extract from a minimal monolingual thesaurus.

It should be stressed that many of these features are optional, enabling a variety of sophisticated uses, and should not deter straightforward use of the basics in simple applications. Detailed advice on all of them may be found in national and international standards—principally ISO 25964 and ANSI/NISO Z39.19 (National Information Standards Organization [2005] 2010)—and are further explicated in Aitchison et al (2000), Broughton (2006a) and Will 2012.

Despite availability of the guidance cited above, few current or past thesauri comply with the standards in every detail. Difficulties and divergences commonly occur in the following aspects:

- Rigorous conformity with guidelines for hierarchical relationships
- Rigorous facet analysis
- In a multilingual thesaurus, when and how to establish equivalence across languages
- When and how to admit complex concepts designated by compound terms
- Adoption of the data model

Surmounting such difficulties demands considerable expertise and time, adding to the expense of thesaurus construction and to doubts about cost-effectiveness, as noted in Sections 4.4 and 4.6 below.

2.3 Definitions

An authoritative definition of “thesaurus” may be found in the international standard ISO 25964-1 (Clause 2.62):

controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.

Although phrased differently, a broadly compatible definition is that in the American standard ANSI/NISO Z39.19-2005 (R2010) (National Information Standards Organization [2005] 2010, Clause 4.1):

A controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators. Relationship indicators should be employed reciprocally.

The above definitions derive their authority from the process of drafting and approving a standard, which requires agreement by a committee of experts and extensive consultation among the user community. But copious alternative definitions exist in a variety of texts, illustrating the

extent of confusion that surrounds the thesaurus. Many are intended to counteract loose use of the term “thesaurus,” which is commonly applied to any sort of knowledge organization system (KOS), such as a subject headings list, or to a set of synonym rings. Conversely, some vocabularies that could properly be described as thesauri may instead be called an ontology or a taxonomy.

2.4 Why the confusion?

Some of the current confusion may be explained by developments that emulate thesaural conventions in other types of KOS. In the 1990s, for example, a “thesaurification” project explored adaptation of some schedules of the UDC (Universal Decimal Classification) (Riesthuis and Bliedung 1991). This did not interfere with the primary function of the UDC as a classification scheme. Around the same time, the *Library of Congress Subject Headings (LCSH)* began to adopt thesaurus tags such as BT, NT and RT in its display (Thesaural use of these tags is illustrated in Figure 2). Today (3 August 2016), a Wikipedia entry for the *LCSH* claims that “The Library of Congress Subject Headings (LCSH) comprise a thesaurus ... of subject headings, maintained by the United States Library of Congress, for use in bibliographic records.” The use of thesaural conventions and BT/NT tags, however, does not make the *LCSH* a thesaurus, as pointed out long ago by Rolland-Thomas (1993) and illustrated more recently by Spero (2008 and 2012). The *LCSH* is fundamentally a subject headings list (defined as a “structured vocabulary comprising terms available for subject indexing, plus rules for combining them into pre-coordinated strings of terms where necessary” – ISO 25964, clause 2.57) rather than a thesaurus. Differences between the way subject headings and thesauri are used are discussed in De Keyser (2012).

Another explanation is that all the national and international standards for thesauri take the form of guidelines rather than mandatory instructions. Adopters, therefore, have a great deal of liberty to cherry-pick only the recommendations that suit the circumstances of their own thesaurus and ignore the rest.

A third part of the explanation is that very often the person charged with sorting out an organization’s information assets has little or no training in knowledge organization. If the decision is to develop an in-house indexing language or a filing structure, it may be built in whatever way comes easiest, and randomly named a “classification scheme” or a “thesaurus” or an “ontology” to suit the fashion of the day. The misnomer “thesaurus” has spread easily this way, leading to much confusion.

Even for the trained information professional, distinguishing between the different types of KOS can be hard. Over the years, many attempts at clarification have been

made (e.g. Fast et al, 2002; Garshol 2004; Hodge 2000; Kless et al 2012). Useful definitions of several KOS types may be found in ANSI/NISO Z39.19 and ISO 25964; ISO 25964-2 (International Organization for Standardization 2013) also brings out the similarities and the differences to provide for in the context of interoperability. Zeng (2008) casts further light by analysing and comparing features of many different types of KOS.

A different sort of confusion surrounds the basic roles of terms versus concepts. From the early days of thesaurus R&D, the basic aim was to index the semantic content of documents rather than the terminological content. Concepts useful for indexing were collected in a thesaurus, where they were organized and their inter-relationships were established. When a hierarchical relationship was established, the reciprocal links between the broader and narrower concepts might usefully have been designated BC (broader concept) and NC (narrower concept). In practice, however, they were named BT (broader term) and NT (narrower term), and this practice was adopted widely. The 1974 edition of ISO 2788 (International Organization for Standardization 1974) attempted to clarify by explaining “the hierarchical relation is represented by the references BROADER TERM (BT), representing the relation of a concept being superordinated, and NARROWER TERM (NT), indicating the reciprocal relation” (International Organization for Standardization (1974) clause 3.4.3). But it was too late—the misnomers have stuck, to this day.

Over the decades, this confusion has led to much misunderstanding among thesaurus users. Also, the software developed for thesaurus management has often adopted a data model in which the hierarchical and associative relationships are established between terms rather than between concepts, and this has impeded thesaurus interoperability. See discussion in Dextre Clarke and Zeng (2011).

3.0 How a thesaurus is used

3.1 For post-coordinate indexing and searching

The original thesaurus purpose and mode of use as declared in the standards is confirmed in many texts, such as Wellisch (1995, 475) “thesauri are primarily intended for indexing as well as for searching and retrieval from post-coordinated systems, in which an indexer may assign several descriptors to documents, while users may combine those descriptors to form search statements.” For more background on post-coordinated systems and their underpinning with controlled vocabularies and Boolean logic, see Sharp (1967), Lancaster (1972) or Dextre Clarke (2008). Today’s continuing demand for quality bibliographic databases supporting Boolean retrieval is upheld by Hjørland (2015).

As originally conceived, the act of consulting the thesaurus either for indexing or for searching can be time-consuming. When the user has worked out key concepts of the document to be indexed or the query to be investigated, he or she needs to find an entry point among the terms and/or groups available, and follow the network of relationships to establish the closest possible match in the thesaurus. Skill as well as patience and subject knowledge are needed, since thesauri vary greatly in quality and in format (see Section 5 below). Nowadays it is hard to find indexers with the requisite training, while trained end-users are very rare indeed. Therefore, modern systems tend to automate both indexing and searching, using an electronic version of the thesaurus.

Thesaurus-based indexing functions may be needed in situations such as library cataloguing, compilation of bibliographic databases and tagging/indexing of image collections. Generally, a software package designed for that application is used, with indexing support capabilities that vary from (at the simple end) speeding up the task of thesaurus navigation to (at the sophisticated end) delivering totally automatic indexing. In between are systems that validate and/or switch the indexer's terms, that select candidate terms for the human indexer to accept or reject and "suggester" systems for social tagging.

Forty years ago Caplan (1978) reported a number of failures in trials of thesaurus-based automatic indexing. More recently Lancaster (1998) provided a more promising account of the techniques available, but concluded (294) "even the most sophisticated of current automatic indexing procedures compare unfavourably with skilled human indexing." Eight years later, Tudhope et al (2006) were still calling for more research. Ten years later, research into metadata enrichment with thesaurus terms was outlined in Tudhope and Binding (2016). Kempf and Neubert (2016) describe several modes of implementation, including one that exploits inter-KOS mappings. Clause 16 of ISO 25964-1 advises on the thesaurus features needed to enable such functions.

Meanwhile, as described in Section 7 below, a new breed of KOS is emerging in the enterprise search sector, loosely named "taxonomy," and stimulating a demand for automatic categorization tools. Since some taxonomies share some features with thesauri (See ISO 25964-2, Clause 19), the associated R&D effort is already yielding progress that can be applied to thesaurus-based indexing. Unfortunately, a great many in-house applications go unreported in the research literature, including research by the vendors of software for automatic categorization. In the experience of this author, the support for thesaurus-based indexing in off-the-shelf library management packages is rarely as effective or user-friendly as it could be. Likewise, the quality of automatic or semi-automated indexing tools varies

greatly and much care is needed to obtain reliable outputs. Further discussion of automatic indexing is outside the scope of this article, particularly since most cases do not use a thesaurus.

Turning now to search applications, here too the electronic medium speeds up thesaurus navigation. Furthermore, with suitable software it enables broadening or narrowing a search at will. Consider, for example, a search for "packaging AND fruit." Relevant results would include any items dealing with the packaging of any of potentially hundreds of different types of fruit. The technique known as "search explosion" exploits the hierarchical relationships in a thesaurus to expand the search statement automatically and cover all those hundreds of fruit types. It is similarly possible to extend a search via associative relationships, and this is usually termed "search expansion." These and other search functions are reviewed in Shiri et al (2002), and further discussed in Shiri (2012). The case study of the *STW Thesaurus for Economics* by Kempf and Neubert (2016) illustrates similar techniques, and other ways in which a thesaurus can be used to enhance retrieval, even when the user is unaware of its support.

Evidently indexing and searching have moved on from the early days, when a thesaurus and its IR system could operate usefully in isolation and even without a computer. Thesaurus use in today's IR applications relies on electronic manipulation, involving transfer of data from one subsystem to another. Success depends on interoperability, i.e., the ability of systems and/or components to exchange information and to use the information that has been exchanged. There are now at least two main contexts for thesaurus interoperability:

- "Vertical integration" of the thesaurus with software for indexing or searching or occasionally some other IR function, as already described;
- "Horizontal engagement" of the thesaurus with another KOS (perhaps another thesaurus, or a subject headings list, or a classification scheme), typically requiring conversion of indexing and search expressions between the languages of the different KOSs.

The vertical context sees a thesaurus transformed from a static map of concepts, terms and relationships to a functioning system. The horizontal context crosses a different boundary, to be described next.

3.2 Networked uses, especially in the semantic web

A single search across multiple databases would be relatively straightforward if all used the same natural language, the same machine protocols and the same indexing language. To overcome the disparities found in real life, two

approaches to interoperability are especially relevant for KOSs, namely inter-vocabulary mappings and linked data.

A mapping is defined as a “relationship between a concept in one vocabulary and one or more concepts in another” (ISO 25964-2, clause 3.41). For example, an equivalence mapping between the concepts labelled “instant coffee” in one thesaurus and “soluble coffee” in another, would establish that they are viewed as identical for semantic purposes. Existence of mappings like this makes it easy to “translate” search queries for use in the corresponding IR systems, and/or to augment the metadata of resources indexed with either thesaurus. When sets of mappings are available between many KOSs, it opens the prospect of extending searches widely and multilingually.

The value of such mappings is demonstrated in the “Metathesaurus” of the Unified Medical Language System (UMLS) <www.nlm.nih.gov/research/umls/>, a semantic tool serving research in biomedicine, health care and related fields. It contains concepts from more than 100 KOSs as well as relationships from within the KOSs and many mappings between their respective concepts. Andrade and Lopes Gines de Lara (2016) assess its usefulness in retrieval from relevant databases. The influence of this construct has led some authors to speak of a “metathesaurus” wherever existing thesauri are integrated, linked or mapped together (Shiri 2012)—and a variety of ways is possible.

Not all mappings are as simple as equivalence. Dextre Clarke (2011a) enumerates a variety of mapping types in-

vestigated in research projects such as Renardus, MACS (Multilingual ACcess to Subjects), CrissCross and Ko-MoHe. ISO 25964-2 (International Organization for Standardization 2013) provides for hierarchical and associative mappings as well as equivalence. Hierarchical mappings are directional—either broader or narrower. Equivalence mappings subdivide into simple or compound; compound equivalence has two subtypes (intersecting or cumulative) while simple equivalence can be qualified as exact or inexact. Figure 3 shows the range of mapping types, with an example of each. Mapping statements should be expressed using the conventional tags (EQ, BM, NM etc) and the symbols shown.

Even more subtlety is possible in applications that need to distinguish between subtypes of hierarchical mapping. See Figure 4.

While thesaurus mapping projects have a much longer history (see, for example, Horsnell (1975) or Hood and Eberman (1990) or Hoppe (1996) reporting on UMLS work that began in 1986), the growth of the Internet and the WWW has made them more widely applicable. Thus, Zeng and Chan (2004) drew attention to opportunities emerging in the internet context and Vizine-Goetz et al (2004) described a labour-saving methodology. Mayr and Petras (2008) illustrated the possibilities. Several other mapping projects were reported in Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization (Boteram et al 2011).

Equivalence	
Simple:	Laptop computers EQ Notebook computers
Exact equivalence:	Aubergines =EQ Egg-plants
Inexact equivalence:	Horticulture ~EQ Gardening
Compound:	
Intersecting compound equivalence:	Women executives EQ Women + Executives
Cumulative compound equivalence:	Inland waterways EQ Rivers Canals
Hierarchical	
Broader:	Streets BM Roads
Narrower:	Roads NM Streets
Associative	e-Learning RM Distance education

Figure 3. Mapping types in ISO 25964-2, with examples of mapping statements.

Hierarchical Subtype	Mapping statement example	Reciprocal example
Generic	rats BMG rodents	rodents NMG rats
Instantial	Paris BMI capital cities	capital cities NMI Paris
Whole-part	fingers BMP hands	hands NMP fingers

Figure 4. Mapping statements that distinguish between subtypes of hierarchical mapping.

Doerr (2000) analysed perceived semantic problems of thesaurus mapping. Confusingly for us today, his use of the term “mapping” differs from the ISO 25964 definition, applying to relationships within one vocabulary rather than between different ones. Thus, he deplored the weakness of thesaurus semantics for hierarchical relationships when compared with class subsumption in an ontology. Subsequent release of SKOS (see note 2) seems to have overcome or at least eased such problems (Tudhope and Binding 2016). According to Isaac and Baker (2015, 2)

The lack of a way to express less formal semantics hindered many early projects that tried to apply Semantic Web technology in the cultural sector by massaging existing knowledge organization systems into formal ontologies. Given the scope of the artifacts considered, this effort required considerable ontological debugging that was ultimately of dubious value. Indeed, most information retrieval scenarios using KOS for searching or browsing collections do not require more than the information that one concept is broader than another.

Establishment of the world wide web (WWW) has brought new opportunities and challenges for IR in general and for KOS use in particular. On the one hand, vast resources have come within our reach; on the other hand, individual resources may be expressed in a multiplicity of languages like the Tower of Babel. As pointed out in the context of the *STW Thesaurus for Economics*, “The Web changes everything” (Kempf and Neubert 2016, 162).

A particular breakthrough for KOS linkage was approval and release of the W3C recommendation *SKOS Reference* (Miles and Bechhofer, 2009) with specific guidance on how to publish mappings between KOSs. SKOS publication followed on from a research report by Miles (2006, 1) aiming “to develop a formal theory of retrieval using controlled vocabularies that have a simple and intuitive structure [such as thesauri, classification schemes, subject heading systems, taxonomies and other types of structured vocabulary], to provide the necessary theoretical foundations for the development of Semantic Web languages and design patterns for distributed retrieval applications.” Since 2009, a number of extensions have been added to SKOS to support interoperability in particular contexts; work on some mapping tools for thesauri is described in note 2.

Turning to the other main interoperability opportunity, the principles of linked data are set out in Tim Berners-Lee’s 2006 paper at www.w3.org/DesignIssues/LinkedData.html. As he explains (1), “The Semantic Web ... is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data. Like the web of hypertext,

the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links [sic] between arbitrary things described by RDF.” For a KOS (such as a classification scheme or a thesaurus) the essential starting point is to publish the whole scheme on the web using resource description framework (RDF) syntax and giving each concept or class a uniform resource identifier (URI). Once that is in place, anyone anywhere can set up a direct link to any concept or class. For example, if a web page or a bibliographic record in a database on the web has been indexed with the thesaurus concept “renewable energy,” the person interested in that concept can move directly from the thesaurus to those and other relevant pages. This opens up the prospect for any thesaurus published on the web to act as a connecting hub for an immense literature in the subject field concerned, without any need to assemble the disparate documents in one collection or database. See note 2 for tools for hand-in-hand application of ISO 25964 and SKOS.

A vision of Wikipedia as the connecting linked data hub for hundreds of thesauri and other KOSs is outlined in a speculative paper from Garcia-Marco (2016). Kempf and Neubert (2016) show how the use of linked open data (LOD) is already paying off for the *STW Thesaurus for Economics*. Baca and Gill (2015) describe the challenges and long development path at the Getty Research Institute leading up to publication on the web of three KOSs that are very influential and widely used in the cultural heritage sector: *The Art & Architecture Thesaurus*, the *Union List of Artist Names* and the *Getty Thesaurus of Geographic Names*. They hope, thereby, to enable potential universal access to information in different formats and languages, about the works of art and countless other exhibits in museums, libraries and galleries around the world. The 2017 release by the European Commission of a new European Interoperability Framework (EIF) <ec.europa.eu/isa2/eif> further emphasizes the opportunities for public administrations to put linked data to work across the countries of Europe.

3.3 Other uses

Although not the primary purpose, thesauri may also be used for precoordinate indexing (Wellisch 1995, ISO 999:1996 (International Organization for Standardization, 1996)). When this is done, users of the precoordinate index (typically found at the back of a book) are not expected to consult a thesaurus (since cross-references to synonyms etc. may be embedded within the index). Conversely, a thesaurus may be used not for indexing but only for searching. This removes the need for compliance with the standards. See Section 5.7 below.

Educationalists sometimes argue that a thesaurus is valuable in its own right, for domain analysis, as a conceptual and terminological guide to a domain, and for development of the mind. Lykke Nielsen (2001, 778) states that “the thesaurus is a tool that helps individual users to get an understanding of the collective knowledge domain.” Broughton (<www.iskouk.org/sites/default/files/ISKOUKGreatDebate-Broughton_0.pptx>, slide 6) argues “the thesaurus teaches us to take a critical and analytical approach to the domain. It makes us think about the nature of concepts, the form of their labels [and about] their relationships,” and (slide 9) “there’s something fundamental about this approach to modelling information domains that should not be lightly abandoned.” More generally Soergel (2014) has argued that the construction of any sort of knowledge organization schema, particularly with entity-relationship modeling, facet analysis and a graphical presentation of concepts, is a useful learning discipline. Still more uses are emerging as the internet pervades the office and everyday living. To satisfy these new uses, however, the standard thesaurus model may need to evolve.

4.0 History of thesaurus development and use

4.1 Origins

To Peter Mark Roget, working in the middle of the nineteenth century, we owe the insight that it would be valuable to supply (Roget 1952, 559 emphasis original) “a collection of the words [the English language] contains and of the idiomatic combinations peculiar to it, arranged, not in alphabetical order, as they are in a dictionary, but according to the *ideas* which they express.” His aim, rather than information retrieval, was to help “find the word, or words, by which [an] idea may be most fitly and aptly expressed.”

An expanding volume of scientific and other scholarly literature in the first half of the twentieth century brought challenges for classification, the orthodox retrieval technology of the times. It led to developments such as faceted classification, post-coordinate indexing, and experiment with various sorts of cards, all of which were to prove helpful when the idea of an IR thesaurus was conceived.

According to Roberts (1984), the first suggestion of using a thesaurus in the context of IR came from Calvin Mooers in 1947. At around the same time C. L. Bernier and E. J. Crane made a similar, independent suggestion, but (Roberts, 272) “expressed the view that a general thesaurus was not an appropriate form for retrieval purposes.” Much experiment followed over the next decade, but none of the various thesaurus approaches described by Joyce and Needham (1958)—e.g., “term lattices”—seems to have prospered, nor come close to the style of thesaurus that was to emerge in 1959. It was after this ges-

tation period that (Roberts 1984, 281) “the first full-scale, operational in-house retrieval thesaurus [was produced] to solve pressing practical problems at E. I. Du Pont Nemours and Co., Inc., Wilmington, U.S.A.” Krooks and Lancaster (1993) credit Eugene Wall with developing the principles that determined the shape of this pioneering compilation.

4.2 Period of ascendancy

More research and development (R&D) followed the 1959 birth of the IR thesaurus, one of the driving forces being the post-war preoccupation of the US military with a need for effective information management. Progress came with publication of the *Thesaurus of ASTLA Descriptors* (Armed Services Technical Information Agency 1960) and of the *Chemical Engineering Thesaurus* (American Institute of Chemical Engineers 1961), followed in 1967 by the landmark *Thesaurus of Engineering and Scientific Terms* (Office of Naval Research 1967), commonly known as *TEST*. A fuller description of these works can be found in Krooks and Lancaster (1993) and in Aitchison and Dextre Clarke (2004).

Widespread use of thesauri continued throughout the 1960s, 1970s and 1980s in IR systems that mostly relied on cards of various types, sizes and materials, including some that were sorted by machines (Dextre Clarke 2008; Sharp 1967). These were post-coordinate systems, which require each document to be indexed by selecting relevant terms from a controlled vocabulary such as a thesaurus. A thesaurus was used too by many of the bibliographic databases that were hosted online by services such as Lockheed’s Dialog system, followed later by CD-ROM distribution. Notable pioneers of construction methodology included Jean Viet, Jean Aitchison and Donald Leatherdale, who each produced a number of influential thesauri. Dextre Clarke (2008) provides a vivid account of how the tools and technology of those times were used.

Further impetus came from development of national and international standards for thesaurus construction, indicating the extent of interest from the information-using community. The most influential, listed in chronological order of their first editions, included:

Deutsches Institut für Normung. DIN 1463 Guidelines for the establishment and development of monolingual thesauri [translated title] 1972 (Now withdrawn, with ISO 25964-1 recommended in its place).

International Organization for Standardization. ISO 2788-1974 Documentation - Guidelines for the establishment and development of monolingual thesauri. 1st ed. International Organization for Standardization: Geneva, 1974 (Now superseded by ISO 25964-1).

American National Standards Institute. ANSI Z39.19-1974 American National Standard Guidelines for thesaurus structure, construction and use. American National Standards Institute: New York, 1974 (Now superseded by ANSI/NISO Z39.19-2005).

International Organization for Standardization. ISO 5964-1985. Documentation - Guidelines for the establishment and development of multilingual thesauri. International Organization for Standardization: Geneva, 1985 (Now superseded by ISO 25964-1).

These and other KOS standards are discussed in Dextre Clarke (2011b), although this article pre-dates publication of the two parts of ISO 25964, in 2011 and 2013 respectively (International Organization for Standards 2011 and 2013).

4.3 Systematization

As noted by Dextre Clarke (2001, 86) “standardisation has not brought uniformity.” Having the status of guidelines rather than mandatory requirements, all the standards left plenty of scope for continuing experiment. While nearly all published thesauri include an alphabetical list of terms (which may be as simple as the extract in Figure 2, or can show additional attributes and relationships) very often the alphabetical list is complemented by other types of display.

The great weakness of any alphabetical list is the need to know a term before one can find the corresponding concept(s). Thus, an alphabetical list does not respect Roget’s vision that it would be useful to arrange terms systematically according to the ideas or concepts they represent. His literary insight applies equally in the context of information retrieval. A search concerning “wood” for example, could equally be expressed using the term “timber,” and an alphabetical list would place these terms far apart even though the underlying concept may well be the same.

The classic *TEST* thesaurus addressed this weakness by providing three indexes: permuted, hierarchical and by subject category. A derived style, slightly more elaborate, was followed in several thesauri designed by Jean Viet, including the influential *Macrothesaurus* from the Organization for Economic Cooperation and Development (OECD). A different approach was adopted by Aitchison, Gomersall and Ireland in their ground-breaking 1969 vocabulary *Thesaurifacet*, comprising a faceted classification fully integrated with a thesaurus. This approach relies on concept-based analysis from the very start, enabling elaboration of the faceted classification and subsequent derivation of a thesaurus. Aitchison and Dextre Clarke (2004) describe how Aitchison progressively refined and enhanced this technique over the decades to follow, designing a long line of thesauri such as the *UNESCO Thesaurus* (Aitchison 1977), the *BSI ROOT Thesaurus* (British Standards Institution 1981) and the *Inter-*

national thesaurus of refugee terms (Aitchison 1996). Biswas and Smith (1989) review a number of other efforts to combine a classification scheme with a thesaurus, especially the “Classaurus” and its variants developed in India by Bhattacharyya, Devadason and others. Broughton (2006a) also advocates facet analysis as the soundest basis for thesaurus construction and claims that “the generation of a thesaurus from its equivalent faceted classification is almost as automatic a process as thesaurus construction can ever hope to be” (Broughton, 2006b, 60).

Rather than a full-blooded classification, the systematic listing of preferred terms in *MeSH (Medical Subject Headings)* was a set of extensive hierarchical “tree structures” with an elaborate expressive notation that served both as a vocabulary look-up device and as a search key in the databases of MEDLARS (Medical Literature Analysis and Retrieval System) and later Medline. The first (1982) edition of the multilingual thesaurus *AGROVOC* (Leatherdale 1982), taking a different approach, avoided the need for a separate hierarchical section by embedding the complete upper and lower hierarchical context of each concept within the alphabetical display. See Figure 5.

HORSES
<i>uf</i> <i>equus caballus</i>
BT1 equidae
BT2 perissodactyla
BT3 mammals
BT4 vertebrates
BT1 livestock
BT2 domestic animals
BT3 animals
NT1 draught horses
NT1 foals
NT1 mares
NT1 racehorses
NT1 saddle horses
NT2 ponies
NT1 stallions
NT2 geldings
rt meat animals

Figure 5. Entry for a preferred term in the English version of AGROVOC’s 1982 edition.

Throughout the 1960s, 1970s and 1980s, much of the experiment was constrained by the need to provide users with printed copies of the thesaurus, and update them regularly. Even after bibliographic databases such as Medline and AGRIS became available online through host services such as Dialog, or on CD-ROM discs, the corresponding thesauri were still widely distributed in hard copy. As late as 1990 the first edition of the influential *Art & Architecture*

Thesaurus was published conventionally, and even followed in 1994 by a second edition (Getty Art History Information Program 1994) in five weighty volumes, each over 500 pages. But after that, only the electronic version has been maintained. From the 1990s onwards, most new thesauri have been published in electronic media only.

If the focus is on an electronic version, not only are the costs and hassle of printed distribution eliminated, but also there is greater freedom to change the presentation frequently in response to feedback, and develop features that support indexing and searching of any associated databases. For example, the *STW Thesaurus for Economics* is nowadays published on the web (see <http://zbw.eu/stw/version/latest/about.en.html>), enabling immediate searching of the EconBiz database and at the same time supporting linked open data applications. Similarly, *AGROVOC* has in the twenty-first century undergone huge redevelopment, exploiting SKOS to enable linked data applications and incorporating some new relationship types in its “agrontology” (Caracciolo and Keizer 2014). Figure 6 shows how an entry in *AGROVOC* looks in 2017, with simultaneous views of hierarchy and all the language equivalents, etc., on one screen, plus easy hyperlinks to all related concepts.

4.4 Maturity, senescence or rejuvenation?

A tailing-off in the popularity of thesauri has occurred from approximately the end of the 1980s, probably due to in-

creasing availability of desktop computers, as well as the rise of the internet (Dextre Clarke 2008). The new technologies have enabled alternative retrieval methods that for most applications appear less expensive than post-coordinate indexing plus thesaurus development and maintenance. From that time onwards, while a good thesaurus works no less effectively than before, its role has been relegated to relatively fewer search applications (Dextre Clarke 2016), such as retrieval from image collections (MacFarlane 2016), cultural heritage collections and bibliographic databases. In these situations, it still brings benefits, especially when implemented in linked data mode (Tudhope and Binding 2016). Shiri (2012) paints an optimistic picture of the opportunities.

Latest versions of the standards Z39-19, ISO 25964-1 and ISO 25964-2 are dated 2005 (reaffirmed 2010), 2011 (confirmed 2017) and 2013 respectively. While in them the basic principles of thesaurus design show little change from previous versions, it is clear the context in which a thesaurus operates has changed markedly. Interoperability is now the key to success—and is reflected in the content of the standards. See note 3 for some clarification of the differences between these standards.

Some authors believe the way relationships are treated in a thesaurus could usefully evolve. Alexiev et al (2014) suggest that interoperability, inferencing capabilities and the reliability of search explosion would all be improved by more rigorous discrimination between the three types of hierarchical relationship allowed by the standards. Hjørland (2016)

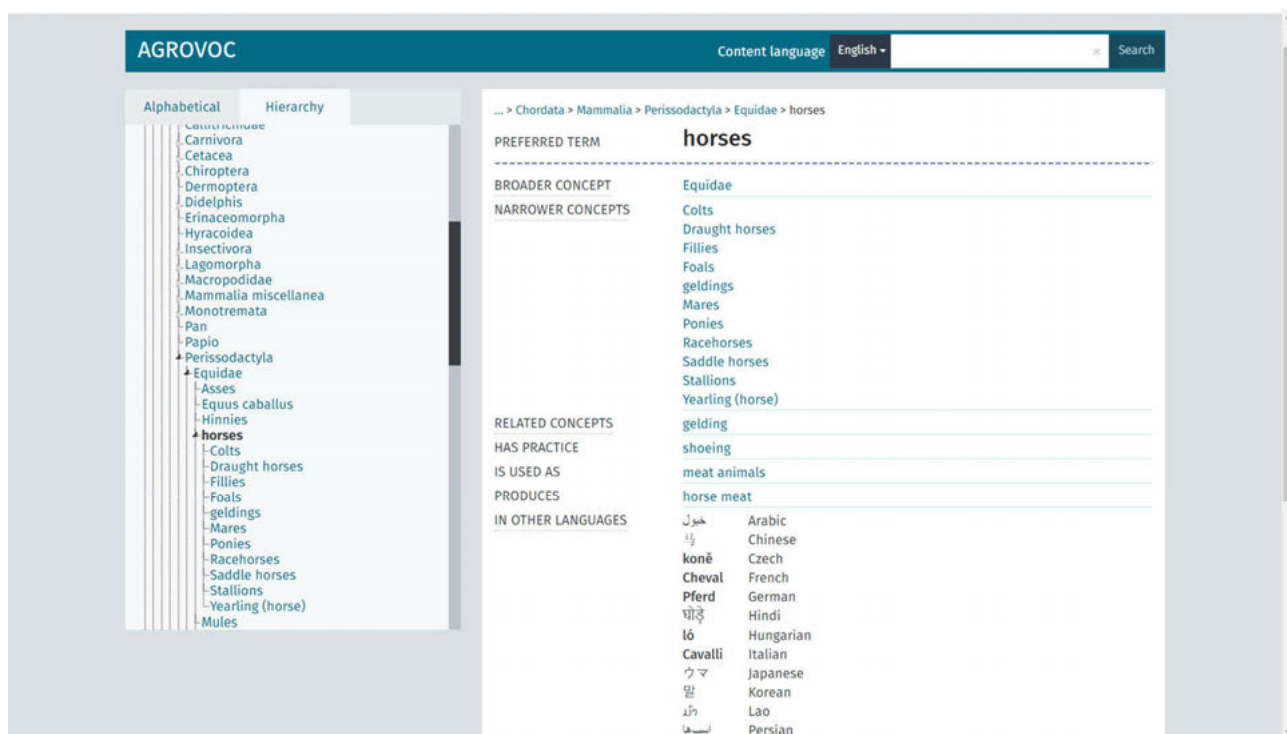


Figure 6. Online display of the entry for “horses” in AGROVOC, January 2017.

asks why thesauri “bundle” different kinds of semantic relations into one relationship type—the associative—and suggests (151) that “thesauri would probably be improved” if they adopted some attributes from ontologies, in particular the avoidance of “standardized limitations on the kind of semantic relations used.” He points out further that the most useful types of relationship to specify may vary from one domain to another. Vernau (www.iskook.org/sites/default/files/190215Debate_1-JudiVernau.mp3) too has called for changes in the approach to relationships.

Enthusiasts for change may like to note that the current standards are already permissive of developments, e.g., the inclusion of new customized relationships, that do not transgress the existing rules. All international standards are reviewed on a five-year cycle, enabling proponents to make the case for revision as soon as such developments have proved their worth. The passage of time will tell whether the thesaurus continues as before in its relatively few niche applications, or whether it blossoms into new networked opportunities, perhaps revitalised by an infusion of ideas from ontologies and other types of KOS. Section 7 below summarizes the challenges and opportunities for continuing exploitation and evolution.

5.0 Types and styles of thesaurus

5.1 Overview

The Basel Register of Thesauri, Ontologies & Classifications (BARTOC) at www.BARTOC.org and the Taxonomy Warehouse <taxonomywarehouse.com> list hundreds of thesauri among other types of KOS. Despite their astonishing variability in aspects such as subject scope, size, specificity, function, format, layout, language, quality of construction, etc., it is hard to divide them into distinct species or types. Much of the variation seems stylistic rather than fundamental, with one style borrowing features from another and a proliferation of hybrids. This section will, therefore, start by describing the “bare minimum” that can be expected in any IR thesaurus, and continue with some discussion of frequently observed differences in style, before discussing some categories of thesaurus that might or might not be considered distinct types.

5.2 The bare minimum

These features are indispensable in a traditional functioning thesaurus:

- For every concept deemed worth indexing/searching, inclusion of as many as possible of the terms that might represent it, with one of these selected as “preferred”;
- Any hierarchical or close associative relationships between the concepts should be shown;
- Some kind of display or index must enable users to look up the terms and concepts.

Taken together these three requirements typically lead to a list of all terms and relationships, with entries alphabetically arranged, in the style of the extract in Figure 2. Alongside these traditional requirements, it is worth noting a trend towards applications in which the thesaurus is implemented behind the scenes; this reduces or obviates the need for any kind of display or, indeed, for designating the preferred term for a concept.

While the vocabulary illustrated in Figure 2 complies with the standards, a more ambitious thesaurus would also incorporate scope notes, history notes, faceted arrays introduced by node labels, concept groups and other optional features. The data model in Figure 1 points to very many opportunities for enhancing a thesaurus in ways that are standards-compliant and supportive of interoperability in networked applications. As to format, the alphabetic list is often supplemented by other displays to help users find the right term, such as a classified display, a set of hierarchical trees, a permuted index or even a graphical display. Thesauri that were developed to serve a particular database sometimes show extras, such as the number of postings for each term. In lieu of explicit display, some of the extra features may be hidden, invoked only as functions of a retrieval system.

5.3 Different styles for different communities

In this section, we discuss presentational differences, which may not be fundamental to thesaurus operation but can still influence user acceptance and hence retrieval effectiveness. In certain domains, a particularly influential thesaurus has influenced the development of subsequent vocabularies. For example, *MeSH* (the *Medical Subject Headings*, list of the National Library Medicine, first issued in 1960) early on developed a set of “tree structures” having a distinctive style of notation directly functional in database retrieval (and still visible in today’s MeSH online, see <https://meshb.nlm.nih.gov/#/treeSearch>). Both the trees and the notation were emulated in later thesauri for medical applications, such as *EMTREE*, the thesaurus of Elsevier’s *Excerpta Medica* database. Similarly, the *Art & Architecture Thesaurus* first published in 1990, with a very distinctive style of facet-driven hierarchical display incorporating “guide terms,” has inspired much thesaurus development work in the heritage sector worldwide.

Other historical influences have been thesaurus maintenance software and preferences of the original designers. Thus, thesauri managed with the CAIRS or the TIKIT

package characteristically presented “stop terms” and “go terms,” and some user communities still look for these in every new thesaurus. Many thesauri funded in the twentieth century by the Commission of the European Communities used the ASTUTE software, which generated alphabetical displays with entries in the style of Figure 5. The style and conceptual approach of pioneers Jean Viet and Jean Aitchison (see 4.3 above) inspired many followers to apply the principles of classification to thesaurus development.

A good many other variations on format and layout of printed thesauri, including some graphical representations, are described and/or illustrated in Foskett (1980). Shiri (2012) provides an update, including screen layouts for electronic thesauri, to be discussed next.

5.4 Electronic thesauri

Arguably, an electronic format is just another stylistic variation, not affecting the fundamentals. Electronic thesauri have been around from the early days of online bibliographic databases such as AGRIS, CAB Abstracts, INSPEC, Engineering Index, ERIC, etc, that chose to provide their search vocabularies as a printed thesaurus and as an electronic version of the same, integrated to greater or lesser degree with the search functions of the database. In such cases, the underlying content and structure of both online and printed versions are the same.

That said, the electronic medium offers enhanced opportunities for thesaurus design, maintenance, presentation and implementation, enabling interactive retrieval functions for the users as described in Section 3 above. To exploit linked data and other semantic web applications, the electronic thesaurus should be published on the web in the format of the W3C standard *SKOS Simple Knowledge Organization System Reference* (see <https://www.w3.org/TR/skos-reference/>). ISO 25964-2 (International Organization for Standardization 2013) gives further advice on semantic interoperability between thesauri and other KOSs. Shiri (2012) discusses several examples and offers guidelines for the design of thesaurus-enhanced search interfaces.

5.5 Multilingual vs monolingual thesauri

All the stylistic variations described so far can apply to monolingual or to multilingual thesauri. The inclusion of more than one language is not just another variable—it makes a big difference to design, maintenance and use. Compare the illustration in Figure 2 with that in Figure 7, for a bilingual thesaurus (English/Spanish). The display illustrated is for use by speakers of English; an alternative, language-inverted display for speakers of Spanish would show all the terms and relationships for that language.

Multilingual thesauri can be subdivided into two types—symmetrical or not. In a symmetrical thesaurus, every concept has a preferred term in each of the languages, and the scope and relational structure is identical in each. In a non-symmetrical thesaurus, not every concept need be represented in all the languages, and the hierarchical structure may vary from one language to another to accommodate cultural differences. See more discussion and examples in Working Group on Guidelines for Multilingual Thesauri of IFLA Classification and Indexing Section (2009) and Hudon (2001).

5.6 Macro- and micro-thesauri

An original aim of the OECD’s *Macrothesaurus* published in 1972 was to “create a documentary language for processing information in the broad field of economic and social development, while striving for compatibility with sectoral thesauri serving agriculture, industry, labour, education, population, science, technology, culture communication, health and the environment” (Viet 1972, v). Both the name and the aim were popular, and so years later the term “macrothesaurus” with a small “m” was borrowed as a generic name for any broad-level thesaurus that either contains or is aligned with a number of “microthesauri” having greater specificity in a more limited field.

The normal situation according to Aitchison et al (2000, 177) is for built-in compatibility, with the specialized “microthesaurus” being “mapped onto, and entirely integrated within, the hierarchical structure of some broader thesaurus, the macrothesaurus.” They acknowledge, however, that sometimes the macrothesaurus is a separate entity, managed independently of any corresponding microthesauri.

In practice, it is not easy to maintain alignment between specialized thesauri used by different communities, unless management is centralized. Among the successful examples today is *EUROVOC*, the Multilingual Thesaurus of the European Union, which is structured into twenty-one “domains,” each of which is subdivided into a number of “microthesauri.” Concepts belong to more than one microthesaurus if appropriate. Each microthesaurus contains a hierarchically structured list of concepts, terms and relationships, and can be downloaded separately (see and browse at <http://eurovoc.europa.eu/drupal/>).

5.7 The search thesaurus

The search thesaurus is one designed for use, not in indexing, but only at the search stage. (see more discussion in Aitchison et al (2000) and Lykke Nielsen (2004)). At first glance, this would not seem to make it very different. And indeed, sometimes a normal standards-compliant thesau-

<p>pesticides</p> <p>es: plaguicidas</p> <p>UF: fumigants</p> <p>BT: agrochemicals</p> <p>NT: fungicides</p> <p>herbicidas</p> <p>insecticidas</p> <p>RT: pests</p> <p>pests</p> <p>es: plagas</p> <p>NT: pest insects</p> <p>plant pests</p> <p>RT: pesticides</p> <p>pigs</p> <p>es: cerdos</p> <p>UF: hogs</p> <p>porkers</p> <p>sows</p> <p>BT: livestock</p> <p>plant pests</p> <p>es: plagas de plantas</p> <p>BT: pests</p>	<p>plant products</p> <p>es: productos de origen vegetal</p> <p>NT: cereals</p> <p>fruits</p> <p>spices</p> <p>vegetables</p> <p>RT: plants</p> <p>plants</p> <p>es: plantas</p> <p>RT: plant products</p> <p>porkers</p> <p>USE: pigs</p> <p>poultry</p> <p>es: aves de corral</p> <p>BT: livestock</p> <p>NT: chickens</p> <p>ducks</p> <p>geese</p> <p>turkeys</p> <p>RT: eggs</p> <p>sheep</p> <p>es: ovinos</p> <p>BT: livestock</p> <p>RT: wool</p>
--	---

Figure 7. Extract from one Alphabetical display of a Spanish/English thesaurus.

Pioneers. Pioneer(s). Early settler(s). Pilgrim(s). Frontiers(man,men). Backwoods(man,men). Early colonist(s). Homesteader(s). Early immigrant(s). *Consider also:* discoverer(s), explorer(s), pathfinder(s), scout(s), trailblazer(s), leader(s). *See also:* Explorers; Pioneering; Scientists.

Figure 8. Sample entry from Knapp’s search thesaurus (Knapp 1993).

rus is applied only at the search stage, and then described as a “search thesaurus.”

A deeper study, however, reminds us of the way a standard thesaurus is designed to work (ISO 25964-1, Clause 4.1): “The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term.” In the case of the thesaurus shown in Figure 2, for example, an indexer would assign the term “pigs” to every item in the collection that deals with pigs or sows or hogs or porkers. The searcher would use only the term “pigs” to retrieve all these items. But if the same tool was being used as a search thesaurus, indexing would not have taken place. The searcher would have to look for “pigs OR sows OR hogs OR porkers.”

Thus, the notion of a “preferred term” is inapplicable to a search thesaurus designed as such. Standards such as ISO 25964 become irrelevant, allowing even greater freedom of content, style and structure. Lopez-Huertas (1997) proposes one example, structured very differently from the standard thesaurus. Another fully worked example is Knapp’s “The contemporary thesaurus of social science terms and synonyms” which attempts to remind readers of many alternative ways of expressing the same idea, using a layout quite different from the standard (see Figure 8). In practice, not many such works have been published.

The converse of the search thesaurus is the “indexing thesaurus,” to be used for indexing and not for search. While applications are sometimes found in which indexing

is enhanced automatically with the help of a thesaurus, a standard thesaurus is usually applied, rather than one designed for indexing alone.

6.0 Performance and evaluation

While the criteria presented by Mader and Haslhofer (2015) apply to a range of KOSs and not specifically thesauri, they do help evaluate interoperability in the context of SKOS use, for any controlled vocabulary. Much earlier, Owens and Cochrane (2004) described four approaches—structural, formative, observational and comparative—to thesaurus evaluation. None of these directly measures the effectiveness with which a thesaurus succeeds in the purpose for which it was intended—retrieving information. Such a measure is difficult if not impossible to devise, partly because the thesaurus is only one of several components in the retrieval system, and partly because there are so many variables in the context of use. Lengthy experiments in the 1960s and 1970s studied the effects on precision and recall as different features of indexing languages were tested, but ultimately failed to provide conclusive support for the use of any controlled vocabulary (Keen 1973; Soergel 1994; Svenonius 1986; Dextre Clarke 2001). Despite efforts over many years, we still do not have definitive proof that development and use of a thesaurus is a worthwhile investment. Dextre Clarke (2016) provides an account of the continuing debate.

The long debate was highlighted at an event run by the UK Chapter of ISKO in February 2015 (see proceedings at <http://www.iskouk.org/content/great-debate>), with a subsequent special issue of *Knowledge Organization* (v. 43, no.3 2016) devoted wholly to questioning the future of thesauri. Despite reservations expressed by Hjørland (2016), that depiction of the future makes it clear the context of KOS use is changing, and the thesaurus evolving to occupy new roles and opportunities. Although quantitative proof of efficacy may be lacking, there is plenty of qualitative evidence of thesauri prospering and supporting users in some key areas of a changing environment. Modes of evaluation may have to adapt to reflect the new context.

7.0 The future of thesauri

The thesaurus as conceived by the current national and international standards is still based on the assumption “that human intellect is usually involved in the selection of indexing terms and in the selection of search terms. If both the indexer and the searcher are guided to choose the same term for the same concept, then relevant documents will be retrieved. This is the main principle underlying thesaurus design” (ISO 25964-1, Introduction vi). Nowadays, opportunities to apply the thesaurus may shrink because the

trained indexer and searcher are increasingly scarce. End-users are largely unaware of thesauri (this is confirmed, for example, by Greenberg (2004)); trained indexers and searchers are usually deemed unaffordable. Areas where the thesaurus seems most likely to survive and flourish include:

- Applications where no other IR technology is effective (e.g. indexing of still images)
- Applications with an income to pay the costs of indexing and thesaurus maintenance (e.g. profitable bibliographic databases)
- Applications with new benefits to spread the costs over more outcomes (e.g., via linked data and/or mapping services)
- Enhanced implementation behind the scenes, so that users get the benefits without the discomforts of look-up (e.g., with automatic-aided indexing);
- Evolved or hybrid KOSs, with new characteristics that are now in demand (e.g., by incorporation of domain-specific relationships).

Examples of developments like these may be found in the special issue of *Knowledge Organization* (2016) mentioned above and in Shiri (2012).

Simultaneously as true thesauri still thrive in the types of application just listed, a parallel future may lie in their gradual transformation under the banner “taxonomy.” This term, long applied to the practice and science of classification and especially the Linnaean classification of biological organisms, has been widely applied since the 1990s to a variety of KOSs found in electronic media. Applications include corporate intranets, online retail sales outlets, digital libraries, public sector advice websites, as well as displacement of the thesaurus in some of its traditional occupations. White (2016) points to the value of KO tools and techniques in some of these contexts.

There’s still little uniformity among the “taxonomies” being developed for such applications, which may be simple heading lists, or may be complex hybrids that combine features from thesauri, traditional classification schemes, faceted schemes, ontologies and other types of KOS. In comparison with the widespread adoption of web search engines, their value is barely recognized. But there certainly is a very large need and opportunity for the principles of knowledge organization to be applied towards helping millions of workers in the knowledge society to find information resources of all kinds. The names we shall find for the emerging hybrid vocabularies are hard to predict, but we can safely say that the thesaurus will pass some of its genes into new tools for searching the cyberworld to come.

8.0 Further reading

Full details of the publications in this list may be found in the references, listed at the end of this article.

a) Useful registers of thesauri (among other types of KOS) may be found in the Basel Register of Thesauri, Ontologies & Classifications at www.BARTOC.org and the Taxonomy Warehouse at taxonomywarehouse.com. The latter site also lists relevant events, blogs, publishers and links to some associated products such as software.

b) Two guides to thesaurus construction are recommended:

- Aitchison, Jean, Alan Gilchrist and David Bawden (2000) *Thesaurus construction and use: a practical manual*.
- Broughton, Vanda (2006) *Essential thesaurus construction*.

The first of these also carries an extensive bibliography. Both guides draw heavily on the then current national and international standards for thesauri: ISO 2788, ISO 5964, BS8723 and ANSI/NISO Z39.19 (of which the first three have since been withdrawn, superseded by ISO 25964).

c) Specialist software is needed for thesaurus construction and maintenance. While ISO 25964-1 and ANSI/NISO Z39.19 both advise on the functionality required, see also the following article and its list of references:

- Will, Leonard (2010) “*Thesaurus Management Software*.”

d) A special issue of *Cataloging & Classification Quarterly* (Roe and Thomas 2004) was devoted to “The thesaurus: review, renaissance and revision,” in which all the articles have useful reference lists. Contents include:

- Aitchison and Dextre Clarke (2004) “The thesaurus: a historical viewpoint, with a look to the future.”
- Greenberg (2004) “User comprehension and searching with information retrieval thesauri.”
- Johnson (2004) “Distributed thesaurus web services.”
- Landry (2004) “Multilingual subject access: the linking approach of MACS.”
- Lykke Nielsen (2004) “Thesaurus construction: key issues and selected readings.”
- Owens and Cochrane (2004) “Thesaurus evaluation.”
- Riesland (2004) “Tools of the Trade: Vocabulary Management Software.”
- Shearer (2004) “A practical exercise in building a Thesaurus.”
- Thomas (2004) “Teach yourself thesaurus: Exercises, readings, resources.”

- Will (2004) “Thesaurus consultancy.”

e) A special issue of *Knowledge Organization* (v. 43, no. 3 2016) was devoted to a continuation of the ISKO-UK debate “This house believes that the traditional thesaurus has no place in modern information retrieval.” All the articles have useful reference lists. Contents include:

- Dextre Clarke (2016) “Origins and trajectory of the long thesaurus debate.”
- Dextre Clarke and Vernau (2016a) “Guest editorial: the thesaurus debate continues.”
- Dextre Clarke and Vernau (2016b) “Questions and answers on current developments inspired by the thesaurus tradition.”
- Garcia-Marco (2016) “Enhancing the visibility and relevance of thesauri in the Web: searching for a hub in the linked data environment.”
- Hjørland (2016) “Does the traditional thesaurus have a place in modern information retrieval?”
- Kempf and Neubert (2016) “The role of thesauri in an open Web: a case study of the STW Thesaurus for Economics.”
- MacFarlane (2016) “Knowledge Organisation and its role in multimedia information retrieval.”
- Tudhope and Binding (2016) “Still quite popular after all those years - the continued relevance of the information retrieval thesaurus.”
- White (2016) “The value of taxonomies, thesauri and metadata in enterprise search.”

f) The interest group NKOS (Networked Knowledge Organization Systems/Services/Structures) runs projects and activities enabling all sorts of KOS as networked interactive information services, especially through the internet. On its website at nkos.slis.kent.edu are links to many relevant publications and the proceedings of past events. The presentations from the NKOS workshops in USA and Europe are especially helpful in pointing to the future of thesauri and other KOSs. For an overview, see Tudhope and Lykke Nielsen 2006.

Notes

1. “Information retrieval” is used here broadly to mean “the activity of obtaining information resources relevant to an information need from one or more collections of information resources” (definition adapted from Wikipedia). It is not limited to use in systems held on computer.
2. As explained on the W3C (World Wide Web Consortium) website, “SKOS [Simple Knowledge Organization Systems] is an area of work developing specifica-

tions and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web.” A key product of this programme is the *SKOS Simple Knowledge Organization System Reference* (Miles and Bechhofer 2009), a common data model for sharing and linking knowledge organization systems via the web. Development work on this specification took place around the same time as the development of BS 8723 and ISO 25964, with regular communication between the corresponding teams, so that a high degree of compatibility was achieved. The main difference between them can be summarized as follows: While ISO 25964-1 serves as a standard for construction of thesauri, SKOS is a standard for publishing thesauri and other types of KOS on the web. Likewise, ISO 25964-2 recommends the sort of mappings that can be established between one KOS and another; SKOS presents a way of expressing these when published to the web.

The data models of these two standards are not identical, because ISO 25964 must provide for the needs of all sorts of thesauri (whether for web use or for other applications) while SKOS [29] must provide for all sorts of KOS (including classification schemes and many others that do not comply with ISO 25964). Good alignment between the two made possible a set of correspondences <ISO25964-SKOSXL-MADS-2013-12-11.pdf> between components of the data models, developed by the same teams of authors. Where the basic SKOS data model lacked a construct corresponding to a feature of the ISO 25964 model, the SKOS-XL <www.w3.org/TR/skos-reference/skos-xl.html> model was used, supplemented by additional proposals where necessary. Care was taken to avoid incompatibility with another project to align SKOS with MADS www.loc.gov/standards/mads/. Based on the documented correspondence table, an RDF schema that provides a machine-readable version for these mappings as well as for the elements from the ISO 25964 model was developed and made available at <http://purl.org/iso25964/skos-thes>. At the time of writing (July 2017), this latter site is unavailable but work is in hand to restore it. More background on all the above developments is provided at <www.niso.org/schemas/iso25964/#skos>.

It should be noted that exploitation of these interoperability standards and opportunities demands skill and attention to detail. Some practical examples and cautionary tales are provided in De Smedt (2012) and Lindenthal (2012).

3. The American standard ANSI/NISO Z39.19 and the International standard ISO 25964 are broadly aligned. Here are some key points of similarity or difference:

- The latest version of Z39.19 is a single document, issued in 2005 and reaffirmed in 2010; whereas ISO 25964 is a two-part standard, of which the first part (ISO 25964-1) was issued in 2011 and confirmed in 2017 while the second (ISO 25964-2) was issued in 2013.
- The scope of Z39.19 is broadly comparable with that of ISO 25964-1, but Z39.19 covers several types of monolingual controlled vocabulary—lists of controlled terms, synonym rings, taxonomies and thesauri—while ISO 25964-1 deals only with thesauri, both monolingual and multilingual.
- ISO 25964-1 provides a data model but Z39.19 does not
- The whole of ISO 25964-2 (ninety-nine pages) deals with interoperability between thesauri and other types of KOS, including classification schemes, taxonomies, subject heading schemes, ontologies, terminologies, name authority lists and synonym rings. The treatment of interoperability in Z39.19 is contained in one clause of eight pages (plus a five-page appendix), in which multilingual thesauri are treated as a special case of interoperability.
- Z39.19 may be downloaded free of charge from the NISO website, whereas each part of ISO 25964 currently (2017) costs 198 Swiss francs, from the ISO store.

In view of the relatively wider scope of ISO 25964 in respect of thesauri, including in-depth treatment of interoperability and provision of a data model, it has been referenced more often than Z39.19 in this article.

References

- Aitchison, Jean. 1977. *UNESCO Thesaurus*. Paris: UNESCO.
- Aitchison, Jean. 1996. *International Thesaurus of Refugee Terms*. 2nd ed. New York: United Nations High Commissioner for Refugees.
- Aitchison, Jean and Stella Dextre Clarke. 2004. “The Thesaurus: A Historical Viewpoint, with a Look to the Future.” *Cataloging & Classification Quarterly* 37, no. 3/4: 5-21.
- Aitchison, Jean and Alan Gilchrist. 1972. *Thesaurus Construction: A Practical Manual*. London: Aslib.
- Aitchison, Jean, Alan Gilchrist and David Bawden. 2000. *Thesaurus Construction and Use: A Practical Manual*. 4th ed. London: Aslib.

- Aitchison, Jean, Alan Gomersall and Ralph Ireland. 1969. *Thesaurifacet: A Thesaurus and Faceted Classification for Engineering and Related Subjects*. Whetstone, Leicester: English Electric.
- Alexiev, Vladimir, Antoine Isaac and Jutta Lindenthal. 2014. "On Compositionality of ISO 25964 Hierarchical Relationships (BTG, BTP, BTI)." Paper presented at the 13th European Networked Knowledge Organization Systems (NKOS) Workshop held during the DL2014 Conference. <https://at-web1.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2014/content/NKOS2014-abstract-alexiev-isaac-lindenthal.pdf>
- American Institute of Chemical Engineers. 1961. *Chemical Engineering Thesaurus: A Wordbook for Use with the Concept Coordination System of Information Storage and Retrieval*. New York: American Institute of Chemical Engineers.
- American National Standards Institute. 1974. *American National Standard Guidelines for Thesaurus Structure, Construction and Use*. ANSI Z39.19-1974. New York: American National Standards Institute.
- Andrade, Juliatti de and Marilda Lopes Ginez de Lara. 2016. "Interoperability and Mapping between Knowledge Organization Systems: Metathesaurus – Unified Medical Language System of the National Library of Medicine." *Knowledge Organization* 43: 107-12.
- Armed Services Technical Information Agency. 1960. *Thesaurus of ASTLA Descriptors*. Arlington, VA: Armed Services Technical Information Agency.
- Baca, Murtha and Melissa Gill. 2015. "Encoding Multilingual Knowledge Systems in the Digital Age: The Getty Vocabularies." *Knowledge Organization* 42: 232-43.
- Berners-Lee, Tim. 2006. "Linked Data." www.w3.org/DesignIssues/LinkedData.html
- Biswas, Subal C. and Fred Smith. 1989. "Classed Thesauri in Indexing and Retrieval: A Literature Review and Critical Evaluation of Online Alphabetic Classaurus." *Library and Information Science Research* 11, no. 2: 109-41.
- Boteram, Felix, Winfried Goedert and Jessica Hubrich, eds. 2011. *Concepts in Context: Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization*. Würzburg: Ergon.
- BSI (British Standards Institution). 1981. *BSI ROOT Thesaurus*. Milton Keynes, England: British Standards Institution.
- Broughton, Vanda. 2006a. *Essential Thesaurus Construction*. London: Facet.
- Broughton, Vanda. 2006b. "The Need for a Faceted Classification as the Basis of All Methods of Information Retrieval." *Aslib Proceedings* 58: 49-72.
- Caplan, Priscilla Louise. 1978. "Thesaurus-Based Automatic Indexing: A Study of Indexing Failure." PhD diss., University of North Carolina.
- Caracciolo, Caterina and Johannes Keizer. 2014. "What KOS can do, with the Proper Tools Available. About AGROVOC, Edited in Vocbench and Used in the AGRIS Web Application." Paper presented at Knowledge Organization - Making a Difference: ISKO UK Biennial Conference. <http://www.iskouk.org/sites/default/files/CaraccioloSlidesISKO-UK2015.pdf>
- De Keyser, Pierre. 2012. "Introduction to Subject Headings and Thesauri." In *Indexing: From Thesauri to the Semantic Web*, ed. Pierre De Keyser. Oxford: Chandos, 1-37.
- De Smedt, Johan. 2012. "Exchanging ISO 25964-1 Thesauri Data using RDF, SKOS and SKOS-XL." Paper presented at the 11th European Networked Knowledge Organization Systems (NKOS) Workshop, Sept 2012. https://at-web1.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2012/presentations/ISO25964-mapping-to-SKOS-XL_TPDL-2012-09-27v6.pdf
- Dextre Clarke, Stella G. 2001. "Organising Access to Information by Subject." In *Handbook of Information Management*. 8th ed., ed. Alison Scammell. London: Aslib, 72-110.
- Dextre Clarke, Stella G. 2008. "The Last 50 Years of Knowledge Organization: A Journey through my Personal Archives." *Journal of Information Science* 34: 427-37.
- Dextre Clarke, Stella G. 2011a. "In Pursuit of Interoperability: Can we Standardize Mapping Types?" In *Concepts in Context: Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization*, ed. Felix Boteram, Winfried Goedert and Jessica Hubrich. Würzburg: Ergon, 91-109.
- Dextre Clarke, Stella G. 2011b. "Knowledge Organization System Standards." In *Encyclopedia of Library and Information Science*. 3rd ed., ed. Marcia J. Bates and Mary Niles Maack. Boca Raton, FL: CRC Press, 3176-83.
- Dextre Clarke, Stella G. 2016. "Origins and Trajectory of the Long Thesaurus Debate." *Knowledge Organization* 43: 138-44.
- Dextre Clarke, Stella G. and Marcia Lei Zeng. 2011. "Standard Spotlight: From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling." *Information Standards Quarterly* 25, no. 1: 20-6.
- Dextre Clarke, Stella G. and Vernau, Judi. 2016a. "Guest Editorial: The Thesaurus Debate Continues." *Knowledge Organization* 43: 135-7.
- Dextre Clarke, Stella G. and Vernau, Judi. 2016a. "Questions and Answers on Current Developments Inspired by the Thesaurus Tradition." *Knowledge Organization* 43:203-9.
- Doerr, Martin. 2000. "Semantic Problems of Thesaurus Mapping." *Journal of Digital Information* 1, no. 8. <http://jodi.ecc.soton.ac.uk/Articles/v01/i08/Doerr/>

- Elsevier Science. 1991- *EMTREE: The Life Science Thesaurus*. Amsterdam: Elsevier Science.
- Fast, Karl. 2003. "Controlled Vocabularies: A Glosso-Thesaurus." *boxesandarrows* (blog), October 25. <http://boxesandarrows.com/controlled-vocabularies-a-glosso-thesaurus/>
- Foskett, Douglas J. 1980. "Thesaurus." *Encyclopedia of Library and Information Science*, ed. Allen Kent, Harold Lancour and J. E. Daily. New York: Marcel Dekker, 30: 416-62.
- Garcia-Marco, Francisco Javier. 2016. "Enhancing the Visibility and Relevance of Thesauri in the Web: Searching for a Hub in the Linked Data Environment." *Knowledge Organization* 43: 193-202.
- Garshol, Lars Marius. 2004. "Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it All." *Journal of Information Science* 30: 378-91.
- Getty Art History Information Program. 1994. *Art & Architecture Thesaurus*. 2nd ed. Oxford: Oxford University Press.
- Gilchrist, Alan. 1971. *The Thesaurus in Retrieval*. London: Aslib.
- Greenberg, Jane. 2004. "User Comprehension and Searching with Information Retrieval Thesauri." *Cataloging & Classification Quarterly* 37, no. 3/4: 103-20.
- Hjørland, Birger. 2015. "Classical Databases and Knowledge Organization: a Case for Boolean Retrieval and Human Decision-Making during Searches." *Journal of the Association for Information Science and Technology* 66, no. 8: 1559-75.
- Hjørland, Birger. 2016. "Does the Traditional Thesaurus Have a Place in Modern Information Retrieval?" *Knowledge Organization* 43: 145-59.
- Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington DC: Council on Library and Information Resources. <http://www.clir.org/pubs/reports/pub91/contents.html>
- Hood, Martha W. and Christine Ebermann. 1990. "Reconciling the CAB Thesaurus and AGROVOC." *LAALD Quarterly Bulletin* 35, no. 3: 181-5.
- Hoppe, Stephan. 1996. "The UMLS – a Model for Knowledge Integration in a Subject Field." In *Compatibility and Integration of Order Systems: Research Seminar Proceedings of the TIP/ISKO Meeting*, ed. Ingetraut Dahlberg and Krystyna Siwek. Warsaw: Wydawnictwo SBP, 97-110.
- Horsnell, Verina. 1975. "The Intermediate Lexicon: An Aid to International Co-Operation." *Aslib Proceedings* 27: 57-66.
- Hudon, Michele. 2001. "Relationships in Multilingual Thesauri." In *Relationships in the Organization of Knowledge*, ed. Carol A Bean and Rebecca Green. Dordrecht: Kluwer, 67-80.
- IFLA Working Group on Guidelines for Multilingual Thesauri. 2009. "Guidelines for Multilingual Thesauri." <http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf>
- ISO (International Organization for Standardization). 1974. *Guidelines for the Establishment and Development of Monolingual Thesauri*. ISO 2788-1974. Geneva: International Organization for Standardization.
- ISO (International Organization for Standardization). 1985. *Guidelines for the Establishment and Development of Multilingual Thesauri*. ISO 5964-1985. Geneva: International Organization for Standardization.
- ISO (International Organization for Standardization). 1996. *Guidelines for the Content, Organization and Presentation of Indexes*. ISO 999:1996. Geneva: International Organization for Standardization.
- ISO (International Organization for Standardization). 2011. *Thesauri for Information Retrieval. Information and Documentation*. Part 1 of *Thesauri and Interoperability with Other Vocabularies*. ISO 25964-1. Geneva: International Organization for Standardization.
- ISO (International Organization for Standardization). 2013. *Interoperability with Other Vocabularies*. Part 2 of *Thesauri and Interoperability with Other Vocabularies*. ISO 25964-2. Geneva: International Organization for Standardization.
- Isaac, Antoine and Thomas Baker. 2015. "Linked Data Practice at Different Levels of Semantic Precision: The Perspective of Libraries, Archives and Museums." *ASIS&T Bulletin*. http://www.asis.org/Bulletin/Apr-15/AprMay15_Isaac_Baker.html
- Joyce, T. and R. M. Needham. 1958. "The Thesaurus Approach to Information Retrieval." *American Documentation* 9: 192-97.
- Keen, E. Michael. 1973. "The Aberystwyth Index Languages Test." *Journal of Documentation* 29: 1-35.
- Kempf, Andreas Oscar and Joachim Neubert. 2016. "The Role of Thesauri in an Open Web: a Case Study of the STW Thesaurus for Economics." *Knowledge Organization* 43: 160-73.
- Kless, Daniel, Simon Milton and Edmund Kazmierczak. 2012. "Relationships and Relata in Ontologies and Thesauri: Differences and Similarities." *Applied Ontology* 7, no. 4: 401-28.
- Knapp, Sara D. 1993. *The Contemporary Thesaurus of Social Science Terms and Synonyms. A Guide for Natural Language Computer Searching*. Phoenix: Oryx.
- Krooks D. A. and Lancaster F. W. 1993. "The Evolution of Guidelines for Thesaurus Construction." *Libri* 43, no. 4: 326-42.
- Lancaster, F. W. 1972. *Vocabulary Control for Information Retrieval*. Washington, D.C.: Information Resources Press.
- Lancaster, F. W. 1998. *Indexing and Abstracting in Theory and Practice*. London: LA Publishing.

- Leatherdale, Donald. 1982. *AGROVOC*. Rome: Food and Agriculture Organization of the United Nations and Commission of the European Communities.
- Lindenthal, Jutta. 2012. "Ambiguities in Representing Thesauri Using Extended SKOS - Examples from Real Life". Paper presented at the 11th European Networked Knowledge Organization Systems (NKOS) Workshop, Sept 2012. https://at-web1.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2012/presentations/TPDL2012_NKOS_Ambiguities_R1.pdf
- Lopez-Huertas, M. J. 1997. "Thesaurus Structure Design: A Conceptual Approach for Improved Interaction." *Journal of Documentation* 53: 139-77.
- Lykke Nielsen, Marianne. 2001. "A Framework for Work Task Based Thesaurus Design." *Journal of Documentation* 57: 774-97.
- Lykke Nielsen, Marianne. 2004. "Thesaurus Construction: Key Issues and Selected Readings." *Cataloging & Classification Quarterly* 37, no. 3/4: 57-74.
- MacFarlane, Andrew. 2016. "Knowledge Organisation and its Role in Multimedia Information Retrieval." *Knowledge Organization* 43: 180-3.
- Mader, Christian and Bernard Haslhofer. 2015. "Quality Criteria for Controlled Web Vocabularies." Paper presented at International Conference on Theory and Practice of Digital Libraries 2011, NKOS Workshop. http://eprints.cs.univie.ac.at/2923/1/tpdl_workshop.pdf
- Mayr, Philipp and Vivien Petras. 2008. "Building a Terminology Network for Search: the KoMoHe Project." In *Metadata for Semantic and Social Applications: Proceedings of the International Conference on Dublin Core and Metadata Applications*, ed. Jane Greenberg and Wolfgang Klas. Singapore: Dublin Core Metadata Initiative, 177-82.
- Miles, Alistair. 2006. "Retrieval and the Semantic Web: Incorporating a Theory of Retrieval Using Structured Vocabularies." PhD diss., Oxford Brookes University. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.7212&rep=rep1&type=pdf>
- Miles, Alistair and Bechhofer, Sean, eds. 2009. "SKOS Simple Knowledge Organization System Reference: W3C Recommendation." <http://www.w3.org/TR/skos-reference>
- NISO (National Information Standards Organization). (2005) 2010. *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. ANSI/NISO Z39.19-2005 (R2010). Bethesda, MD: National Information Standards Organization.
- National Library of Medicine. 1960-. *Medical Subject Headings (MeSH)*. Bethesda, MD: National Library of Medicine.
- Office of Naval Research. 1967. *Thesaurus of Engineering and Scientific Terms (TEST)*. New York: Engineers Joint Council and US Department of Defense.
- Owens, Lesley Ann and Pauline Atherton Cochrane. 2004. "Thesaurus Evaluation." *Cataloging & Classification Quarterly* 37, no. 3/4: 87-102.
- Riesthuis, Gerhard J. A. and Steffi Bliedung. 1991. "The-saurification of the UDC." In *Tools for Knowledge Organization and the Human Interface. Proceedings of the 1st International ISKO Conference*, ed. Robert Fugmann. Frankfurt/Main: Indeks Verlag, 2, 109-17.
- Roberts, N. 1984. "The Pre-History of the Information Retrieval Thesaurus." *Journal of Documentation* 40: 271-85.
- Roe, Sandra K. and Alan R. Thomas, eds. 2004. "The Thesaurus: Review Renaissance and Revision." Special issue, *Cataloging & Classification Quarterly* 37, nos. 3-4.
- Roget, Peter. 1952. *Everyman's Thesaurus of English Words and Phrases*, ed. D. C. Browning. London: J. M. Dent.
- Rolland-Thomas, P. 1993. "Thesaural Codes: An Appraisal of their Use in the Library of Congress Subject Headings." *Cataloging & Classification Quarterly* 16, no. 2: 71-91.
- Sharp, J. R. 1967. "Information Retrieval." In *Handbook of special librarianship and information work*. 3rd ed., ed. Wilfred Ashworth. London: Aslib, 141-232.
- Shiri, Ali. 2012. *Powering search. The Role of Thesauri in New Information Environments*. Medford, New Jersey: Information Today.
- Shiri, Ali Asghar, Crawford Revie and Gobinda Chowdhury. 2002. "Thesaurus-Enhanced Search Interfaces." *Journal of Information Science* 28: 111-22.
- Soergel, Dagobert. 1974. *Indexing Languages and Thesauri: Construction and Maintenance*. Los Angeles, CA: Melville.
- Soergel, Dagobert. 1994. "Indexing and Retrieval Performance: The Logical Evidence." *Journal of the American Society for Information Science* 45: 589-99.
- Soergel, Dagobert. 2014. "Knowledge Organization for Learning." In *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects: Proceedings of the Thirteenth International ISKO Conference 19-22 May 2014, Kraków, Poland*, ed. Wieslaw Babik. Advances in Knowledge Organization 14. Würzburg: Ergon, 22-32.
- Spero, Simon E. 208. "LCSH is to Thesaurus as Doorbell is to Mammal: Visualizing Structural Problems in the Library of Congress Subject Headings." Paper presented at the DCMI International Conference on Dublin Core and Metadata Applications in Berlin, 22-26 September 2008. <http://dcpapers.dublincore.org/pubs/article/view/937>
- Spero, Simon E. 2012. "What, if Anything, is a Subdivision?" In *Facets of Knowledge Organization: Proceedings of the ISKO UK Second Biennial Conference*, ed. Alan Gilchrist and Judi Vernau. Bingley, UK: Emerald, 69-83.
- Svenonius, Elaine. 1986. "Unanswered Questions in the Design of Controlled Vocabularies." *Journal of the American Society for Information Science* 37: 331-40.
- Tudhope, Douglas and Ceri Binding. 2016. "Still Quite Popular after All Those Years - the Continued Relevance of

- the Information Retrieval Thesaurus.” *Knowledge Organization* 43: 174-79.
- Tudhope, Douglas, Traugott Koch and Rachel Heery. 2006. “Terminology Services and Technology: JISC State of the Art Review.” <http://www.ukoln.ac.uk/terminology/JISC-review2006.html>
- Tudhope, Douglas and Marianne Lykke Nielsen. 2006. “Introduction to Knowledge Systems and Services. *New Review of Hypermedia and Multimedia* 12, no. 1: 3-9.
- Viet, Jean. 1972. *Macrothesaurus for Information Processing in the Field of Economic and Social Development*. Paris: OECD Development Centre.
- Vizine-Goetz, Diane, Carol Hickey, Andrew Houghton and Roger Thompson. 2004. “Vocabulary Mapping for Terminology Services.” *Journal of Digital Information* 4, no. 4. <https://journals.tdl.org/jodi/index.php/jodi/article/view/114/113>
- Wellisch, Hans H. 1995. *Indexing from A to Z*. 2nd ed. New York: H W Wilson.
- White, Martin. 2016. “The Value of Taxonomies, Thesauri and Metadata in Enterprise Search.” *Knowledge Organization* 43: 184-92.
- Will, Leonard. 2010. “Thesaurus Management Software.” In *Encyclopedia of Library and Information Sciences*. 3rd ed., ed. Marcia J. Bates and Mary Niles Maack. Boca Raton, FL: CRC Press, 5238-46.
- Will, Leonard. 2012. “The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus.” *Bulletin of the American Society for Information Science and Technology* 38, no. 4. https://www.asis.org/Bulletin/Apr-12/AprMay12_Will.pdf
- Zeng, Marcia Lei. 2008. “Knowledge Organization Systems (KOS).” *Knowledge Organization* 35: 160-82.
- Zeng, Marcia Lei and Lois Mai Chan. 2004. “Trends and Issues in Establishing Interoperability among Knowledge Organization Systems.” *Journal of the American Society for Information Science and Technology* 55: 377-95.

Astronomy's Three Kingdom System: A Comprehensive Classification System of Celestial Objects†

Steven J. Dick

Library of Congress, 101 Independence Ave SE, Washington, DC 20540, USA

<stevenjdick@comcast.net>



Steven J. Dick served as NASA Chief Historian and Director of the NASA History Office from 2003 to 2009. He was the 2014 Baruch S. Blumberg NASA/Library of Congress Chair in Astrobiology at the Library of Congress's John W. Kluge Center. In 2013, he testified before the United States Congress on the subject of astrobiology. He is the author or editor of twenty-three books, including *Discovery and Classification in Astronomy: Controversy and Consensus* (Cambridge, 2013), and *Classifying the Cosmos: How We Can Make Sense of the Celestial Landscape* (Springer, 2019). In 2009, minor planet 6544 Stevendick was named in his honor.

Dick, Steven J. 2019. "Astronomy's Three Kingdom System: A Comprehensive Classification System of Celestial Objects." *Knowledge Organization* 46(6): 460-466. 22 references. DOI:10.5771/0943-7444-2019-6-460.

Abstract: Although classification has been an important aspect of astronomy since stellar spectroscopy in the late nineteenth century, to date no comprehensive classification system has existed for all classes of objects in the universe. Here we present such a system, and lay out its foundational definitions and principles. The system consists of the "Three Kingdoms" of planets, stars and galaxies, eighteen families, and eighty-two classes of objects. Gravitation is the defining organizing principle for the families and classes, and the physical nature of the objects is the defining characteristic of the classes. The system should prove useful for both scientific and pedagogical purposes.

Received: Accepted: 9 May 2019

Keywords: classification, celestial objects, Three Kingdom System, kingdom, family, class, astronomical

† Derived from the article of similar title in the ISKO Encyclopedia of Knowledge Organization, Version 1.0, published 2019-05-08. Article category: KOS, specific (domain specific)

1.0 Introduction to the Three Kingdom System

This article introduces a classification system of celestial objects developed by the author. In contrast to biology, physics and chemistry, and despite a long and distinguished history of classifying specific objects such as stars and galaxies, astronomy lacks a comprehensive classification system for what has become a veritable celestial zoo. What would such a system look like, and based on what principles? Here we present a system devised for pedagogic use over the last several decades (Figure 1) but that will also be useful for scientific purposes. This so-called "Three Kingdom" system begins with the three "kingdoms" of planets, stars and galaxies, stipulates six "families" for each kingdom, and distinguishes eighty-two distinct "classes" of astronomical objects. Like biology, it is hierarchical, extending from kingdom to family to class, with the possible extension to further categories lower in the hierarchy such as type and subtype. As in biological classification it occasion-

ally adds an intermediate subfamily level wherever useful. With the benefit of hindsight, and with utility in mind, the system incorporates some classes as they have historically been defined, and adds others as they might be defined in a more coherent and consistent system.

In constructing such a system, one immediately runs into the problem of how to define the categories of kingdom, family and class. The three kingdoms adopted here (planets, stars, galaxies) are the three canonical divisions adopted in astronomy textbooks for almost a century, since it became clear that galaxies were indeed a separate realm from our Milky Way Galaxy, as determined by the American astronomer Edwin Hubble in the early 1920s. For each kingdom, six astronomical families are delineated, based on the object's origin (proto-), location (circum- and inter-), subsidiary status (sub-) and tendency to form systems (systems), in addition to the "central" family (planet, star or galaxy) with respect to which the other families are defined. These considerations give rise to astronomy's

eighteen families, and the symmetry of the six families of each kingdom reflects their physical basis in gravity's action in all three kingdoms.

For a more general introduction to astronomical classification and its issues see Buta, Corwin et al. (2007), DeVorkin (1981), Dick (2013; 2018), Feigelson (2012), Gray and Corbally (2009), Morgan (1937; 1988), Morgan and Keenan (1973) and Sandage (2005).

2.0 Defining astronomy's eighty-two classes

The Three Kingdom System contains eighty-two classes of objects, as delineated in Figure 1.

But this begs the question: How does one define a class of astronomical objects? More specifically, how does one recognize a new class of objects? We have tackled these questions in previous books, including *Discovery and Classification in Astronomy: Controversy and Consensus* (Dick 2013), and *Classifying the Cosmos: How We Can Make Sense of the Celestial Landscape* (Dick 2018), in which the Three Kingdom System is laid out in full and the history and science of each class is described.

One way of approaching the question of the definition of class is by looking at history, where (exceptions like stars and galaxies notwithstanding) classification has often been ad hoc, haphazard and historically contingent on circumstance. If astronomical history demonstrates anything, it is that the classification of astronomical objects has been based on many characteristics, depending on the state of knowledge and the needs of a particular community at the time. For example, planets could be divided according to their physical nature (terrestrial, gas giant and ice giants) or as the recent discovery of planetary systems has taught us, by orbital characteristics (highly elliptical or circular), proximity to their parent star ("hot Jupiters") and so on. Historically, binary stars have often been classified by the method of observation as visual, spectroscopic, eclipsing and astrometric, or (after more information became known) by the configuration or contents of the system, such as a white dwarf binary, or by the dominant wavelength of its electromagnetic radiation, as in an X-ray binary. While these overlapping systems have served astronomers well and illustrate how the same object may be classified in many ways, such designations are the source of much confusion among students, not to mention indecipherable to the public.

History also demonstrates that at the time of discovery, by the very nature of the problem, it is sometimes difficult to decide whether a new class of object has been discovered. Perhaps by analogy with the Earth's moon, Galileo decided relatively quickly that the four objects he first saw circling Jupiter in 1610 were satellites, proof that the moon was not unique, but a member of a class of circumplane-

tary objects (even if he did not speak in terms of "class"). But the object he first saw surrounding Saturn was not at all obviously a ring, and awaited the interpretation of Christiaan Huygens more than forty years later. Even in the late twentieth-century it was not immediately evident that pulsars were neutron stars, or that quasars were active galactic nuclei, both qualifying in the end for new class status.

Inconsistency notwithstanding, the criterion that astronomers have most often used in the astronomical literature for determining class status—and the one we adopt for the Three Kingdom system—is the physical nature of the object. In the planetary Kingdom, for example, rather than orbital characteristics, the definition of planetary classes in our own solar system has been based on their physical characteristics as rocky, gaseous or icy in composition; pulsar planets have also been distinguished by being inferred as physically very different again due to the extreme nature of their environment and probable different origin. As we have noted, new classes of planets will undoubtedly be uncovered as observations of extrasolar planets progress, but thus far not enough is known about their physical nature to do so. Many of the extrasolar planets discovered so far are believed to be gas giants; many are close to their stars and thus called "hot Jupiters." The first terrestrial extrasolar planets have also been claimed, in the form of "super-Earths" and the first rocky transiting system, known as CoRoT-7b.

This history indicates that a comprehensive classification system for astronomy can perhaps do no better than to use the typological definition of "class" largely discarded by biologists (Mayr 1988, 337): "membership in a class is determined strictly on the basis of similarity, that is, on the possession of certain characteristics shared by all and only members of that class. In order to be included in a given class, items must share certain features which are the criteria of membership or, as they are usually called, the 'defining properties.' Members of a class can have more in common than the defining properties, but they need not. These other properties may be variable—an important point in connection with the problem of whether or not classes may have a history."

But what is the unit of classification for astronomy? For physics, it is elementary particles. For chemistry, it is the elements defined by atomic number in the Periodic Table. For biology, it is species at the macro level, giving rise to biology's "five kingdoms," still favored by some microbiologists, and genetic sequences of 16S ribosomal RNA at the molecular level, giving rise to Carl Woese's "three domains" of Archaea, Bacteria and Eucarya—favored by most molecular biologists.¹ For astronomy, the unit of classification adopted here is the astronomical object itself, and with some theoretical justification. For as strong and

Astronomy's 82 Classes

Kingdom of the Planets	Kingdom of the Stars	Kingdom of the Galaxies
Family: Protoplanetary Class P 1: Protoplanetary Disk Family: Planet Class P 2: Terrestrial (rocky) Class P 3: Gas Giant Class P 4: Ice Giant Class P 5: Pulsar Planet Family: Circumplanetary Class P 6: Satellite Class P 7: Ring Class P 8: Radiation Belt Family: Subplanetary Class P 9: Dwarf Planet Class P 10: Meteoroid Subfamily: Small Bodies of Solar System Class P 11: Minor Planet/ Asteroid Class P 12: Comet Class P 13: Trans-Neptunian Objects Family: Interplanetary Medium Class P 14: Gas Class P 15: Dust Subfamily: Energetic Particles Class P 16: Solar Wind Class P 17: Anomalous Cosmic Ray	Family: Protostellar Class S 1: Protostar Family: Star Subfamily: Pre-Main Sequence Class S 2: T Tauri Class S 3: Herbig Ae/Be Subfamily: Main Sequence (H burning - Luminosity Class V) Class S 4: Dwarf Class S 5: Subdwarf Subfamily: Post-Main Sequence (He burning and higher elements) Class S 6: Subgiant (Luminosity Class IV) Class S 7: Giant (Luminosity III) Class S 8: Bright Giant Class II) Class S 9 Supergiant (Lumin. Class I) Class S 10 Hypergiant (Lumin. Class 0) Subfamily: Evolutionary Endpoints Class S 11 Supernova Class S 12 White Dwarf Class S 13 Neutron Star/Pulsar Class S 14 Black Hole Family: Circumstellar Class S 15: Debris disk Class S 16: Shell (dying stars) Class S 17: Planetary Nebula Class S 18: Nova Remnant Class S 19: Core Collapse Supernova	Family: Protogalactic Class G 1: Protogalaxy Family: Galaxy Subfamily: Normal Class G 2 Elliptical Class G 3 Lenticular Class G 4 Spiral Class G 5 Irregular Subfamily: Active Class G 6 Seyfert Class G 7 Radio Galaxy Class G 8 Quasar Class G 9 Blazar Family: Circumgalactic Class G 10 Satellites and Stellar Streams Class G 11 Galactic Jet Class G 12 Galactic Halo Family: Subgalactic Class G 13 Subgalactic Object Family: Intergalactic Medium Subfamily: Gas Class G 14 Warm Hot IGM Class G 15 Lyman alpha blobs Subfamily: Dust Class G 16 Dust

Figure 1. The Three Kingdom (3K) System. From Dick (2019, xx-xxi); reproduced with permission.

weak forces are dominant in particle physics, and as the electromagnetic force is dominant in chemistry (except for nuclear chemistry), so in astronomy is it the weakest but most far-reaching force of gravity that predominantly acts on and shapes these astronomical objects. Though other considerations such as hydrostatics and gas and radiation pressure come into play, gravity is the determining factor for the structure and organization of planets, stars and galaxies, their families and classes of objects. To put it another way, the strong interaction holds protons and neutrons together and allows atoms to exist; the electromagnetic interaction holds atoms and molecules together and allows the Earth to exist; and the gravitational interaction holds astronomical bodies together and allows the solar system, stellar systems and galactic systems to exist.² Gravity is thus a prime candidate—the one adopted here—to serve as the chief organizing principle for a comprehensive classification system for all astronomical objects.

Where does such a definition of class lead in the construction of a classification system? In the “kingdom of the stars,” stellar spectra were first classified on what turned out

to be a temperature sequence, a system devised at Harvard in the late nineteenth-century with its familiar O, B, A, F, G, K and M stars and so on. Spectra were later classified on a luminosity scale, devised at Yerkes Observatory in the 1940s, the so-called MKK (Morgan-Keenan-Kellman) system with its dwarfs, giants and supergiants.³ Which to choose to delineate “classes” for stars in a more comprehensive system for astronomical objects? We have adopted the Yerkes/MKK system (now known as the MK system) as a more evolved two-dimensional system based on spectral lines sensitive not only to temperature, but also to surface gravity (g) and luminosity. As astronomers Richard Gray and Christopher Corbally recently put it in their magisterial volume *Stellar Spectral Classification* (2009, 10), in connection with the luminosity classes, “Stars readily wanted to be grouped according to gravity as well as according to temperature, and this grouping could be done by criteria in their spectra.” The resulting luminosity classes (main sequence, subgiant, giant, bright giant and supergiant labeled from Roman numeral V to I respectively), together with the stellar endpoint classes (supernova, white dwarf, neutron star and

Astronomy's 82 Classes (cont.)

<p>Family: Systems</p> <p style="padding-left: 20px;">Class P 18: Planetary Systems/ Exoplanets</p> <p style="padding-left: 20px;">Class P 19: Asteroid Groups</p> <p style="padding-left: 20px;">Class P 20: Meteoroid streams</p> <p style="padding-left: 20px;">Subfamily: Trans-Neptunian Systems</p> <p style="padding-left: 40px;">Class P 21: Kuiper Belt</p> <p style="padding-left: 40px;">Class P 22: Oort Cloud</p>	<p style="padding-left: 20px;">Class S 20: Stellar Jet</p> <p style="padding-left: 20px;">Class S 21: Herbig-Haro Object</p> <p style="padding-left: 20px;">[See also Protoplanetary Disk (P 1); Planetary System, (P 18) Kuiper Belt (P 21) Oort Cloud (P 22)]</p> <p>Family: Substellar</p> <p style="padding-left: 20px;">Class S 22: Brown dwarf</p> <p>Family: Interstellar Medium</p> <p style="padding-left: 20px;">Subfamily: Gas (99%)</p> <p style="padding-left: 40px;">Class S 23: Cool Atomic Cloud (H I)</p> <p style="padding-left: 40px;">Class S 24: Hot Ionized Cloud (H II)</p> <p style="padding-left: 40px;">Class S 25: Molecular Cloud (H₂)</p> <p style="padding-left: 40px;">Class S 26: White Dwarf Supernova Remnant</p> <p style="padding-left: 20px;">Subfamily: Dust (1%)</p> <p style="padding-left: 40px;">Class S 27: Dark Nebulae</p> <p style="padding-left: 40px;">Class S 28: Reflection Nebulae</p> <p style="padding-left: 20px;">Subfamily: Energetic Particles</p> <p style="padding-left: 40px;">Class S 29: Stellar Wind</p> <p style="padding-left: 40px;">Class S 30: Galactic Cosmic Rays</p> <p>Family: Systems</p> <p style="padding-left: 20px;">Class S 31: Binary Star</p> <p style="padding-left: 20px;">Class S 32: Multiple Star</p> <p style="padding-left: 20px;">Class S 33: Association (OB)</p> <p style="padding-left: 20px;">Class S 34: Open Cluster</p> <p style="padding-left: 20px;">Class S 35: Globular Cluster</p> <p style="padding-left: 20px;">Class S 36: Population</p>	<p>Subfamily: Energetic Particles</p> <p style="padding-left: 20px;">Class G 17 Galactic Wind</p> <p style="padding-left: 20px;">Class G 18 Extragalactic Cosmic Rays</p> <p>Family: Systems</p> <p style="padding-left: 20px;">Class G 19 Binary</p> <p style="padding-left: 20px;">Class G 20 Interacting</p> <p style="padding-left: 20px;">Class G 21 Group</p> <p style="padding-left: 20px;">Class G 22 Cluster</p> <p style="padding-left: 20px;">Class G 23 Supercluster</p> <p style="padding-left: 20px;">Class G 24 Filaments & Voids</p>
--	---	--

Figure 1 (cont.)

black hole) not only have significance in the evolutionary sequence but also have a real history of discovery that can be uncovered. W. W. Morgan delineated these luminosity classes to begin with, because he realized each grouping of stars formed a sequence of near constant log g (surface gravity) (Gray and Corbally 2009, 9-10; Morgan 1937, 380 ff.). Thus, gravity as a sculpting force for stars was recognized already by the founders of the MKK system as the dominating force for the luminosity classes.

The choice of luminosity for stellar classes does not subordinate the Harvard system of spectral types. To the contrary, Harvard spectral types are still an integral part of the system. As the originators of the Yerkes/MKK system argued, it is simply the case that their system contains more information and better represents the physical nature of stars, as astronomers gradually separated them (over the thirty years from 1910 to 1940) into supergiants, bright giants, giants and subgiants. In other words, since 1943 with the Yerkes/MKK system, modern astronomy has a formal two-dimensional temperature-luminosity system with distinct classes, building on the Hertzsprung-Russell diagram,

which was literally a two-dimensional plot of temperatures versus luminosities when it was first constructed around 1914. Both the Harvard and the Yerkes systems are represented in the full designation of a star, as in Sirius (A1V) as a main sequence star with Harvard spectral type A1.

Thus, choices for class status become more clear-cut once there is a guiding principle such as physical meaning, which goes to the heart of Morgan's quest for "the thing itself." Again in the stellar kingdom, for the interstellar medium instead of "diffuse nebulae" (a morphological classification), classes in the Three Kingdom System are distinguished according to physical constitution of the nebulae: gas (cool atomic neutral hydrogen, hot ionized hydrogen and molecular) and dust (reflection nebulae). These categories are used in astronomy and subsume classifications based on morphology that are historically contingent. In the galactic kingdom, galaxy morphologies (elliptical, lenticular, spiral, barred spiral and irregular) laid out by Edwin Hubble in the 1920s also reflect compositional differences (as Morgan's galaxy classification system showed), so the principle of physical meaningfulness still holds.

3.0 Classification principles in the Three Kingdom System

As we have stipulated, by definition kingdoms are delineated by the three central prototypes of objects in the universe—planets, stars and galaxies, as enshrined in canonical textbooks since the 1950s. Families are delineated by the various manifestations of the gravitational force acting on astronomical objects, e.g., protoplanetary, planetary, circumplanetary, subplanetary, interplanetary and systems. As in any classification system, there will be ambiguities of placement in lower taxon levels. These can be mitigated by a system of classification principles. For the Three Kingdom System, these include the following when it comes to the determination of classes and the placement of objects in classes:

- 1) Classes are delineated based on the physical nature of the object, defined as physical composition wherever possible.
- 2) An object should always be placed in its most specific class.
- 3) To the extent possible, classes already in use are retained, as in the luminosity classes of the MK system and the Hubble classes for galaxies, supplemented by new knowledge.
- 4) The recommendations of the International Astronomical Union are followed; e.g., a dwarf planet is not a class of planet.
- 5) Potential, but unverified, classes are not included.

Figure 1 shows the result of applying these principles to astronomical objects. For those who do not recognize their favorite objects, it is likely because they exist at a taxonomic level below that of “class.” The plethora of variable stars, for example, are not classes of objects in this system, on the same level as giant and dwarf stars and so on. Rather, they are types of these stars that could be elaborated in a more complete system.

It is important to emphasize that classification in astronomy has similarities and differences with classification in biology, chemistry and physics. The most obvious difference between the classes (species) in biology and the classes in astronomy, at least as depicted in our Three Kingdom System, is the sheer number of species. E. O. Wilson, the Harvard naturalist who is one of the chroniclers of the diversity of life, has estimated that by 2009, 150 years after Darwin's *Origin of Species*, some 1.8 million species had been discovered and described, out of perhaps tens of millions that now exist. And this does not include what Wilson (in a rare astronomical analogy employed in the domain of biology) calls the “dark matter” of the microscopic universe, which could be tens or hundreds of millions of species of sub-visible organisms.⁴

The number of “species” or classes in astronomy is obviously put to shame by the effusive and creative diversity of biology, no matter how one defines class or what classification system one uses. In terms of number, astronomy's classes, at least as defined in the Three Kingdom System, are more comparable to elements in chemistry (ninety-three natural and fifteen artificial), or to the phyla (thirty-two) and classes (ninety) in just one of Lynn Margulis's five kingdoms (*Animalia*) of biology, which contains almost a million species by itself. Any such comparison depends not only on how one defines a class of astronomical objects, but also whether the classes as defined here in the Three Kingdom System are really analogous to species in the biological hierarchy of classification, or to elements in the linear classification. That is also a matter of definition, and in part a subjective matter based on relation to higher and lower categories in the system. One can argue whether a giant star of Luminosity Class III in the MK system should be called a class or a type, but one cannot argue that a particular member of the class, a type of giant star such as an RR Lyrae, for example, should be placed at a higher level in the system than the class of which it is a member.

This classification exercise also illustrates a problem that astronomical taxonomy has in common with biological taxonomy: classification characteristics do not necessarily conform to evolutionary relationships. The class of giants as defined by the MK system definition was not precisely the same as the class of giants that Henry Norris Russell declared about 1910, nor is it entirely coextensive with the evolutionary states of the giant stars as known today. Russell's definition (and the Mt. Wilson system) was based on size and luminosity, as determined by their distances and apparent magnitudes, which could be converted to luminosity. The MKK definition was based on spectroscopy, in particular “line ratios” defined by standard stars. If an unclassified star matched the standard in a spectroscopic sense, it became a member of that class, such as a giant, without regard to its internal structure or evolutionary status. While luminosities and MK definitions are still used, today astrophysicists often think of giant stars and other stellar classes in terms of their evolutionary state, which for a giant is normally undergoing core helium fusion, but varies depending on the star's mass and where it stands in the spectral temperature sequence. Moreover, a particular class may be adjusted based on new data; in the early 1990s the Hipparcos satellite determined distances ten times more accurate than ground-based parallaxes, and correspondingly more accurate luminosities. The data showed that many of the luminosities were in error, and in the post-Hipparcos, and now the Gaia spacecraft era, the modern concept of a giant star (core helium fusion with shell hydrogen burning via the CNO cycle) is by no means

co-extensive with MK class III defined by spectral line ratios. Nevertheless, the general classes of stars remain, but with a broader definition than determined by the MK system.

In short, astronomical classes have evolved in a way analogous to biology, where “the way it looks” (the phenotype) was primary in the five kingdom classification embraced by zoologists, as opposed to the deeper structure based on genetic makeup (the genotype). But whereas in biology Woese’s “three domain” system caused an uproar in biology with its finding of a completely new domain of life and different relationships for parts of the classification system, the classification of stars by how they physically operate rather than by how they appear has thus far led to broader thinking with only minor adjustments.⁵

4.0 Uses of the system and future development

A good classification system must not only be useful but should also lead to deeper understanding and advance its subject. The uses of the Three Kingdom System are at least threefold, all of which may potentially lead to deeper understanding for different audiences.

First, for scientific purposes, as a comprehensive system for all astronomical objects based on consistent physical principles, the Three Kingdom System brings a consistent set of classification principles to discussions such as the status of Pluto as a planet. It suggests that the definition of a planet should not be based primarily on hydrostatic equilibrium, or roundness, or dynamical considerations, but on physical constitution—just as stellar classification was based on consistent physical principles as determined by spectroscopy. Other criteria may indeed enter any classification decision, but they should be secondary. The Three Kingdom System thus brings consistency to astronomical classification, and more clarity in making classification decisions. In the process it might also, over the longer term, bring consistency to astronomical nomenclature as far as taxa such as class and type are concerned.

Secondly, again for scientific purposes, the symmetric structure of the Three Kingdom System facilitates comparisons at three different scales. In the comparison of families across kingdoms, one can ask, for example, how the interplanetary, interstellar and intergalactic media compare, and analyze what this tells us about the nature of the cosmos. Similarly, for protoplanetary, protostellar and protogalactic processes, and so on. Such comparisons are sometimes already made, but the Three Kingdom System cries out for such comparison in a systematic way. Comparisons of classes across kingdoms may also prove enlightening. Planetary rings, stellar rings and galactic rings in the form of stellar streams have much in common as broken up remains, but at vastly different scales and energies.

Similarly, for planetary, stellar and galactic jets, or subgalactic, substellar and subplanetary objects. However, since the bedrock definition of a class is that at least one representative object must have been observed, we have not included a class of planetary jets, even though the discovery of brown dwarf jets in 2007 led to speculation that planetary jets might exist during the accretion phase of gas giants. Based on symmetry among families in the three kingdoms, we might also predict the existence of such jets, as well as other objects. While some might argue that volcanic eruptions or water spouts from Europa or Enceladus might qualify as jets, this does not seem to me quite analogous to stellar and galactic jets formed by energetic processes. But one could argue.

Thirdly, there is an educational advantage for the teaching of astronomy. The Three Kingdom System allows students to perceive immediately where an object fits in the scheme of astronomical objects. In assessing a new discovery, for example, whether the object is a type, class, family or kingdom should help a student to see its relative importance in the astronomical zoo. Thus, definitive proof of a new kingdom in astronomy would be vastly more important than, say, a new type of subgiant star. Moreover, the decision as to whether a particular class should be placed in a particular family can lead to fruitful discussion among students, and maybe even scientists. For example, the question of whether a globular cluster is circumgalactic or not will lead students to realize that these objects are not found just surrounding the galaxy, but also within the galaxy, and so on.

Finally, as new discoveries are made in astronomy the Three Kingdom System may well be elaborated. For the most part, the additions and revisions will be made at the class and type level, for example, as new classes of planets are discovered, or new classes of baryonic dark matter objects are revealed, or newly detected objects are analyzed such as the mysterious “G objects” at the center of our galaxy that look like gas clouds but behave like stars (W. M. Keck Observatory 2018). It is not out of the question that a new family could be added, though this seems unlikely given our definition of family. At the kingdom level, surprisingly, one can already glimpse a possible new entry: the universe itself may be one of a class of objects in what has been called the multiverse. Because this is a kingdom that, so far, we have not seen, but only inferred from concepts like the anthropic principle, it has not been included in the Three Kingdom System at present. Only time will tell. More fundamentally we must always remember we are classifying baryonic objects composed of protons, electrons and neutrons, and that baryonic matter constitutes only 4.6 % of the matter and energy content of the universe. Non-baryonic dark matter is 23%, and dark energy (believed to be responsible for the accelerating universe) is

72%. But we have no idea what that dark matter and dark energy may be. Classification of the objects that we know notwithstanding, plenty of work remains for future astronomers based on what we do not yet know.

Finally, it is essential to emphasize that because all classes and classification systems are socially constructed, the Three Kingdom System for astronomy is not the only system that could be proposed. But in the end, like the other classification systems, its *raison d'être* and its staying power are dependent on its accuracy, simplicity and utility, both in scientific and pedagogical terms. Such features are an asset for astronomical classes and classification systems in general.

Notes

1. On the “three domain” versus “five kingdom” controversy in biology see especially Sapp (2009). On classification in physics and chemistry see Gordin (2004), Pickering (1984) and Gell-Mann (1994).
2. Davies (2007), especially chapter 4. Isaac Asimov has made the same point in his popular books; for example, Asimov (1992, 263).
3. For more on these classification systems for stars see Dick (2013, chapter 4). A recent popular account of the development of the Harvard system is Sobel (2016).
4. Wilson (2010, xi). In 2011 a group of biologists using a novel analysis estimated 8.7 million eukaryotic species exist, give or take a million. Eukaryotic species contain a nucleus, in contrast to prokaryotes. (Strain 2011).
5. Taxonomy has also evolved, see Mayr (1982, 145), for stages in classification, and microtaxonomy vs macrotaxonomy.

References

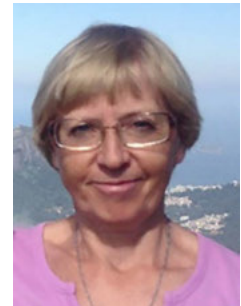
- Asimov, Isaac. 1992. *Atom: Journey Across the Subatomic Cosmos*. New York: Penguin.
- Buta, Ronald J., Harold G. Corwin, Jr., and Stephen C. Odewahn. 2007. *The de Vaucouleurs Atlas of Galaxies*. Cambridge: Cambridge University Press.
- Davies, Paul. 2007. *Cosmic Jackpot: Why Our Universe is Just Right for Life*. Boston: Houghton-Mifflin.
- DeVorkin, David H. 1981. “Community and Spectral Classification in Astrophysics: The Acceptance of E. C. Pickering’s System in 1910.” *Isis* 72: 29-49
- Dick, Steven J. 2013. *Discovery and Classification in Astronomy: Controversy and Consensus*. Cambridge: Cambridge University Press.
- Dick, Steven J. 2018. *Classifying the Cosmos: How We Can Make Sense of the Celestial Landscape*. New York: Springer.
- Feigelson, Eric. 2012. “Classification in Astronomy: Past and Present.” In *Advances in Machine Learning and Data Mining for Astronomy*, ed. Michael J. Way and Jeffrey D. Scargle. London: CRC Press, 3-10.
- Gell-Mann, Murray. 1994. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. New York: W. H. Freeman.
- Gordin, Michael D. 2004. *A Well-Ordered Thing: Dmitrii Mendeleev and the Shadow of the Periodic Table*. New York: Basic Books.
- Gray, Richard O. and Christopher J. Corbally. 2009. *Stellar Spectral Classification*. Princeton: Princeton University Press.
- Mayr, Ernst. 1982. *The Growth of Biological Thought: Diversity, Evolution and Inheritance*. Cambridge, MA: Harvard University Press.
- Mayr, Ernst. 1988. *Toward a New Philosophy of Biology*. Cambridge, MA: Harvard University Press.
- Morgan, William Wilson 1937. “On the Spectral Classification of the Stars of Types A to K”. *Astrophysical Journal* 85, no. 5: 380-97.
- Morgan, William Wilson and P. C. Keenan. 1973. “Spectral Classification,” *Annual Reviews of Astronomy and Astrophysics* 11: 29-50
- Morgan, William Wilson. 1988. “A Morphological Life,” *Annual Reviews of Astronomy and Astrophysics* 26: 1-9.
- Pickering, Andrew. 1984. *Constructing Quarks: A Sociological History of Particle Physics*. Edinburgh: Edinburgh University Press.
- Sandage, Alan. 2005. “The Classification of Galaxies: Early History and Ongoing Developments,” *Annual Reviews of Astronomy and Astrophysics* 43: 581-624.
- Sapp, Jan. 2009. *The New Foundations of Evolution*. Oxford: Oxford University Press.
- Sobel, Dava. 2016. *The Glass Universe: How the Ladies of the Harvard Observatory Took the Measure of the Stars*. New York: Viking.
- Strain, Daniel. 2011. “8.7 Million: A New Estimate for All the Complex Species on Earth.” *Science* 333, no. 6046: 1083.
- W. M. Keck Observatory. 2018, “More Mystery Objects Detected Near Milky Way’s Supermassive Black Hole,” *phys.org* (blog), June 7. <https://phys.org/news/2018-06-mystery-milky-supermassive-black-hole.html>
- Wilson, Edward Osborne. 2010. Foreword to *Kingdoms and Domains: An Illustrated Guide to the Phyla on Earth* by Lynn Margulis and Michael Chapman. Amsterdam: Elsevier, lxi-lxii.

Soil Classification*

Aleksandra A. Nikiforova

Lomonosov Moscow State University, Faculty of Soil Science,
1-12 Leninskie Gory, 119991 Moscow, Russian Federation,
<nikifsoil@gmail.com>

Aleksandra Aleksandrovna Nikiforova is a research fellow at Faculty of Soil Science, Lomonosov Moscow State University (MSU), Russia. She received her PhD in soil mapping at the MSU and a specialist degree at the Department of Soil Geography of the MSU. Her main research interests are related to the theory of universal soil-landscape classification and multiscale soil-landscape GIS mapping, as well as its practical implementation. One of her significant achievements is the creation, together with other authors, of *The Agroecological Soil-Reclamation map of the Nonchernozem belt of the European part of Russia* on a scale of 1: 1,500,000.



Nikiforova, Aleksandra A. 2019. "Soil Classification." *Knowledge Organization* 46(4): 467-488. 127 references. DOI:10.5771/0943-7444-2019-6-467.

Abstract: Soil classification is a long-debated issue. Despite the accumulation of empirical data and appearance of modern computer technologies, soil classification problems remain unresolved and relevant for discussion. The main problem is the creation of a universal soil classification system. The causes of soil classification problems are analyzed and a solution based on contemporary theories of classification and the general systems theory (open system) approach is presented. I discuss the purposes and the current state of soil classification, as well as unresolved issues such as: what definition of soils should be the basis for a universal soil classification system, should soil classification systems be genetic or morphological, how to make them evolutionary, and others. The common features of officially recognized national and international soil classification systems and some underdeveloped ones are reviewed, as well as those in which they differ from each other. It is shown that the shortcomings of soil classification systems are largely related to neglecting the essential character of soils, namely, its dual systemic nature to be not only an independent natural body (that is, a system), but also the result of interaction and interrelation of soil-forming factors (that is, an element of the system), ignoring the rules for logical division of concepts and replacing the differentiating criteria with diagnostic criteria. The theoretical basis and advantages of the "soil-landscape classification system" being developed by the author are outlined. To solve soil classification problems, an outside perspective is needed, that is, the use of classiology and the systems approach.

Received: 27 February 2019; Accepted: 11 April 2019; Revised 24 June 2019

Keywords: soil classifications, soils, natural science

* I would like to express my deep gratitude to professor Birger Hjørland, as well as to anonymous reviewers for their valuable and constructive suggestions for editing this article. I am also very grateful to Mr. Mikhail M. Borisov for creating the first online version of the "soil-landscape classification system" and to my son Fyodor Nikiforov for his great help in preparing the figures for the article. Derived from the article of similar title in the ISKO Encyclopedia of Knowledge Organization, Version 1.0, published 2019-02-26, last edited 2019-03-19 Article category: KO in specific domains

1.0. Introduction

Soil classification is a long-debated issue. The first scientific version of the soil classification system developed by Vasily Dokuchaev was published in 1886. In 1962, Muir assessed the situation with regard to soil classification as follows: "soil classification is still in an elementary stage of development." Much time has passed since then, but one cannot say that the situation has changed fundamentally. Despite the fact that during this time a lot of empirical data accumulated and modern computer technologies appeared and were used, the transition to a qualitatively new stage in the development of soil classification did not happen. Soil classification problems remain the same; they are still unresolved and relevant for discussion. The main

problem is the creation of a universal (that is, basic, unified, global, generally accepted) soil classification system (soil classification system hereinafter referred to as SCS). This was confirmed in 2010 at the 19th World Congress of Soil Science, where, by decision of the Council of the International Union of Soil Science (IUSS), a Working Group on Universal Soil Classification was officially established (https://www.iuss.org/index.php?article_id=525). However, a universal SCS has not yet been created.

In the present text I use contemporary theories of classification (classiology¹) (Frické 2016; Hjørland 2017; Mill 1882; Parrochia 2017; Parrochia and Neuville 2013; Pokrovsky 2014; Rozhkov 2012; Rozova 1986; Subbotin 2001) and the general systems theory (open system) approach (or, if shortened, the systems approach²), developed

by Bertalanffy (1968). Examples of attempts to apply the systems approach in natural science can be found in the works of Chorley and Kennedy (1971), Juma (1999), Karpachevsky (1981, 240-245), Phillips (1998), and Solntsev (1981). Such an interdisciplinary approach allowed us to see the shortcomings of the existing SCSs from an outside perspective and to propose the creation of a new, fundamentally different SCS, namely the “soil-landscape classification system” (hereinafter referred to as SLCS).

Officially recognized national and international and some so-called “underdeveloped” SCSs (or schemes) (Krasilnikov and Arnold 2009, 319) proposed by various soil scientists, which are interesting from a scientific point of view are discussed here. However, I do not cover the history of the development of these systems, since its detailed description can be found in many works, including the work of Krasilnikov and Arnold (45-335). At the same time, the following SCSs are not the subject of discussion: 1) narrow-focused, simple characterizations in which soils are divided using one criteria, for example, land use type, topography, age, parent materials, or color (Hartemink 2015, 131); 2) engineering (technical) used for practical purposes; 3) outdated and extinct; 4) folk; and, 5) numerical (quantitative) based on mathematical and statistical methods (pedometric approaches) (Hole and Hironaka 1960; Hughes et al. 2014; McBratney and Gruijter 1992; Parrochia 2017; Rayner 1966; Rozhkov 2011; Verheyen et al. 2001). Numerical SCSs are excluded from consideration for the following reasons. First, they relate to a large, specific, and independent section of soil classification and, therefore, require special attention. Secondly, the problem of scientific classification cannot be solved with the help of mathematical methods, since first of all the theory of classification should be developed (Rozova 1986, 196). Moreover, the development of a numerical soil classification is considered as an “auxiliary” and/or additional task in relation to the problem of developing a genetic soil classification (Sokolov 2004, 185). However, what purposes are set for soil classification?

2.0. Purposes of soil classification

There are scientific (theoretical, fundamental) and practical (applied) purposes of soil classification. For example, Arnold (2002) states: “Applied uses and scientific knowledge have both been major purposes of soil classification.” The most frequently mentioned scientific purposes are:

- Providing a common scientific language to facilitate the comparison, exchange, and extrapolation of soil information, results, and experience on agricultural and environmental issues among scientists by correlating and harmonizing officially recognized SCSs and unification of soil nomenclature (Brevik et al. 2016; Cline 1949; De

Bakker 1970; Hempel et al. 2013; IUSS Working Group WRB 2015, 5; Láng et al. 2013).

- Improving the scientific understanding of the genesis of soils by reflecting the relationship between soils and the environment (Beckmann 1984; Cline 1949; De Bakker 1970; Hartemink 2015; Kubična 1958; Muir 1962; Riecken 1963; Sokal 1974; Zonneveld 1959, quote according to De Bakker 1970).
- Identification and reflection of the main stages of soil formation and, on this basis, prediction of their behavior under different uses and management (Beckmann 1984; Kellogg 1963; Kovda 1973, 377-428).
- Discovery, display, and explanation of the basic laws of soil formation (Basinski 1959; Smith 1965; Sokolov 2004, 170).
- Defining the soil science paradigm to indicate the path to the future development and progress of soil science (Ibáñez and Boixadera 2002; Kiryushin 2011, 8).
- Providing the basis for developing soil map legends (De Bakker 1970; IUSS Working Group WRB 2015, 12-21; Rozova 1986, 67).
- Unification of diagnostic methods and development of a methodology for the identification of soils (Tyurin 1957, quote according to Basinski 1959).

Nevertheless, in spite of these scientific purposes, the overwhelming majority of officially recognized SCSs in development had mainly practical purposes, first of all, supporting soil surveys (mapping) (Arnold 2002; Baruck et al. 2016). For practical purposes, the already developed SCSs are used for inventory of soils and solving applied problems in agriculture, land use, engineering, and environmental surveys, construction, operation of roads, underground utilities, in the fields of geology, hydrology, forestry, etc. (De Bakker 1970; Riecken 1963; Sokolov 2004, 171). Before proceeding to the analysis of the unresolved issues of soil classification, discussed in soil science, let us dwell on the current state of soil classification.

3.0. The current state of soil classification

In 2001, Langohr characterized the state of soil classification as follows: it has a poor reputation and is often called useless because of too many classification systems changing too often, containing too many characteristics, requiring too complex data, having too complicated terminology, and not having common accurate soil names. In 2012, Rozhkov drew attention to the weak theoretical base of SCSs: “The existing soil classification systems do not completely satisfy the principles of classiology. The violation of logical basis, poor structuring, low integrity, and inadequate level of formalization make these systems verbal schemes rather than classification systems *sensu stricto*.”

It can be said that little has changed since then. At the same time, it is encouraging that there is a growing desire to change the situation and make progress in soil classification (Brevik et al. 2016; Hempel et al. 2013; Ibáñez and Boixadera 2002; Nikiforova and Fleis 2018; Sokolov 2004, 170). The priority is to create a universal SCS, which is considered a challenge due to the continuous nature, extreme complexity and high spatial variability of soils, as well as due to the wide variety of soil-forming factors and incomplete soil data (Fridland 1986, 9; Heuvelink and Webster 2001; Ibáñez and Boixadera 2002).

The continuous nature and transitional forms of soils and the impossibility of unambiguously attributing them to a particular class (Rozova 1986, 97) create difficulties in soil classification. It should be said that this problem is acute not only in soil science, but also, for example, in geobotany and petrography (96). There are different points of view on how to solve this dilemma, but we will focus on the philosophical ones. To begin with, according to Rozova (95-96), “formal-logical criteria for good classification require a clear definition of the boundaries between classes of objects” and “if this condition cannot be met, the classification procedure cannot be implemented either.” At the same time, philosophy does not give a definite answer how to establish these boundaries. On the one hand, in order to do this, it is proposed to get rid of intermediate classes in the process of building a classification system, thereby ignoring the transitional forms of objects (97). On the other hand, “in a situation when it is necessary to theoretically ‘grasp’ the development of an ‘object,’ the transitional forms should be separated into special independent classes.” In general, philosophers conclude that the concept of a continuum of objects does not deny the possibility of their classification (see Rozova 1986, 156).

The situation is complicated by a misunderstanding of what classification is and what is the difference between the terms classification, classifying, and classification system (Ibáñez and Boixadera 2002; Rozova 1986, 194; Sokal 1974). For example, classification is often confused with classifying and is understood as the allocation of soils in accordance with a specific classification system (IUSS Working Group WRB 2015, 13). It can also be understood as the arranging of soils into classes for a specific (scientific, environmental, engineering, agronomic) purpose (Jones et al. 2005, 25) and as combining soils with similar properties into groups (Nagy et al. 2016). In this article, the term classification refers to the logical division of a set into subsets (disjoined classes and subclasses), whereas the term classifying to the identification of objects in accordance with the already developed classification systems. An example of such an understanding of classification and classifying in soil science is the statement by Sokolov (2004, 177). He stresses that there is a need of “a clear understanding of the differences between

the development of classification and the identification of objects in accordance with the classification system already prepared.” In addition, the terms classification and classification system, on the one hand, and mapping, zoning and map legend, on the other hand, are also often misunderstood and are used as synonyms (Krasilnikov, Martí, Arnold and Shoba 2009, 3; Narayanan et al. 1992), although the former are the basis for the latter (Avery 1973; Buol et al. 1980, 343; 345; De Bakker 1970; Rozova 1986, 67; Schelling 1970; Subbotin 2001, 59).

Soil classification problems are mainly associated with a lack of theoretical justification (Ibáñez and Boixadera 2002; Rozova 1986, 196; Sokolov 2004, 165). However, most current publications on soil classification are devoted to other topics, namely:

- History of creation and description of SCSs (Anderson and Smith 2011; Gennadiyev and Gerasimova 1996; Isbell 1992; Krasilnikov, Martí, Arnold and Shoba 2009; Paton and Humphreys 2007; Simonson 1989).
- Comparison and correlation (or harmonization) of SCSs (Hughes et al. 2017; 2018; Gerasimova and Khitrov 2012; Krasilnikov, Martí, Arnold and Shoba 2009; Lebedeva et al. 1999; Mazhitova et al. 1994; Michéli 2008; Murashkina et al. 2005; Shi et al. 2010; Shoba 2002; Zádorová and Penížek 2011).
- Diagnostics and classifying of soils (Deressa et al. 2018; Gobin et al. 2000; Lebedeva and Gerasimova 2012).
- Technological and mathematical (statistical) methods of classification and classifying of soils, as well as of revising, updating, and improving of the current SCSs (Da Silva et al. 2014; Hartemink and Minasny 2014; Hempel et al. 2013; Nagy et al. 2016; Ogen et al. 2017; Teng et al. 2018; Vasques et al. 2014).

4.0 Unresolved issues of soil classification

In this section, the following unresolved issues of soil classification, which are discussed in soil science, are covered: is a single universal SCS required, what definition of soils should be the basis for a universal SCS, what is the basic unit (minimal object with homogeneous properties) of soil classification (hereinafter referred to as BUSC), should SCSs be genetic or morphological, what is genetic soil classification, what type and method of constructing SCSs is more fruitful, and how to make SCSs evolutionary.

4.1 Is a single universal soilclassification system required?

There are two main points of view on the need for a single universal SCS. The first point of view is that there should be different SCSs for different purposes (Cline 1962; Ibáñez

and Boixadera 2002), and the second point of view is that there should be one SCS, which serves as the basis for practical (applied) SCSs (Fridland 1986, 6; Sokolov 2004, 166). In defense of the second point of view, Fridland (1986, 6) emphasizes that a basic SCS should provide the basis for the integrity of soil science through a common language and the most effective use of results of soil science in other sciences and in practice. At the same time, Rozova (1986, 215), based on her philosophical position, believes that in the future, a unified system of a fundamental and applied nature may emerge.

4.2 What definition of soils should be the basis for their classification?

It is widely recognized that classification without a precise definition of its object is impossible. Dokuchaev, commonly regarded as the founder of soil science, gave the first scientific definition of soils. However, as evidenced by the constant appearance of old and new definitions, there is the need to improve Dokuchaev's definition or give another one. For example, in a recent study, Hartemink (2015) analyzes eighty-one definitions of soils and suggests another. However, these new definitions, in contrast to Dokuchaev's one, for the most part only list the diagnostic properties of soils, leaving their essential character without proper attention. Their essential nature lies in their duality—on the one hand, the soils are independent natural bodies (that is, systems), and on the other hand they are the result of the interaction and interrelationship of soil-forming factors (that is, elements of systems). As an example, we give the definitions of soils presented in the explanatory notes to the “world reference base for soil resources” (WRB), the U.S. soil taxonomy and Russian soil classification system:

- For WRB, soil is: any material within two meters of the Earth's surface that is in contact with the atmosphere, excluding living organisms, areas with continuous ice not covered by other material, and bodies of water deeper than two meters. The definition includes continuous rock, paved urban soils, soils of industrial areas, cave soils as well as subaqueous soils. Soils under continuous rock, except those that occur in caves, are generally not considered for classification. In special cases, the WRB may be used to classify soils under rock, for example for palaeopedological reconstruction of the environment. (IUSS Working Group WRB 2015, 4)
- For “U.S. Soil Taxonomy” (Soil Survey Staff 1999): “Soil ... is a natural body comprised of solids (minerals and organic matter), liquid, and gases that occurs on the land surface, occupies space, and is characterized by one or both of the following: horizons, or layers, that are

distinguishable from the initial material as a result of additions, losses, transfers, and transformations of energy and matter or the ability to support rooted plants in a natural environment.”

- For the Russian soil classification system (Shishov et al. 2004): “The soil is a natural or natural-anthropogenic solid-phase body, exposed on the land surface, formed as a result of long-term interaction of the processes leading to the differentiation of the original mineral and organic material into horizons.”

It should be recalled that there are two main versions of the definition of soils proposed by Dokuchaev, which have a similar first part (namely, the soil is an independent natural body) and are distinguished by their second part. These two parts of the definition reflect the dual nature of the soil. The second part of the well-known first version, which is commonly used every day: “Each soil is the product of the aggregate activity of parent material, climate, vegetation, and topography” (Dokuchaev 1879, 1). This second part of the first version varies in other works of Dokuchaev, since he returned to it many times over many years. Much more rarely is this second part used, namely: soils are “those daily or outward horizons of rocks ... which are more or less changed naturally by the common effect of water, air and various kinds of living and dead organisms” (Dokuchaev 1886, 227). A comparison of these second parts of the two versions of the definition shows that, unlike the second part of the first version, the second part of the second version corresponds to the systems approach, despite the fact that it does not use its terminology (Nikiforova and Fleis 2018). However, at present, only the first part of Dokuchaev's definition is used as the basis for soil classification, while the second is either not used at all, or its use is only declared (see Buol et al. 1980, 17; 320; Florea 2012; Lebedeva and Gerasimova 2009). For example, Jenny (1941, 1-21) draws attention to the fact that most soil scientists deal only with the soil as such (that is, with the soil as an independent natural body), but not with the soil as a part of a wider system, namely the natural landscape or the environment,” however, “often it is not sufficiently realized that the boundary between soil and environment is artificial.” In turn, Karpachevsky (1981) expresses the following view: “An analysis of the soil definition given by V.V. Dokuchaev shows that although soil is a special natural body ... it should always be considered as a subsystem of the other natural systems. There is no soil out of these systems. This provision, explicitly or implicitly, normally provides the foundation of all scientific researches of soils.” Fridland (1986, 9) considers the relationship of soils with soil formation factors to be their main property. However, it is the second version of the definition of Dokuchaev, which is currently used to

study the landscape in Russia. Moreover, mainly because of this version, Dokuchaev is considered to be the founder of Russian landscape science, despite the fact that he never used the term landscape in his works.

The definition of soils affects the set of objects that are proposed for inclusion in SCSs. For example, in addition to natural terrestrial soils, it is proposed to include in SCSs: 1) regolith and groundwater, which together with the soil form an integrated natural body that supports life on Earth (see Krasilnikov and Arnold 2009, 329); 2) superficial friable rocks, redeposited and artificially accumulated soils, as well as underwater bottom formations located at a shallow depth and serving as a substrate for green plants (Fridland 1986, 8-9); and, 3) all exogenous bodies characterized by fertility, since they are genetically related to soils by gradual transitions, perform biospheric ecological functions of soils, and are objects of economic activity, cartography, and accounting (Sokolov 1991).

4.3. What is the basic unit of soil classification?

The definition of soils and BUSCs are usually considered as different tasks. The following are often referred to as such BUSCs: prisms of a certain section, soil individuals, pedons, polypedons, soil profiles, solums, three-dimensional natural bodies, etc. Moreover, each of these BUSCs can have different content (AFES 1998; Avery 1973; Buol et al. 1980, 17; Ibáñez and Boixadera 2002; Krasilnikov, Martí and Arnold 2009, 16; Sokolov 1978; 2004, 175). Fridland (1986, 9) names the following requirements to BUSCs: they must: 1) not depend on any classification system; 2) be sufficiently homogeneous, indivisible within classification (this should be controlled by the disappearance of their connection with the soil-forming factors); and, 3) be three-dimensional bodies. At the same time, according to Sokolov (2004, 176), the declared BUSCs should not affect the result of the classification process and the true BUSCs are, as a rule, soil images and natural laws of soil formation.

4.4 Should soil classification systems be genetic or morphological?

To begin with, in philosophy and science, including soil science, there is no generally accepted concept of genetic classification, and the term genetic classification can be understood differently (Krasilnikov, Martí and Arnold 2009, 11; Rozova 1986, 59). As a result, in soil science, the concepts of genetic and morphological classifications are often replaced by each other (Nikiforova et al. 2019). Therefore, we need to clarify what this term means. In soil science, the term genetic classification refers to a classification in which modern soils are divided according to soil

formation conditions (or soil-forming factors), which determine the genesis and properties of soils. At the same time, the term genetic classification system refers to a classification system that reflects these soil formation conditions.

There are two main opposite approaches to soil classification: morphological, that is, focused on the diagnostic properties of soils and, above all, diagnostic horizons (Bridges 1990), and genetic, of which the former became dominant (Hartemink 2015). On the one hand, compared to morphological ones, genetic classification systems provide a deeper understanding of the genesis of the classification objects and a forecast of possible changes in them (Dupré 2006, 31, quote according to Hjørland 2017, 108). Here is what Kubiěna writes in this connection: “the knowledge of the genesis of a property is very important in systematics since only by this can a property or a unit of properties be fully known and understood ... describing things in nature without any efforts to understand them means only a beginning of science, not science itself” (Kubiěna 1958). On the other hand, it is widely believed that the soil genesis can be reflected in SCSs both directly (through the soil-forming factors or landscape features) and indirectly, in a “hidden” form (through the diagnostic soil properties) (Basinski 1959; Smith 1983; Lebedeva and Gerasimova 2009; Rozanov 1982). Moreover, indirect reflection is usually considered more correct due to the widespread notion that soils should be classified as such, regardless of the soil-forming factors, that is, in the same way as other natural objects (IUSS Working Group WRB 2015, 4; Leeper 1952); otherwise, instead of soils, more general concepts, such as landscapes, geobiospheres, and ecosystems, will be classified (Beckmann 1984; Sokolov 1978).

As another argument against genetic soil classification, the fact is advanced that soils reflect not only current soil formation conditions, but also past ones, due to which the dependence of soil properties on soil-forming factors is not always linear (Krasilnikov, Martí and Arnold 2009, 10; Phillips 1996; Targulian and Goryachkin 2004). Finally, the genesis of soils is considered to be based on implicit knowledge and, therefore, a shaky basis for soil classification (Nachtergaele et al. 2002). As a result, today SCSs based on grouping soil profiles as combinations of diagnostic horizons are considered as genetic (Krasilnikov, Martí and Arnold 2009, 11-12).

The refusal to include soil formation conditions in a SCS is connected, in our opinion, with the unwillingness to mix genetic (landscape) features with the diagnostic properties of soils as independent natural bodies, which is quite understandable. However, such a “mixing” simply will not happen if we use genetic (landscape) features as differentiating criteria, and diagnostic properties as diag-

nostic ones. The fact is that differentiating criteria are considered to be essential (internal) properties of objects, “which are causes of many other properties; or, at any rate, which are sure marks of them” (Mill 1882, 872), whereas diagnostic ones are considered to be formal (external), in many cases, morphological properties of objects of classification, which are determined by differentiating criteria. Properties without any content of essential character cannot be considered as differentiating (Muir 1962). In addition, differentiating criteria serve to divide objects into classes and subclasses, whereas diagnostic ones serve to identify them (Rozova 1986, 18, 25, 95; Subbotin 2001, 28-29, 55-57).

It should be emphasized that, following Dokuchaev, many soil scientists were in favor of including genesis in SCSs (Basher 1997; Basinski 1959; Cline 1962; Dobrovolskii 2005; Florea 2012; Juilleret et al. 2016; Knox 1965; Smith 1983; Sokolov 1991). Explaining this position, Florea (2012) stresses that genesis “helps to the understanding of the soil cover in landscape, contributing to a more efficient and of high quality soil survey.” In 1965, Knox dreamed of a SCS based on some kind of soil-landscape units. Another weighty argument for including genesis in SCSs is that modern society needs more and more information about the environment, including information on landscape features (Krasilnikov and Arnold 2009, 329). Therefore, the advantage of many morphological SCSs is that they already contain landscape features, however, not on a systematic basis. See, for example, the “U.S. Soil Taxonomy” (Bockheim et al. 2014; Smith 1986; Soil Survey Staff 1999).

4.5 What type and method of constructing soil classification systems is more fruitful?

Another unresolved issue is the type (hierarchical, non-hierarchical) and method of constructing SCSs. On the one hand, it is stated (Nachtergaele et al. 2002): “[R]igid hierarchic ranking may result in a false sense of correctness not suited for many of the soil studies undertaken and often leading to a loss of soil information.” It is also believed that hierarchical systems are “subjective, expert-dependent structures, which facilitate the search and recall of objects within the system rather than being a reflection of any real organization of entities into natural groups” (see Krasilnikov, Martí and Arnold 2009, 11). On the other hand, hierarchical structures are considered irreplaceable because they “optimize the flow of information” (Ibáñez and Boixadera 2002), may constitute a system of objective laws of soil formation reflecting their subordination (Sokolov 1991), and help “to more holistically combine soil formation factors with soil geography and pattern” (Miller and Schaeztl 2016).

As for the method of constructing SCSs, it is believed there are two ways: descending (top-down, segregating, analytic, and usually genetic) and ascending (bottom-up, aggregating, synthetic, non-genetic) (Arnold 2002; Manil 1959; Muir 1962). Arnold (2002) considers that it is possible to use both methods. He writes: “it is possible to start with the domain and divide it and subdivide it and so on” and “it is also possible to group the individuals, then group the groups, and so on.” However, there is also another point of view. For example, Sokolov (2004, 176) states that “if we set ourselves the purpose of creating a classification that would be a synthesis of our knowledge of soils and reflect the basic laws of soil formation, then it can only be built as top-down.”

4.6 How to make soil classification systems evolutionary?

There is still no clear answer to the question of how to make SCSs evolutionary³ (dynamic, non-static), although the need to resolve this issue is recognized (Basinski 1959; Pokrovsky 2014; Rozanov 1977, 4; 1982; Schelling 1970). For example, according to Schelling (1970), we currently classify “merely momentary glimpses” of soils, which is not enough. To make SCSs evolutionary, Manil (1959) proposes including paleopedological characteristics at the lower categories of SCSs, and Kovda (1973, 377-428) proposes using the soil age and the stages of soil development as criteria for soil division. Mamai (2005) believes that statistical and dynamic classifications together should constitute one system. Finally, according to the philosopher Subbotin (2001, 61), for the classification system to be evolutionary, it must have a time axis of coordinates.

5.0 Officially recognized national and international and some underdeveloped soil classification systems

In the introductory part of this section, it should be said that SCSs are the result of the consensus among experts and, therefore, they are “closed systems,” which are developed, adopted, and changed by the institutions responsible for soil classification and/or soil mapping (Krasilnikov, Martí, Arnold and Shoba 2009, 35). This distinguishes them from biological systems that are “open and grow continuously over time with the inputs of the whole scientific community involved in the detection of new taxa.” It should also be borne in mind that, as is in all other sciences, in soil science, classification organizes the knowledge accumulated at the moment (Smith 1965). This means that it develops and improves with the development of soil science, with the expansion and deepening of knowledge about soils (Rozova 1986, 51).

Some structure features of officially recognized national and international SCSs, as well as some underdeveloped SCSs, namely taxonomic levels (or, in accordance with contemporary theories of classification, degrees or orders; the same applies to hierarchical levels), levels of archetypes and criteria for division of soils are shown in Figures 1 and 2, respectively. In more detail, features of these SCSs are described below.

5.1 Common features

In general, SCSs are similar and not fundamentally different from each other. They are characterized by the presence of taxonomic levels (including levels of archetypes), confusion between differentiating and diagnostic criteria, the lack of objective rules for the selection and ranking of criteria for division of soils, as well as violation of the rules for logical division of concepts.

5.1.1 Presence of taxonomic levels, including levels of archetypes

According to Shreyder (1983) and Krasilnikov, Martí and Arnold (2009, 5-30), almost all SCSs are intuitively based on the concept of archetypes, that is, original central images or concepts, prototypes of soils. Many of the archetypes existed before the advent of modern scientific classification systems. For example, Krasilnikov, Martí and Arnold (2009, 18) explain the meaning of this term in soil classification as follows:

Most natural classifications grew from pre-scientific ones, mostly non-verbal concepts of archetypes ... At the initial stage of the development of modern soil classification, soil types in the sense of V.V. Dokuchaev and his successors were archetypes. The names of soil types were mainly borrowed from folk soil classifications: the words *chernozem*, *solod*, *solonetz*, *rhendzina* were used by Russian, Ukrainian and Polish peasants for ages. The use of indigenous soil names reinforced the use of the archetypes in scientific soil classifications.

Currently in soil science, archetypes are considered the basic taxonomic units, represented mainly by soil types, as well as series and reference groups. Archetypes are characterized by sets of features. For example, in the Russian school of soil science, soils of the same genetic type are similar in: 1) input of organic substances and their transformation and decomposition; 2) decomposition of the mineral mass and synthesis of mineral and organo-mineral neoformations; 3) migration and accumulation of substances; 4) soil profile structure; and, 5) measures to im-

prove and maintain soil fertility (Rode 1975, 254). In addition, archetypes form the initial basic taxonomic levels of SCSs after which they are grouped and/or divided, forming higher and lower levels. Usually in SCSs, there is only one level of archetypes; two levels (one for landscapes and one for soil profiles) are present in SCSs using the concept of soil series related to landscapes and parent materials (Krasilnikov, Martí and Arnold 2009, 24-26). Thus, the creation of SCSs does not begin from the zero-level represented by the initial set (universe) of soils but from the archetype levels and then continues in an upward and/or downward direction. Most of the officially recognized SCSs have levels obtained because of dividing archetypes and their grouping. For example, in the Russian SCS (Shishov et al. 2004) and the "U.S. Soil Taxonomy" (Soil Survey Staff 1999), the upper hierarchical levels (sections, trunks and orders, sub-orders, respectively) are built by grouping archetypes represented by soil types and great groups of soils, respectively, whereas lower levels (subtypes, genres and subgroups, families, etc.) are built by dividing archetypes. However, this method of building SCSs contradicts the concept of a hierarchical classification system. Therefore, SCSs having levels of archetypes can be called "pseudo-hierarchical," since they only seem to be hierarchical, but, in fact, they are not.

It should also be said that some SCSs, for example, the French SCS (AFES 1998) and the WRB (IUSS Working Group WRB 2015) are considered reference databases without any or little hierarchy (Krasilnikov, Martí and Arnold 2009, 41). There are also classification systems created in the form of tables, but this can be considered an exception to the general rule. An example is the SCS of the Republic of South Africa (Soil Classification Working Group 1977).

5.1.2 Confusion between differentiating and diagnostic criteria

Differentiating criteria are among the most important classification elements that determine the success of the development and operation of a natural classification system (Subbotin 2001, 29) and ultimately its scientific character (Mill 1882, 872). Differentiating criteria are used for classification of objects, and diagnostic criteria are used for their identification (classifying). However, in existing SCSs, diagnostic criteria or a mixture of differentiating and diagnostic criteria replace differentiating ones. Moreover, in most cases, in officially recognized SCSs, diagnostic criteria play a leading role in this mixture, whereas in underdeveloped SCSs, the opposite is often the case. As a result, these SCSs are artificial rather than natural and genetic, and do not solve most of the scientific problems facing soil classification.

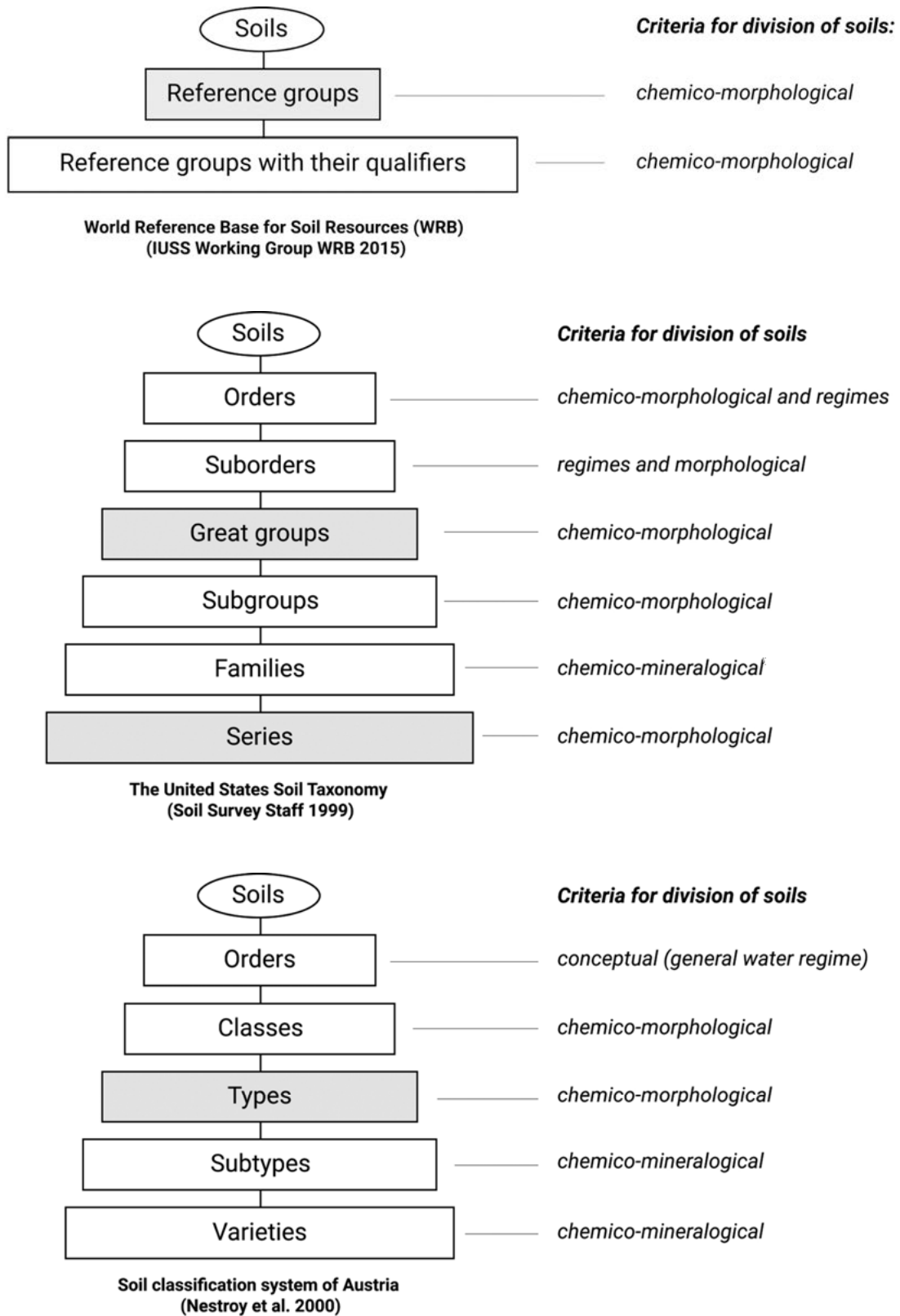


Figure 1. Structure features of some officially recognized national and international SCSs. Note: the book by Krasilnikov, Martí, Arnold and Shoba (2009) was used in the preparation of the figure.

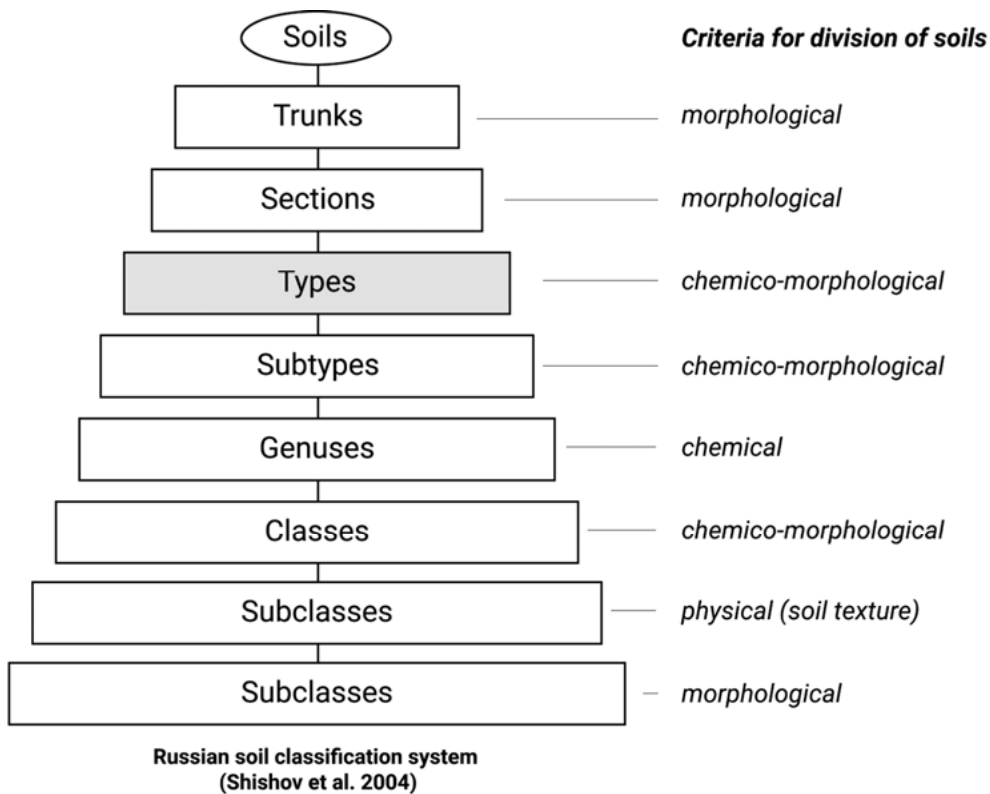
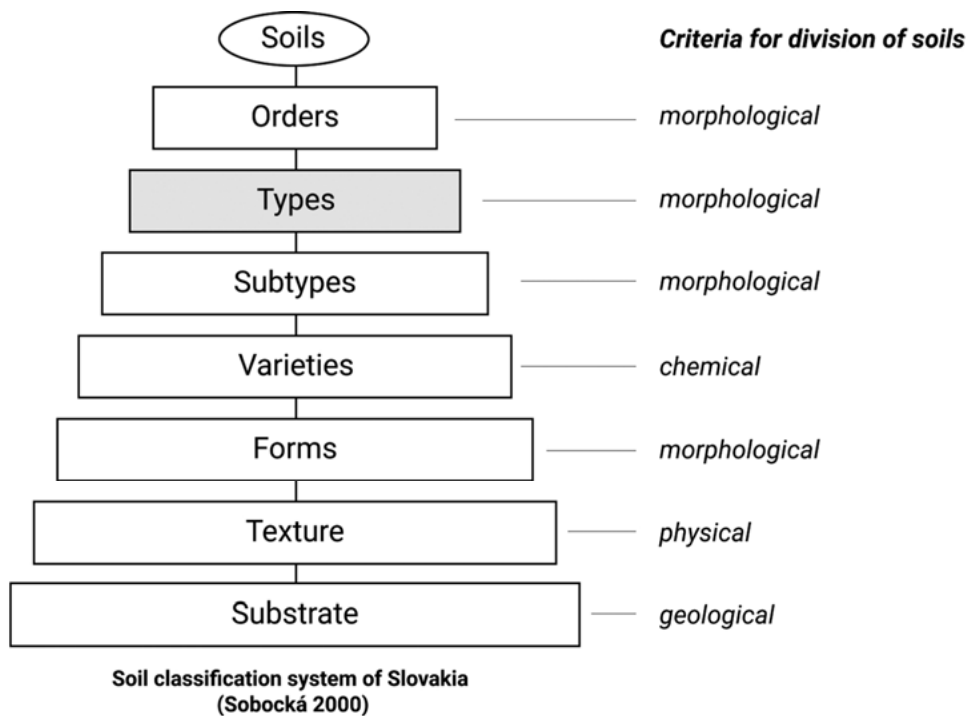
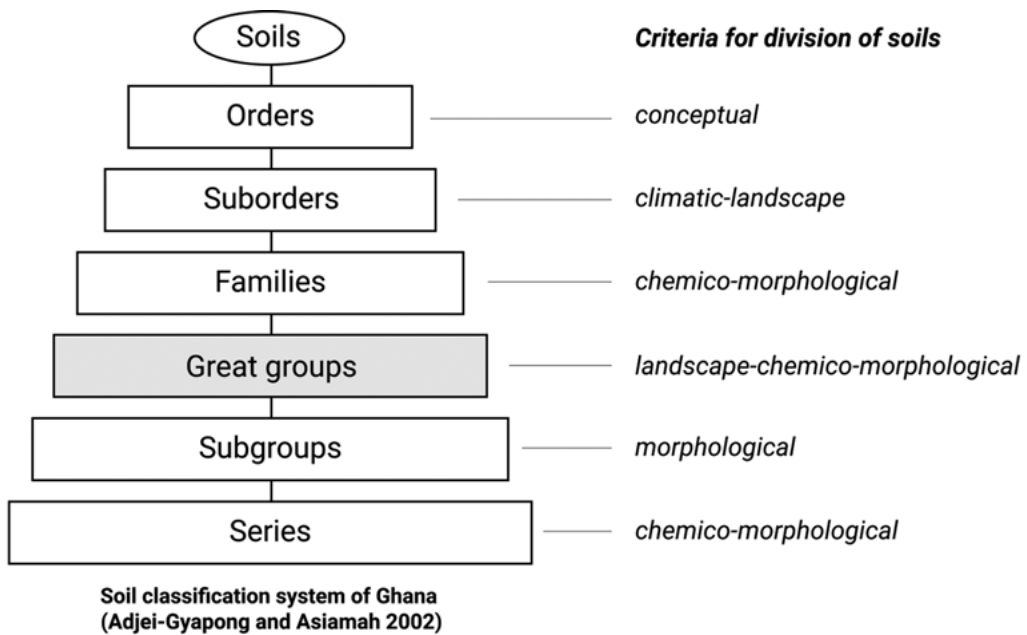
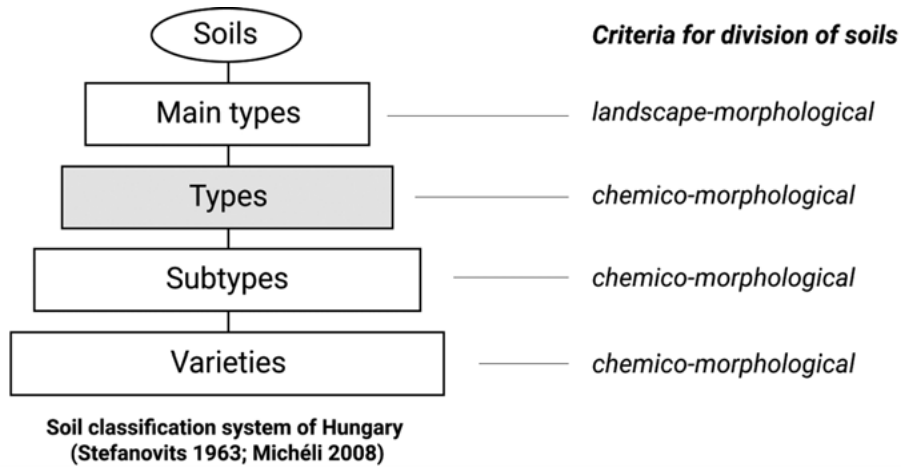


Figure 1 (cont.)




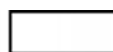

 — universe of soils
  — taxonomic levels
  — levels of archetypes

Figure 1 (cont.)

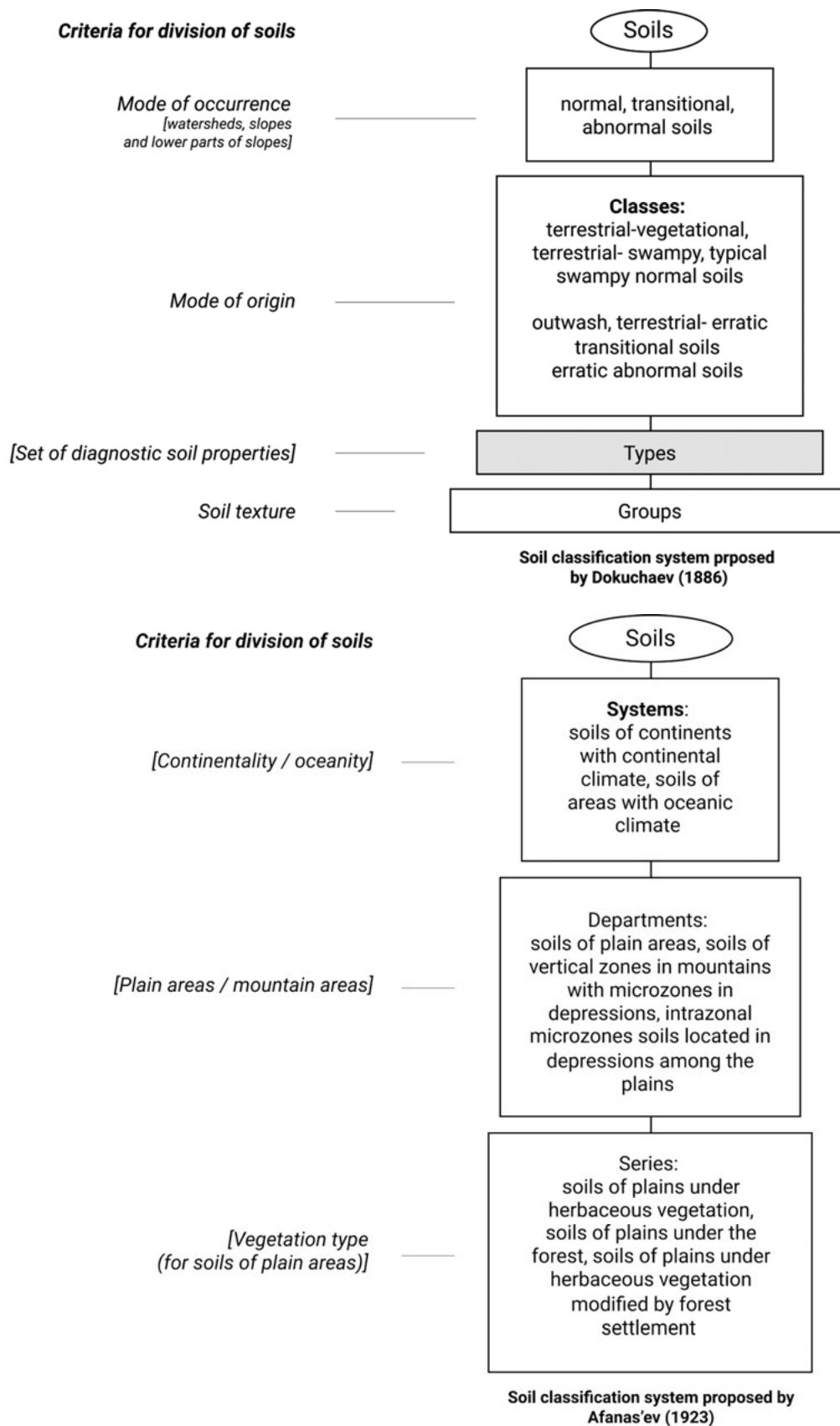


Figure 2. Structure features of some underdeveloped SCSs. Note: Criteria for division of soils enclosed in square brackets are not directly named as such by the authors of SCSs but extracted from the explanatory notes to these SCSs.

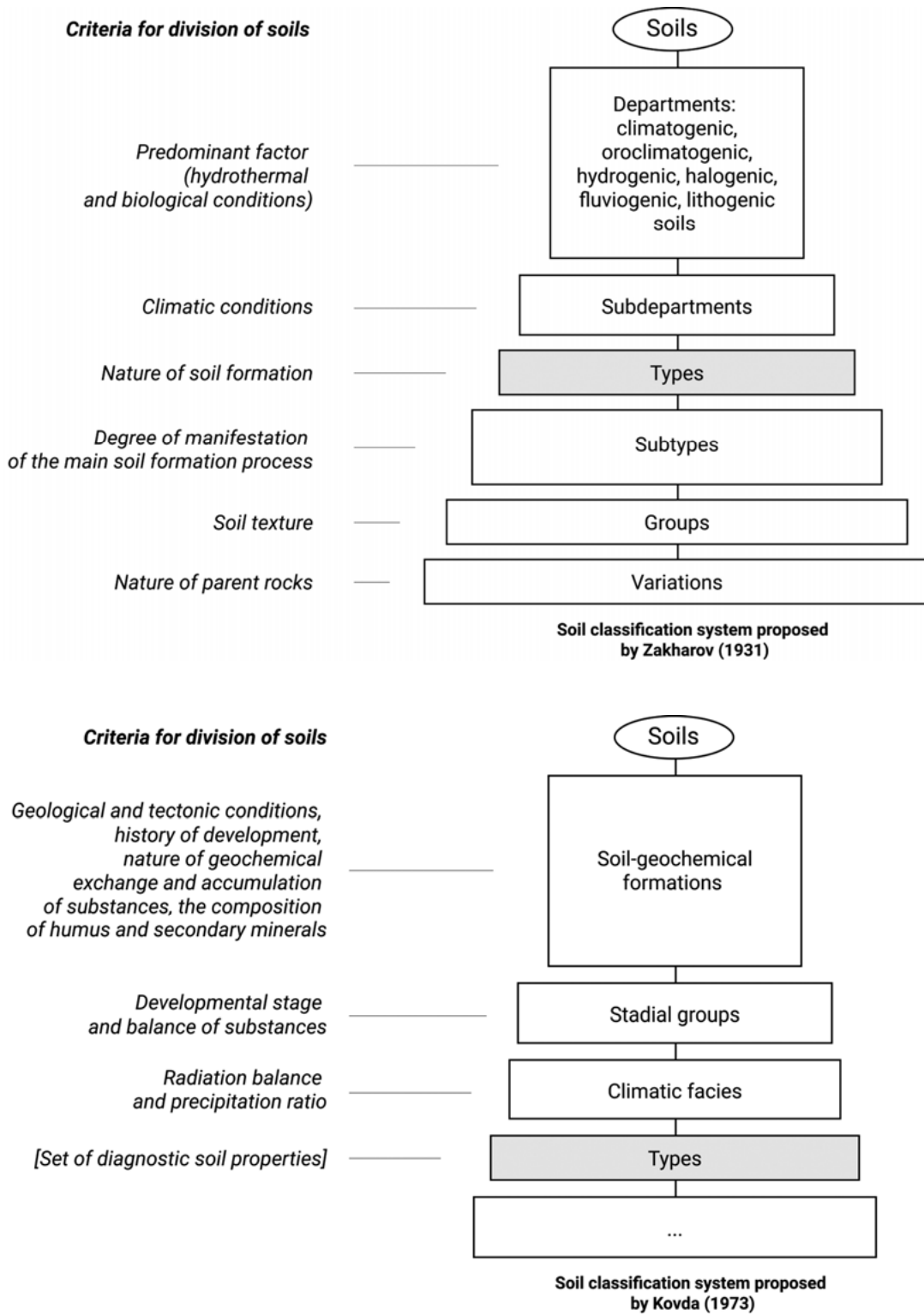


Figure 2 (cont.)

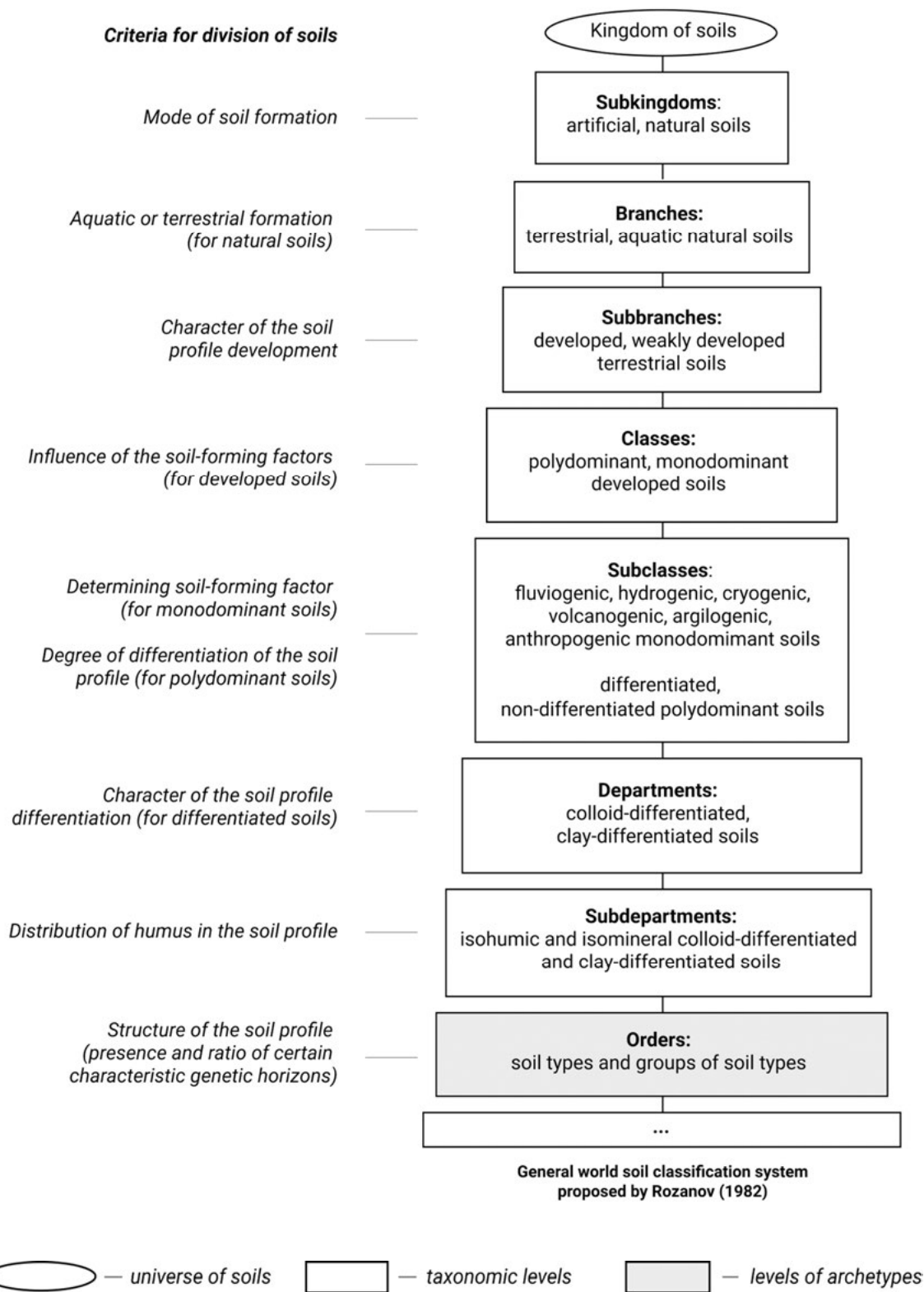


Figure 2 (cont.)

5.1.3 Violations of the rules for logical division of concepts

When using diagnostic criteria instead of differentiating criteria (or when using a set of differentiating criteria instead of a single differentiating one) at a time (that is, when dividing one class of objects), the rules for logical division of concepts are inevitably violated, and this makes the classification systems logically incorrect (Armand 1975, 141-151; Arnold 2002; Sokal 1974). There are many examples of such violations in SCSs:

- 1) in the Russian SCS, at the upper level of trunks, criteria for division of soils are the ratio of lithogenesis and soil formation and, at the same time, the nature of parent materials (Shishov et al. 2004);
- 2) in the German SCS, at the level of classes, criteria for division of soils are a similar stage of soil evolution and the dominant pedological processes (Ad-hoc-Arbeitsgruppe Boden 2005); and,
- 3) in the WRB, at the level of reference soil groups, they are mainly “characteristic soil features produced by primary pedogenetic process, except where special soil parent materials are of overriding importance” (IUSS Working Group WRB 2015, 8).

A rare exception is the SCS proposed by Rozanov (1982), which, at least at the upper levels, follows the rules for logical division of concepts; however, since it has an archetype level, it cannot be called purely genetic.

5.1.4 Lack of objective rules for the selection and ranking of criteria for division of soils

Currently, there are no objective rules for the selection and ranking of criteria for soil division (Nachtergaele et al. 2002); however, according to Rozova (1986, 163), no classification system can exist without such rules. Thus, it can be concluded that most SCSs are: 1) morphological (non-genetic); 2) artificial (not natural); 3) empirical (non-fundamental), that is, “based on several factors at the same level of categorization” (Manil 1959); 4) formal and descriptive (or descriptive with explanations), since “the qualitative diversity of the analyzed objects is simply stated” or partially explained in them (Rozova 1986, 54; 56-57); 5) pseudo-hierarchical; and, 6) static with some, if any, evolutionary elements. Evolutionary elements are present, for example, in the German SCS (Ad-hoc-Arbeitsgruppe Boden 2005), which follows the Kubiěna’s scheme “from the simplest poorly developed soils to the most complex, polygenetic ones” (Krasilnikov and Arnold 2009, 123). Another example is the SCS of the United Kingdom (Avery 1980), in which soil development stages are included at the highest level.

However, why are SCSs artificial? According to classiology, artificial systems are empirical, not based on a substantial theory, and simply document the similarities and differences between objects (Subbotin 2001, 69-70; Hjørland 2017, 111). They only help to achieve the visibility of many soils and ensure the effectiveness of their search; however, they do not reveal their nature (Rozova 1986, 204). In addition, soils in them are not “in an order according to their essential character” (Robinson 1950, 153, quote according to Muir 1962). According to Kubiěna (1958), this means that SCSs are built using a synthetic, rather than an analytical approach to the criteria for division of soils, and the presence of archetypes in them confirms this. In this regard, Kubiěna (1958, italics in original) notes: “Every artificial system of grouping is *only* possible by synthesis and by avoiding any kind of analytical approach.” He also emphasizes (1958) the “important role of analysis (in its wider sense) in soil research and the need to avoid synthesis as much as possible if the aim is the establishment of a natural system of soils and not just a rapid grouping.”

In conclusion of this section, it should be added that, out of the officially recognized SCSs, only three cover the whole world. These are the “U.S. Soil Taxonomy” (Soil Survey Staff 1999), the WRB (IUSS Working Group WRB 2015) and the French SCS (AFES 1998). The first two SCSs are used all over the world, and the third is only potentially suitable for the classification of world soils (Krasilnikov and Arnold 2009, 328).

5.2 Features for which soil classification systems differ from each other

The features for which SCSs differ from each other are:

1. Criteria for division of soils, which are usually represented by various “soil and environmental parameters” (Baruck et al. 2016, 6), for example, the morphological and chemical properties of soils, characteristics of climatic and soil regimes, as well as their combinations (Krasilnikov, Martí and Arnold 2009, 12; 26);
2. Methods for determining and measuring criteria for division of soils;
3. Archetypes, which are represented by various soil types, groups, series, etc.;
4. The number of taxonomic levels;
5. Taxonomic names;
6. Names of soils; and,
7. Objects that are included in them in addition to soils. For example, in addition to terrestrial natural soils, different SCSs include soil-like superficial bodies (subaquatic soils, bare rocks, soils strongly transformed by

agricultural activities, urban soils, and transported materials) and their combinations (Krasilnikov and Arnold 2009, 329).

6.0 Existing proposals for solving soil classification problems

To solve soil classification problems, the following is usually suggested:

- Correlation and harmonization of officially recognized SCSs. In 1998, this task was assigned to the WRB (IUSS Working Group WRB 2015), recommended by IUSS as a soil correlation system for all soil scientists (Nachtergaele et al. 2002).
- Development of a theoretical basis for soil classification. This is considered one of the most important tasks of soil science (Ibáñez and Boixadera 2002; Sokolov 2004, 165). In this regard, Polynov (1933, 45, quote according to Sokolov 2004, 165) notes: “[I]f the classification does not satisfy, then it is obvious that the theory is not completely consistent.” Rozova (1986, abstract) expresses a similar position: “The basis of the classification problem is the need to transfer science from the empirical stage of development to the theoretical one.”
- Objectivization of the soil classification process. On the one hand, it is believed that the use of innovative pedometric approaches, usually called objective, should greatly assist in the development of a universal SCS (Hempel et al. 2013; McBratney et al. 2003; Michéli et al. 2016; Nachtergaele et al. 2002). On the other hand, it is believed that the arguments of pedometricians are unjustified, and that “developments in pedometrics cannot replace the lack of theoretical studies” (Ibáñez and Boixadera 2002).
- Development of improved quantitative diagnostics (Hartemink 2015; Krasilnikov and Arnold 2009, 132; Nagy et al. 2016). For example, according to Nagy et al. (2016), “The application of faster, efficient, and more objective measurements can bring revolution to the classification of soils.”
- Creation of SCSs in the process of mapping and on its basis (Rozanov 1977, 4).

7.0 The “soil-landscape classification system”

The analysis of officially recognized and some underdeveloped SCSs, as well as definitions of soils from the point of view of classiology and the systems approach allowed us to identify their disadvantages and propose an interdisciplinary approach to the creation of a universal SCS (Nikiforova and Fleis 2018; Nikiforova et al. 2019). This ap-

proach was tested on the example of the European part of Russia in the process of multiscale soil-landscape GIS mapping (Fleis et al. 2016; Nikiforova et al. 2014; 2018). As a result, the scheme of SLCS was developed, which is fundamentally different from the existing SCSs and overcomes their shortcomings. This can be seen in Figure 3, if one compares it with Figures 1 and 2. The main features of SLCS are listed below:

- SLCS is based on the following definitions of the concepts natural soil and natural landscape developed by the author:

Natural soil is a material system and, at the same time, a derived element of a higher order material system, namely the natural landscape. Natural landscape consists of both the soil itself and the basic elements—rocks, air, water, and living and dead organisms. All landscape elements are material substances with homogeneous properties and are interrelated and interconnected with each other. The boundaries of natural landscapes and associated soils coincide (Mamai 2005, 31; 38). This follows from the systemic (that is, from the point of view of the systems approach) definition of natural soils and landscapes. Therefore, soil is a unique landscape element, since only soil arises and develops by interaction and interrelation of all other elements (Solntsev [1948] 2006). For example, air as one of the basic landscape elements cannot arise because of the interaction and interrelation of soil, water, rocks and organisms; the same applies to all other basic landscape elements.

- SLCS combines soil and landscape classification systems and, therefore, has two classification objects, which are at the same time its BUSCs. These objects are natural landscape system and its derived element—the soil, which is at the same time a self-sufficient system.
- SLCS is being developed as a complete hierarchy—from the general to the particular and from top to bottom, starting with the “zero” level, represented by the initial sets (universes) of all-natural landscapes and soils, and ending when BUSCs, that is, soil and landscape individuals, are reached.
- Its hierarchical levels are not taxonomic and have numbers instead of names.
- The successive division of natural landscapes in it is carried out in accordance with differentiating criteria and leads to the simultaneous division of associated soils.
- The differentiating criteria are determined by the essential character of soils, which consists in the fact that soils are, on the one hand, material systems, and on the

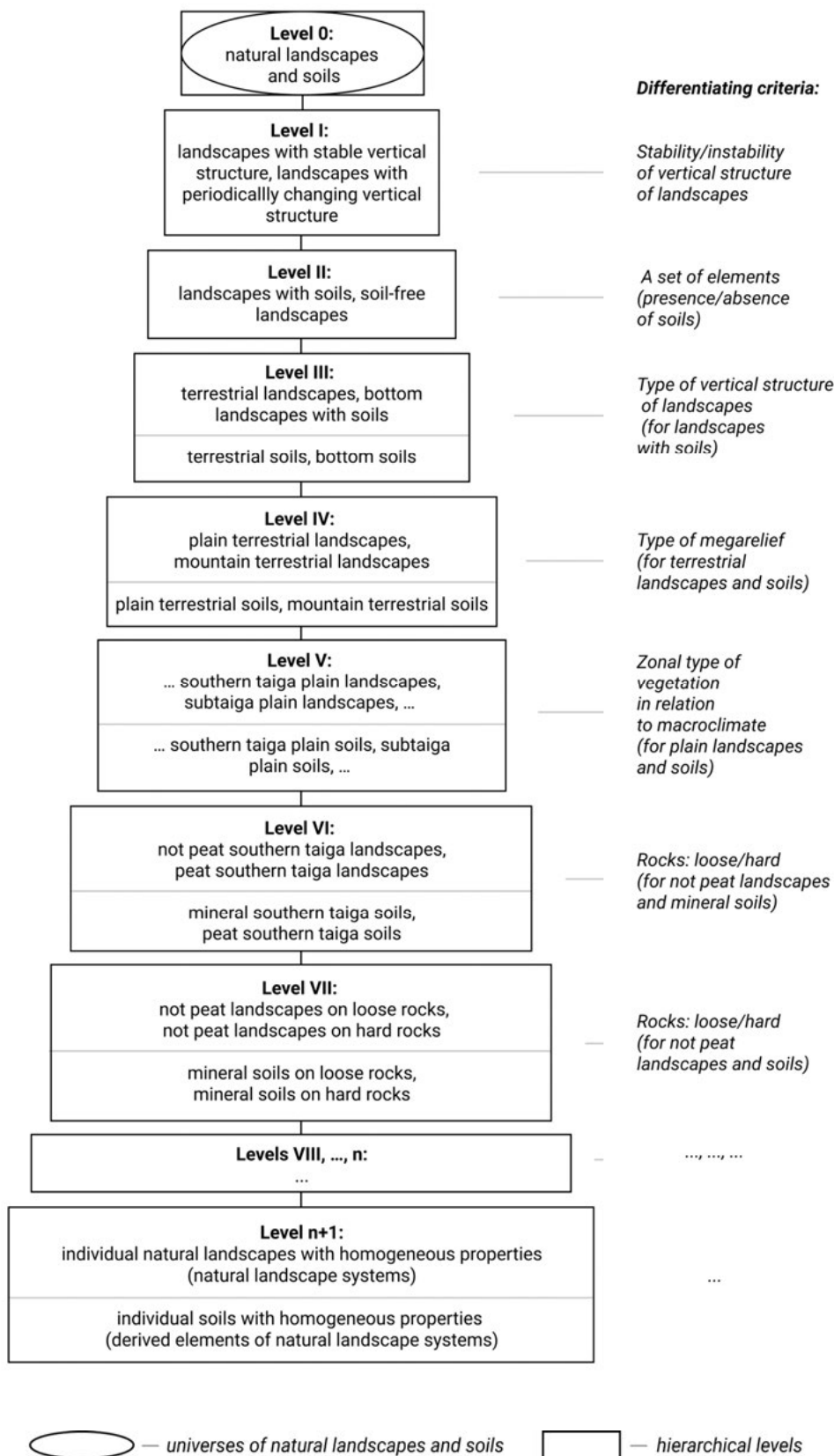


Figure 3. Structure features of SLCS.

other, derived elements of natural landscapes (material systems of a higher order). At the first three levels of classification, differentiating criteria are the main features of natural landscape systems, and at lower levels (in classification branches with soils), they are essential properties of the basic landscape elements.

- Differentiating and diagnostic criteria are separated in SLCS. Differentiating criteria determine diagnostic criteria and are used to divide soils and landscapes into classes (that is, for their classification). Diagnostic criteria are diagnostic properties of soils and landscapes and are used for the identification (classifying) soils and landscapes. The relation and subordination of the concepts described above can be represented in the form of the following chain: essential character of soils→differentiating criteria→diagnostic criteria→diagnostic properties. Diagnostic criteria are defined for all developed classes and subclasses of landscapes and associated soils, as well as for the soil as such, and presented in an online version of SLCS (http://geocnt.geonet.ru/en/landscapes_classification_first.step).
- The selection and ranking of differentiating criteria are subject to the rules developed (Nikiforova et al. 2019).
- SLCS functions as a classification system, as well as a diagnostics system.
- Due to the separation of differentiating and diagnostic criteria in SLCS, it is possible to identify the relationship between the features of landscapes and the properties of their basic elements, on the one hand, and the diagnostic properties of soils and landscapes, on the other. Thanks to this, the USC is able to solve scientific problems.
- SLCS integrates information on natural soils and landscapes.
- SLCS forms a new soil and landscape nomenclature, which reflects soil and landscape properties. Full names of soils and landscapes are obtained by combining their names at all hierarchical levels of a certain branch.
- SLCS includes natural terrestrial and bottom landscapes with and without soils, which allows determining the conditions under which the process of soil formation begins, and, therefore, to find the boundaries beyond which soil formation is impossible.
- SLCS is being developed in the process of multiscale soil-landscape GIS mapping.
- SLCS pursues both scientific and practical (applied) purposes and is, therefore, intended for scholars and practitioners in various fields of human activity who use soil and landscape information in their work.

In general, SLCS can be characterized as natural, genetic, hierarchical, and static. In the future, it is expected that

SLCS will combine the basic classification system with practical ones and will be interactive. It can also be interactive and evolutionary (i.e., have a time axis of coordinates), which will make it possible to distinguish between new and “old” (that is, completed due to changes in the properties of the main elements of the landscape) diagnostic soil properties, which are evidence of current and past soil-forming processes, respectively. In addition, we consider SLCS as a basis for a SCS with anthropogenic soils and landscapes. All this can significantly contribute to the inventory, modeling, and forecasting of natural soils and landscapes.

8.0 Conclusion

It can be concluded that none of the officially recognized national and international and underdeveloped SCSs can serve as the basis for creating a universal SCS, because they do not achieve most of scientific purposes that are set for them. To solve soil classification problems, an outside perspective is needed, that is, the use of classiology and the systems approach. Such an interdisciplinary approach allowed us to identify the causes of failures in creating a universal SCS, use both parts of the second version of the definition of soils proposed by Dokuchaev (1886) as a basis for SLCS, overcome most of the shortcomings of the existing SCSs, and suggest a way to make progress in soil classification.

Notes

1. Rozhkov (2012): “Classiology can be defined as a science studying the principles and rules of classification of objects of any nature. The development of the theory of classification and the particular methods for classifying objects are the main challenges of classiology.”
2. Blauberg and Iudin (2000): General systems theory (open system) approach is “a trend in methodology based on studying objects as systems.” According to von Bertalanffy (1968), a system is an entity, consisting of closely interrelated and interacting elements, and a system element is a minimal structural system unit with homogeneous properties; elements of material systems are material substances.
3. In soil science, as in Russian landscape science (Mamai 2005), there is a fairly clear separation of genetic and evolutionary classifications. Genetic classifications include those in which soils are subdivided into classes and subclasses depending on the conditions of their formation (that is, soil-forming factors), and evolutionary those in which soils are subdivided depending on the main stages of their formation and development over time. However, in other sciences, as a rule, there is no such separation

between genetic and evolutionary classifications (see Gnoli 2018).

References

- Ad-hoc-Arbeitsgruppe Boden. 2005. *Bodenkundliche Kartieranleitung*. 5th ed. Stuttgart: Schweizerbart.
- Adjei-Gyapong, T. and R.D. Asiamah. 2002. "The Interim Ghana Soil Classification System and Its Relation with the World Reference Base for Soil Resources." In *Quatorzième reunion du Sous-Comité ouest et centre africain de corrélation des sols pour la mise en valeur des terres, Abomey, Bénin, 9-13 October 2000*. Rapport sur les ressources en sols du monde 98. Rome: Organisation des Nations Unies pour l'alimentation et l'agriculture.
- Afanas'ev, Ya .N. 1923. "Zonal'nye sistemy pochv." *Zapiski Goretskogo sel'skokhozyaystvennogo instituta* 1, no. 2 (neofitsial'naya): 1-87.
- AFES (Association Française pour l'Étude du Sol). 1998. *A Sound Reference Base for Soils: The 'Référentiel Pédologique'*, trans. J.M. Hodgson, N.R. Eskenazi and D. Baize. Techniques et pratiques. Paris: Institut national de la recherche agronomique. Translation of *Référentiel pédologique*, 1990.
- Anderson, Darwin W. and C.A. Scott Smith. 2011. "A History of Soil Classification and Soil Survey in Canada: Personal Perspectives." *Canadian Journal of Soil Science* 91: 675-94.
- Armand, D.L. 1975. *Nauka o landshafte: Osnovy teorii i logiko-matematicheskije metody*. Moscow: Mysl.
- Arnold, Richard W. 2002. "Soil Classification Principles." In *Soil Classification 2001*, ed. E. Michéli, F.O. Nachtergaele, R.J.A. Jones. and L. Montanarella. European Soil Bureau Research Report 7. Luxembourg: Office for Official Publications of the European Communities, 3-9
- Avery, B.W. 1973. "Soil Classification in the Soil Survey of England and Wales." *Journal of Soil Science* 24: 324-38.
- Avery, B.W. 1980. *Soil Classification for England and Wales: Higher Categories*. Soil Survey Technical Monograph no. 14. Harpenden: Soil Survey.
- Basher, L.R. 1997. "Is Pedology Dead and Buried?" *Australian Journal of Soil Research* 35: 979-94.
- Baruck, Jasmin, Othmar Nestroy, Giacomo Sartori, Denis Baize, Robert Traidl, Borut Vrščaj, Esther Brämig, Fabian E. Gruber, Kati Heinrich and Clemens Geitner. 2016. "Soil Classification and Mapping in the Alps: The Current State and Future Challenges." *Geoderma* 264, part B: 312-31.
- Basinski, J.J. 1959. "The Russian Approach to Soil Classification and Its Recent Development." *Journal of Soil Science* 10: 14-26.
- Beckmann, G.G. 1984. "The Place of 'Genesis' in the Classification of Soils." *Australian Journal of Soil Research* 22: 1-14.
- Bertalanffy, Ludwig von. 1968. *General System Theory: Foundations, Development and Applications*. New York: George Braziller.
- Blauberg, I.V. and E.G. Iudin. 2000. "Systems approach." In *Collins Dictionary of Sociology*, 3rd ed. English translation of the article in the Great Soviet Encyclopedia, 1970-1979, 3rd ed. <http://encyclopedia2.thefreedictionary.com/systems+approach>
- Bockheim J.G., A.N. Gennadiyev, A.E. Hartemink and E.C. Brevik. 2014. "Soil-Forming Factors and Soil Taxonomy." *Geoderma* 226-227: 231-7.
- Brevik, Eric C., Costanza Calzolari, Bradley A. Miller, Paulo Pereira, Cezary Kabala, Andreas Baumgarten and Antonio Jordán. 2016. "Soil Mapping, Classification, and Pedologic Modeling: History and Future Directions." *Geoderma* 264, part B: 256-74.
- Bridges, E.M. 1990. *Soil Horizon Designations*. Technical Paper 19. Wageningen: International Soil Reference and Information Center.
- Buol, S.W., F.D. Hole and R.J. McCracken. 1980. *Soil Genesis and Classification*. 2nd ed., Ames: Iowa State University Press.
- Cline, M.G. 1949. "Basic Principles of Soil Classification". *Soil Science* 67: 81-91.
- Cline, M.G. 1962. "Logic of the New System of Soil Classification." *Soil Science* 96, no. 1: 17-22.
- Chorley, Richard J. and Barbara A. Kennedy. 1971. *Physical Geography: A Systems Approach*. London: Prentice-Hall International.
- Da Silva, Alessandra Fagioli, Maria João Pereira, João Daniel Cameiro, Zimback Célia Regina Lopes, Barbosa Paulo Milton Landim and Amílcar Soares. 2014. "A New Approach to Soil Classification Mapping Based on the Spatial Distribution of Soil Properties". *Geoderma* 219-220: 106-16.
- De Bakker, H. 1970. "Purposes of Soil Classification." *Geoderma* 4, no. 3: 195-208.
- Deressa, Abdenna, Markku Yli-Halla, Muktar Mohamed and Lemma Wogi. 2018. "Soil Classification of Humid Western Ethiopia: A Transect Study along a Toposequence in Didessa Watershed." *Catena* 163: 184-95.
- Dobrovolskii, G.V. 2005. "Soil Classification Principles: Soil Bodies and the Environment." *Eurasian Soil Science Supplement* 38, no 1: 1-5.
- Dokuchaev, V.V. 1879. *Kartografiya russkikh pochv: Ob'yasnitelnyi Tekst k Pochvennoy Karte Yevropeyskoy Rossii, izdannoy Departamentom zemledeliya i sel'skoy promyshlennosti*. Sankt-Peterburg: Tipografiya Kirshbauma.
- Dokuchaev, V.V. 1886. *Materialy k otsenke zemel Nizhegorodskoy gubernii. Yestestvenno-istoricheskaya chast: Otchyot Nizhego-*

- rodskomu gubernskomu zemstvu. Vypusk 1. Glavnye momenty v istorii otsenok zemel Yevropeyskoy Rossii, s klassifikatsionnykh russkikh pochv.* Sankt-Peterburg: Tipografiya Ye. Yevdokimova.
- Fleis, M.E., A.A. Nikiforova, M.V. Nyrtsov, M.M. Borisov and A.G. Khropov. 2016. "Sozdaniye masshtabnogo ryada pochvenno-landshaftnykh kart v geoinformatsionnoy srede." *Izvestiya RAN. Seriya Geograficheskaya* 1: 147-55.
- Florea, Nicolae. 2012. "Soil Facies: Geographic Local-Regional Completing of Soil Taxonomic Units." *Revue Roumaine de Géographie* 56, no. 2: 127-135.
- Frické, Martin. 2016. "Logical Division." *Knowledge Organization* 43: 539-49.
- Fridland, V.M. 1986. *Problemy geografii, genezisa i klassifikatsii pochv.* Moscow: Nauka.
- Gennadiyev, A.N. and M.I. Gerasimova. 1996. "Evolution of Approaches to Soil Classification in Russia and the United States: Stages of Divergence and Convergence." *Eurasian Soil Science* 28, no. 11: 15-24.
- Gerasimova, M.I. and N.B. Khitrov. 2012. "Comparison of the Results of Soil Profiles Diagnostics Performed in Three Classification Systems." *Eurasian Soil Science* 45, no. 12: 1087-94.
- Gnoli, Claudio. 2018. "Genealogical Classification." *ISKO Encyclopedia of Knowledge Organization*. <https://www.isko.org/cyclo/genealogical>
- Gobin, Anne, Paul Campling, Jozef Deckers and Jan Feyen. 2000. "Quantifying Soil Morphology in Tropical Environments: Methods and Application in Soil Classification." *Soil Science Society of America Journal* 64, no. 4: 1423-33.
- Hartemink, Alfred E. 2015. "The Use of Soil Classification in Journal Papers between 1975 and 2014." *Geoderma Regional* 5, no. 8: 127-39.
- Hartemink, Alfred E. and Budiman Minasny. 2014. "Towards Digital Soil Morphometrics." *Geoderma* 230-231: 305-317.
- Hempel, Jonathan, Erika Michéli, Phillip Owens and Alex McBratney. 2013. "Universal Soil Classification System Report from the International Union of Soil Sciences Working Group." *Soil Horizons* 54, no. 2: 1-6.
- Heuvelink, G.B.M. and R. Webster. 2001. "Modelling Soil Variation: Past, Present, and Future." *Geoderma* 100, no. 3-4: 269-301.
- Hjørland, Birger. 2017. "Classification." *Knowledge Organization* 44: 97-128.
- Hole, Francis D. and M. Hironaka. 1960. "An Experiment in Ordination of Some Profiles." *Soil Science Society of America Proceedings* 24: 309-12.
- Hughes, Philip A., Alex B. McBratney, Budiman Minasny and Sebastian Campbell. 2014. "End Members, End Points and Extragrades in Numerical Soil Classification." *Geoderma* 226-227: 365-75.
- Hughes, Philip, Alex B. McBratney, Jingyi Huang, Budiman Minasny, Erika Michéli and Jonathan Hempel. 2017. "Comparisons between USDA Soil Taxonomy and the Australian Soil Classification System I: Data Harmonization, Calculation of Taxonomic Distance and Inter-Taxa Variation." *Geoderma* 307: 198-209.
- Hughes, Philip, Alex B. McBratney, Budiman Minasny, Jingyi Huang, Erika Michéli, Jonathan Hempel and Edward Jones. 2018. "Comparisons between USDA Soil Taxonomy and the Australian Soil Classification System II: Comparison of Order, Suborder and Great Group Taxa." *Geoderma* 322: 48-55.
- Hbáñez, Juan José and Jaime Boixadera. 2002. "The Search for a New Paradigm in Pedology: A Driving Force for New Approaches to Soil Classification." In *Soil Classification 2001*, eds. E. Michéli, F.O. Nachtergaele, R.J.A. Jones and L. Montanarella. European Soil Bureau Research Report 7. Luxembourg: Office for Official Publications of the European Communities, 93-110.
- Isbell, R.F. 1992. "A Brief History of National Soil Classification in Australia Science. Since 1920s." *Australian Journal of Soil Research* 30, no. 6: 825-842.
- IUSS (International Union of Soil Sciences) Working Group WRB. 2015. *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. World Soil Resources Reports 106. Place FAO.
- Jenny, Hans. 1941. *Factors of Soil Formation: A System of Quantitative Pedology*. McGraw-Hill Publications in the Agricultural Sciences. New York: McGraw-Hill.
- Jones, Arwin, Luca Montanarella and Robert Jones, eds. 2005. *Soil Atlas of Europe*. European Soil Bureau Network, European Commission. Luxembourg: Office for Official Publications of the European Communities.
- Juilleret, Jérôme, Stefaan Dondeyne, Karen Vancampenhout, Jozef Deckers and Christophe Hissler. 2016. "Mind the Gap: A Classification System for Integrating the Subsoil into Soil Surveys." *Geoderma* 264, pt. B: 332-9.
- Juma, Noorallah G. 1999. *Introduction to Soil Science and Soil Resources*. Vol. 1 of *The Pedosphere and Its Dynamics: A Systems Approach to Soil Science* Edmonton: Salman Productions.
- Karpachevsky, L.O. 1981. *Les i lesnye pochvy*. Moscow: Leningradskaya Promyshlennost.
- Kellogg, Charles E. 1963. "Why a New System of Soil Classification?" *Soil Science* 96: 1-5.
- Kiryushin, V.I. 2011. *Klassifikatsiya pochv i agroecologicheskaya tipologiya zemel*. Sankt-Peterburg: Lan.
- Knox, Ellis G. 1965. "Soil Individuals and Soil Classification." *Proceedings of Soil Science Society of America* 29, 79-84.

- Kovda, V.A. 1973. *Obschchaya teoriya pochvoobrazovatel'nogo protsesssa*. Vol. 2 of *Osnovy ucheniya o pochvakh*. Moscow: Nauka.
- Krasilnikov, Pavel and Richard Arnold. 2009. *Soil Classifications and Their Correlations*. Pt. 2 of *A Handbook of Soil Terminology, Correlation and Classification*, ed. Pavel Krasilnikov, Juan-José Ibáñez Martí, Richard Arnold and Serghei Shoba. London: Earthscan, 45-335.
- Krasilnikov, Pavel, Juan-José Ibáñez Martí and Richard Arnold. 2009. *Theoretical Bases of Soil Classifications*. Pt. 1 of *A Handbook of Soil Terminology, Correlation and Classification*, ed. Pavel Krasilnikov, Juan-José Ibáñez Martí, Richard Arnold and Serghei Shoba. London: Earthscan, 5-43.
- Krasilnikov, Pavel, Juan-José Ibáñez Martí, Richard Arnold and Serghei Shoba, eds. 2009. *A Handbook of Soil Terminology, Correlation and Classification*. London: Earthscan.
- Kubišna, W.L. 1958. "The Classification of Soils." *Journal of Soil Science* 9, no. 1: 9-19.
- Láng, Vince, Márta Fuchs, István Waltner and Erika Michéli. 2013. "Soil Taxonomic Distance, a Tool for Correlation: As Exemplified by the Hungarian Brown Forest Soils and Related WRB Reference Soil Groups." *Geoderma* 192, no. 1: 269-76.
- Langohr, Roger. 2001. "Facing Problems in the Discipline of Soil Classification. Conclusions Based on 35 Years Practice and Teaching." In *Soil Classification*, eds. E. Michéli, F.O. Nachtergaele, R.J.A. Jones and L. Montanarella. Soil Bureau Research Report No. 7. Luxembourg: Office for Official Publications of the European Communities, 15-25.
- Lebedeva, I.I., V.D. Tonkonogov and M.I. Gerasimova. 1999. "Diagnostic Horizons in Substantive-Genetic Soil Classification Systems." *Eurasian Soil Science* 32, no. 9: 959-65.
- Lebedeva, I.I. and M.I. Gerasimova. 2009. "Factors of Soil Formation in Soil Classification Systems". *Eurasian Soil Science* 42, no. 12: 1412-18.
- Lebedeva, I.I. and M.I. Gerasimova. 2012. "Diagnostic Horizons in the Russian Soil Classification System." *Eurasian Soil Science* 45, no. 9: 823-33.
- Leeper, G.W. 1952. "On Classifying Soils." *Journal of the Australian Institute of Agricultural Science* 18: 77-80.
- Mamai, I.I. 2005. *Dinamika i funkcionirovaniye landschafton*. Moscow: Izdatel'stvo Moskovskogo Universiteta.
- Manil, G. 1959. "General Considerations on the Problem of Soil Classification." *Journal of Soil Science* 10: 5-13.
- Mazhitova, G.G., C.L. Ring, J.P. Moore, S.V. Gubin and C.A.S. Smith. 1994. "Correlation of Soil Classification for Northeastern Russia, America and Russia". *Eurasian Soil Science* 26, no. 5: 50-62.
- McBratney, A.B. and J.J. de Gruijter. 1992. "A Continuum Approach to Soil Classification by Modified Fuzzy k-Means with Extragrades." *Journal of Soil Science* 43, no. 1: 159-75.
- McBratney, A.B., M.L.M. Santos and B. Minasny. 2003. "On Digital Soil Mapping." *Geoderma* 117, nos. 1/2: 3-52.
- Michéli, Erica. 2008. *Improvement and International Correlation of the Hungarian Soil Classification System*. Research Report 46513. Hungarian Scientific Research Foundation.
- Michéli Erika, Vince Láng, Phillip R. Owens, Alex McBratney and Jon Hempel. 2016. "Testing the Pedometric Evaluation of Taxonomic Units on Soil Taxonomy: A Step in Advancing Towards a Universal Soil Classification System." *Geoderma* 264, part B: 340-9.
- Mill, John Stuart. 1882. *A System of Logic, Ratiocinative and Inductive*. 8th ed. New York: Harper.
- Miller, B.A. and R.J. Schaetzl. 2016. "History of Soil Geography in the Context of Scale." *Geoderma* 264, part B: 284-300.
- Murashkina, M., R.J. Southard and G.N. Koptsik. 2005. "Soil-Landscape Relationships in the Taiga of Northwestern Russia Highlight the Differences in the US and Russian Soil Classification Systems." *Soil Science* 170, no. 6: 469-80.
- Muir, J.W. 1962. "The General Principles of Classification with Reference to Soils." *Journal of Soil Science* 13: 22-30.
- Nachtergaele, Freddy, Frank R. Berdink and Jozef Deckers. 2002. "Pondering Hierarchical Soil Classification Systems." In *Soil Classification 2001*, ed. E. Michéli, F.O. Nachtergaele, R.J.A. Jones and L. Montanarella. European Soil Bureau Research Report 7. Luxembourg: Office for Official Publications of the European Communities, 71-9.
- Nagy, Judit, Adam Csorba, Vince Lang, Marta Fuchs and Erika Micheli. 2016. "Digital Soil Morphometrics Brings Revolution to Soil Classification." In *Digital Soil Morphometrics*, ed. Alfred Hartemink and Budiman Minasny. Springer, 365-81.
- Narayanan, Ram Mohan, S.E. Green and D.R. Alexander. 1992. "Soil Classification Using Midinfrared Off-Normal Active Differential Reflectance Characteristics." *Photogrammetric Engineering and Remote Sensing* 58, no. 2: 193-9.
- Nestroy, O., O.H. Dannenberg, M. English, A. Gessl, E. Herzenberger, W. Kilian, P. Nelhiebel, E. Pecina, A. Pehjamberger, W. Schneider and J. Wagner, J. 2000. "Systematische Gliederung der Boden Osterreichs (Osterreichische Bodensystematik 2000)." *Mitteilungen der Osterreichischen Bodenkundlichen Gesellschaft* 60: 1-104.
- Nikiforova, Aleksandra A. and Maria E. Fleis. 2018. "A Universal Soil Classification System from the Perspective of the General Theory of Classification: A Review." *Bulletin of Geography: Physical Geography Series* 14: 5-13.
- Nikiforova, Aleksandra A., Maria E. Fleis and Mikhail M. Borisov. 2014. "Towards Methodologies for Global Soil Mapping." In *GlobalSoilMap: Basis of the Global Spatial Soil*

- Information System. Proceedings of the 1st GlobalSoilMap Conference, Orléans, France, 7-9 October 2013*, ed. Dominique Arrouays, Neil McKenzie, Jon Hempel, Anne Richer de Forges and Alex B. McBratney. Leiden: CRC Press/Balkema, 291-4.
- Nikiforova, Aleksandra A., Maria E. Fleis and Maksim V. Nyrtsov. 2018. "Sozdaniye kart prirodnykh landshaftnykh sistem v srede GIS." In *Understanding and Monitoring Processes in Soils and Water Bodies*, ed. Viktor G. Sychev and Lothar Mueller. Vol. 2 of *Novel Methods and Results of Landscape Research in Europe, Central Asia and Siberia*. Moscow: FGBNU VNII agrokhimii: 29-34.
- Nikiforova, Aleksandra A., Olaf Basian, Maria E. Fleis, Maxim V. Nyrtsov, Aleksandr G. Khropov. 2019. "Theoretical Development of a Natural Soil-Landscape Classification System: An Interdisciplinary Approach." *Catena* 177, no. 6: 238-45.
- Ogen, Yaron, Naftaly Goldshleger and Eyal Ben-Dor. 2017. "3D Spectral Analysis in the VNIR-SWIR Spectral Region as a Tool for Soil Classification." *Geoderma* 302: 100-10.
- Parrochia, Daniel. 2017. "Mathematical Theory of Classification." *Knowledge Organization* 45: 184-201.
- Parrochia, Daniel and Pierre Neuville. 2013. *Towards a General Theory of Classifications*. Bâsel: Birkhäuser.
- Paton, T.R. and G.S. Humphreys. 2007. "A Critical Evaluation of the Zonalistic Foundations of Soil Science in the United States. Part I: The Beginning of Soil Classification." *Geoderma* 139, nos. 3/4: 257-67.
- Phillips, Jonathan. 1996. "Deterministic Complexity, Explanation and Predictability in Geomorphic Systems." In *The Scientific Nature of Geomorphology. Proceedings of the 27th Binghamton Geomorphology Symposium*, ed. Bruce L. Rhoads and Colin E. Thorn. New York: John Wiley, 315-36.
- Phillips, Jonathan D. 1998. "On the Relations between Complex Systems and the Factorial Model of Soil Formation (With Discussion)." *Geoderma* 86, nos. 1/2: 1-21.
- Pokrovsky, M.P. 2014. *Vvedeniye v klassiologiyu*. Yekaterinburg: IGG UrO RAN.
- Rayner, J.H. 1966. "Classification of Soils by Numerical Methods." *Journal of Soil Science* 17: 79-92.
- Riecken, F.F. 1963. "Some Aspects of Soil Classification in Farming." *Soil Science* 96, no. 1: 49-61.
- Rode, A.A. ed. 1975. *Tolkovnyi slovar po pochvovedeniyu*. Moscow: Nauka.
- Roazanov, B.G. 1977. *Pochvennyi pokrov zemnogo shara*. Moscow: MGU.
- Roazanov, B.G. 1982. "Skhemy obshchey klassifikatsii pochv mira." *Pochvovedenie* 8: 121-8.
- Rozhkov, V.A. 2011. "Formal Apparatus of Soil Classification." *Eurasian Soil Science* 44, no. 12: 1289-303.
- Rozhkov, V.A. 2012. "Classiology and Soil Classification." *Eurasian Soil Science* 45, no. 3: 221-30.
- Rozova, S.S. 1986. *Klassifikatsionnaya problema v sovremennoy nauke*. Novosibirsk: Nauka, Sibirskoye otdeleniye.
- Schelling, J. 1970. "Soil Genesis, Soil Classification and Soil Survey." *Geoderma* 4: 165-193.
- Shi, X.Z., D.S. Yu, S.X. Xu, E.D. Warner, H.J. Wang, W.X. Sun, Y.C. Zhao and Z.T. Gong. 2010. "Cross-Reference for Relating Genetic Soil Classification of China with WRB at Different Scales". *Geoderma* 155, nos. 3/4: 344-50.
- Shishov, L.L., V.D. Tonkonogov, I.I. Lebedeva and M.I. Gerasimova. 2004. *Klassifikatsiya i diagnostika pochv Rossii*. Smolensk: Oykumena).
- Shoba, S.A. ed. 2002. *Soil Terminology and Correlation*. 2nd ed. Petrozavodsk: Centre of the Russian Academy of Sciences.
- Shreyder, Yu. A. 1983. "Sistematika, tipologiya, klassifikatsiya" In *Teoriya i metodologiya biologicheskikh klassifikatsiy*. Moscow: Nauka, 90-100.
- Simonson, Roy W. 1989. *Historical Highlights of Soil Survey and Soil Classification with Emphasis on the United States, 1899-1970*. Technical Paper 18. Wageningen: International Soil Reference and Information Centre.
- Smith, G.D. 1965. "Lectures on Soil Classification." *Pedologie* 4: 1-134.
- Smith, G.D. 1983. "Historical Development of Soil Taxonomy: Background." In *Concepts and Interactions Vol. 1 of Pedogenesis and Soil Taxonomy*, ed. L.P. Willing, N.E. Smeck and G.F. Hall. Amsterdam: Elsevier, 23-49.
- Smith, G.D. 1986. *The Guy Smith Interviews: Rationale for Concepts in Soil Taxonomy*. Soil Management Support Services Technical Monograph 11. Washington, DC: U.S. Dept. of Agriculture.
- Sobocká, J. ed. 2000. *Morfogenetický klasifikačný systém pôd Slovenska. Bazálna referenčná taxonómia*. Bratislava: Výskumný ústav pôdozvedectva a ochrany pôdy.
- Soil Science Society of South Africa Soil Classification Working Group. 1977. *Soil Classification: A Binomial System for South Africa*. Science Bulletin 390. Pretoria: Department of Agriculture Technical Survey.
- USDA Natural Resources Conservation Service. 1999. *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. 2nd ed. Agriculture Handbook 436. Washington, DC: U.S. Dept. of Agriculture.
- Sokal, Robert R. 1974. "Classification: Purposes, Principles, Progress, Prospects." *Science* 185, no. 4157: 1115-23.
- Sokolov, I.A. 1978. "O bazovoy klassifikatsii gochv." *Pochvovedenie* 8: 113-23.
- Sokolov, I.A. 1991. "Bazovaya Substantivno-Geneticheskaya Klassifikatsiya pochv." *Pochvovedenie* 3: 107-21.
- Sokolov, I.A. 2004. *Teoreticheskiye problemy geneticheskogo pochvovedeniya*. 2nd ed. Novosibirsk: Gumanitarnye tehnologii.
- Solntsev, N.A. (1948) 2006. "The Natural Geographic Landscape and Some of Its General Rules." Trans. Alexander

- V. Khoroshev and Serge Andronikov. In *Foundation Papers in Landscape Ecology*, ed. John A. Wiens, Michael R. Moss, Monica G. Turner and David J. Mladenoff. New York: Columbia University Press, 19-27.
- Solntsev, V.N. 1981. *Sistemnaya organizatsiya landshaftov: Problemy metodologii i teorii*. Mysl.
- Stefanovits, P. 1963. *Magyarország talajai*. 2nd ed. Budapest: Akadémiai Kiadó.
- Subbotin, A.L. 2001. *Klassifikatsiya*. Moscow: Institut filosofii RAN.
- Targulian, V.O., Goryachkin, S.V. 2004. "Soil Memory: Types of Record, Carriers, Hierarchy and Diversity." *Revista mexicana de ciencias geológicas* 21: 1-8.
- Teng, Hongfen, Raphael A. Viscarra Rossel, Zhou Shi and Thorsten Behrens. 2018. "Updating a National Soil Classification with Spectroscopic Predictions and Digital Soil Mapping." *Catena* 164: 125-134.
- Vasques, G.M., J.A.M. Dematte, Raphael A. Viscarra Rossel, L. Ramirez-Lopez and E.S. Terra. 2014. "Soil Classification Using Visible/Near-Infrared Diffuse Reflectance Spectra from Multiple Depths." *Geoderma* 223-225: 73-8.
- Verheyen, Kris, Dries Adriaens, Martin Hermy and Seppe Deckers. 2001. "High-Resolution Continuous Soil Classification Using Morphological Soil Profile Descriptions". *Geoderma* 101, nos. 3/4: 31-48.
- Zádorová, Tereza and Vít Penížek. 2011. "Problems in Correlation of Czech National Soil Classification and World Reference Base 2006." *Geoderma* 167-168: 54-60.
- Zakharov, S.A. 1931. *Kurs pochvovedeniya*. 2nd ed. Moscow: Gosudarstvennoye Izdatel'sto Sel'skokhozyaystvennoy i Kolkhozno-Kooperativnoy Literatury.

Books Recently Published

Compiled by J. Bradford Young

DOI:10.5771/0943-7444-2019-6-489

- Alli, Mostafa. 2019. *Result Page Generation for Web Searching: Emerging Research and Opportunities*. Hershey, PA: Information Science Reference.
- Arfini, Selene. 2019. *Ignorant Cognition: A Philosophical Investigation of the Cognitive Features of Not-Knowing*. Cham: Springer.
- Balfe, Thomas, Joanna Woodall and Claus Zittel. 2019. *Ad Vivum?: Visual Materials and the Vocabulary of Life-Likeness in Europe before 1800*. Leiden: Brill.
- Benmarhnia, Tarik, Pierre-Marie David and Baptiste Godrie, eds. 2019. *Les sociétés de l'expérimentation: Enjeux épistémologiques, éthiques et politiques*. Québec: Presses de l'Université du Québec.
- Bernard, Andreas. 2019. *Theory of the Hashtag*, trans. Valentine A. Pakis. Cambridge, UK: Polity Press. Translation of *Diktat des Hashtags*.
- Camlot, Jason and Katherine McLeod, eds. 2019. *CanLit Across Media: Unarchiving the Literary Event*. Montreal: McGill Queen's University Press.
- Caron, Philippe. 2019. *L'enjeu des métadonnées dans les corpus textuels: Un défi pour les sciences humaines*. Rennes: Presses universitaires de Rennes.
- Caudill, David S., Shannon N. Conley, Michael E. Gorman and Martin Weinel, eds. 2019. *The Third Wave in Science and Technology Studies: Future Research Directions on Expertise and Experience*. Cham: Palgrave Macmillan.
- Chakrabarti, Arindam. 2019. *Realisms Interlinked: Objects, Subjects and Other Subjects*. London: Bloomsbury Academic.
- Chowdhury, G. G. and Sudatta Chowdhury. 2019. *Data and Information: Organization and Access*. London: Facet Publishing.
- Dumontet, Carlo. 2019. *Determining the Format of Books: An Introduction*. Amersham: Biblio-Graphica. 2 vols.
- Ferraro, Angela. 2019. *La réception de Malebranche en France au XVIIIe siècle: Métaphysique et épistémologie*. Paris: Classiques Garnier.
- Gidney, Kristen. 2019. *RDA Beta Toolkit: A Pictorial Glossary*, by Blinky Kill. [Canberra, ACT]: [Kristen Gidney].
- Glattfelder, James B. 2019. *Information-Consciousness-Reality: How a New Understanding of the Universe can Help Answer Age-Old Questions of Existence*. Cham: Springer Open.
- Gottlieb, Peter and David W. Carmicheal, eds. 2019. *Leading and Managing Archives and Manuscripts Programs*. Chicago: Society of American Archivists.
- Haider, Jutta and Olof Sundin. 2019. *Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life*. Abingdon, Oxon.: Routledge.
- Hetherington, Stephen, ed. 2019. *Epistemology: The Key Thinkers*. 2nd ed. London: Bloomsbury Academic.
- Jolibert, Bernard. 2019. *Science, religion, philosophie: Trois manières d'appréhender le monde*. Paris: L'Harmattan.
- Kelp, Christoph. 2019. *Good Thinking: A Knowledge First Virtue Epistemology*. New York: Routledge.
- Lehner, Ulrich L. and Ronald K. Tacelli. 2019. *Wort und Wahrheit: Fragen der Erkenntnistheorie*. Stuttgart: Verlag W. Kohlhammer.
- Lucarelli, Anna, Alberto Petrucciani and Elisabetta Viti, eds. 2019. *Viaggi a bordo di una parola: Scritti sull'indicizzazione semantica in onore di Alberto Cheti*. Roma: Associazione italiana biblioteche.
- Margolis, Sarah. 2019. *Accessibility of Big Data Imagery for Next Generation Machine Learning Applications*. Silver Spring, MD: U.S. Department of Commerce, National Oceanic and Atmospheric Administration.
- Mocombe, Paul Camy. 2019. *Haitian Epistemology*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Raftopoulos, Athanassios. 2019. *Cognitive Penetrability and the Epistemic Role of Perception*. [Cham]: Springer International.
- Santos, Boaventura de Sousa and Antoni Aguiló. 2019. *Aprendizajes globales: Descolonizar, desmercantilizar y despatriarcalizar desde las epistemologías del sur*. Barcelona: Icaria Editorial.
- Scott, Suzanne. 2019. *Fake Geek Girls: Fandom, Gender, and the Convergence Culture Industry*. New York: New York University Press.
- Yılmaz Şentürk, Elif. 2019. *Arşivlerde tanımlama ve üstveri: Osmanlı dönemi belgeleri*. Güngören, İstanbul: Hiperlink Yayınlar.

KNOWLEDGE ORGANIZATION

KO

Official Journal of the International Society for Knowledge Organization

ISSN 0943 – 7444

International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation

Publisher

Ergon – ein Verlag in der Nomos Verlagsgesellschaft mbH
 Waldseestraße 3-5
 D-76530 Baden-Baden
 Tel. +49 (0)7221-21 04-667
 Fax +49 (0)7221-21 04-27
 Sparkasse Baden-Baden Gaggenau
 IBAN: DE05 6625 0030 0005 0022 66
 BIC: SOLADES1BAD

Editor-in-chief (Editorial office)

KNOWLEDGE ORGANIZATION
 Journal of the International Society for Knowledge Organization
 Richard P. Smiraglia, Editor-in-Chief
 smiragli@uwm.edu

Instructions for Authors

Manuscripts should be submitted electronically (in Microsoft® Word format) in English only via ScholarOne at <https://mc04.manuscriptcentral.com/jisko>. Manuscripts that do not adhere to these guidelines will be returned to the authors for resubmission in proper form.

Manuscripts should be accompanied by an indicative abstract of approximately 250 words. Manuscripts of articles should fall within the range 6,000-10,000 words. Longer manuscripts will be considered on consultation with the editor-in-chief.

A separate title page should include the article title and the author's name, postal address, and E-mail address. Only the title of the article should appear on the first page of the text. Contact information must be present for all authors of a manuscript.

To protect anonymity, the author's name should not appear on the manuscript.

Criteria for acceptance will be appropriateness to the field of knowledge organization (see Scope and Aims), taking into account the merit of the contents and presentation. It is expected that all successful manuscripts will be well-situated in the domain of knowledge organization, and will cite all relevant literature from within the domain. Authors are encouraged to use the KO literature database at <http://www.isko.org/lit.html>.

The manuscript should be concise and should conform to professional standards of English usage and grammar. Authors whose native language is not English are encouraged to make use of professional academic English-language proofreading services. We recommend Vulpine Academic Services (vulpineacademic@gmail.com).

Manuscripts are received with the understanding that they have not been previously published, are not being submitted for publication elsewhere, and that if the work received official sponsorship, it has been duly released for publication. Submissions are refereed, and authors will usually be notified within 6 to 8 weeks.

Under no circumstances should the author attempt to mimic the presentation of text as it appears in our published journal. Instead, please follow these instructions.

In Microsoft® Word please set the language preference ("Tools," "Language") to "English (US)" or "English (UK)."

The entire manuscript should be double-spaced, including notes and references.

The text should be structured with decimally-numbered subheadings (1.0, 1.1, 2.0, 2.1, 2.1.1, etc.). It should contain an introduction, giving an overview and stating the purpose, a main body, describing in sufficient detail the materials or methods used and the results or systems developed, and a conclusion or summary.

Author-generated keywords are not permitted.

Footnotes are not allowed. Endnotes are accepted only in rare cases and should be limited in number; all narration should be included in the text of the article. Do not use automatic footnote formatting. Instead, insert a superscript numeral (Format, Font, Superscript) and create the text of the note manually in a separate list at the end of the manuscript, before the reference list.

Paragraphs should include a topic sentence, a developed narrative and a conclusion; a typical paragraph has several sentences. Paragraphs with tweet-like characteristics (one or two sentences) are inappropriate.

Italics are permitted only for phrases from languages other than English, and for the titles of published works.

Bold type is not permitted.

Em-dashes should not be used as substitutes for commas. Dashes must be inserted manually (Insert, Advanced Symbol, Em-dash) with no spaces on either side.

Do not use automatic formatting of any kind. To indent, use the ruler. Do not use tabs under any circumstances. For a bulleted list, indent the list using the ruler, then insert bullets (Insert, Advanced Symbol, bullet). Do not use automatically-numbered paragraphs.

Illustrations should be embedded within the document. Photographs (including color and half-tone) should be scanned with a minimum resolution of 600 dpi and saved as .jpg files. Tables should contain a number and caption at the bottom, and all columns and rows should have headings. All illustrations should be cited in the text as Figure 1, Figure 2, etc. or Table 1, Table 2, etc.

Examples of classification arrays should be configured as figures and set into the document as .jpgs; they should not be entered as editable text.

Remove all active hyperlinks, including those from reference formatting software (if hovering over the text with a mouse produces a gray highlight, the text is hyperlinked; remove the link "Insert," "Hyperlink," "Remove link").

Reference citations within the text should have the form: (Author year). For example, (Jones 1990). Specific page numbers are required for quoted material, e.g. (Jones 1990, 100). A citation with two authors would read (Jones and Smith 1990); three or more authors would be: (Jones et al. 1990). When the author is mentioned in the text, only the date and optional page number should appear in parentheses: "According to Jones (1990), ..." or "Smith wrote (2010, 146): ..." A subsequent page reference to the same cited work (e.g., to Smith 2010) should have the form "(229)." There is never a comma before the date.

In-text citations should not be routinely placed at the end of a sentence or after a quotation, but an attempt should be made to work them into the narrative. For example:

"Jones (2010, 114) reported statistically significant results.

"Many authors report similar data; according to Matthews (2014, 94): "all seven studies report means within $\pm 5\%$."

In-text citations should precede block quotations, and never are placed at the end of a block-quotation.

References should be listed alphabetically by author at the end of the article. Reference lists should not contain references to works not cited in the text. Websites mentioned in passing in the text should be identified parenthetically with their URLs but not with references unless a specific page of a specific website is being quoted.

Author names should be given as found in the sources (not abbreviated, but also not fuller than what is given in the source). Journal titles should not be abbreviated. Multiple citations to works by the same author should be listed chronologically and should each include the author's name. Articles appearing in the same year should have the following format: "Jones 2005a, Jones 2005b, etc."

Proceedings must be identified fully by title, editor, and details of publication.

Journal issue numbers are given only when a journal volume is not through-paginated. References for published electronic resources should be accompanied by either a URL or DOI but not in lieu of actual publication data; access dates are not allowed.

Unpublished electronic resources may use an access date in lieu of a date of publication. In cases of doubt, authors are encouraged to consult *The Chicago Manual of Style* 17th ed. (or online), author-date reference system (chapter 15).

Examples:

Dahlberg, Ingetraut. 1978. "A Referent-Oriented, Analytical Concept Theory for INTER-CONCEPT." *International Classification* 5: 142-51.

Howarth, Lynne C. 2003. "Designing a Common Namespace for Searching Metadata-Enabled Knowledge Repositories: An International Perspective." *Cataloging & Classification Quarterly* 37, nos. 1/2: 173-85.

Pogorelec, Andrej and Alenka Saupel. 2006. "The Alternative Model of Classification of Belles-Lettres in Libraries." *Knowledge Organization* 33: 204-14.

Schallier, Wouter. 2004. "On the Razor's Edge: Between Local and Overall Needs in Knowledge Organization." In *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference 13-16 July 2004 London, UK*, edited by Ia C. McIlwaine. Advances in knowledge organization 9. Würzburg: Ergon Verlag, 269-74.

Smiraglia, Richard P. 2001. *The Nature of 'a Work': Implications for the Organization of Knowledge*. Lanham, Md.: Scarecrow.

Smiraglia, Richard P. 2005. "Instantiation: Toward a Theory." In *Data, Information, and Knowledge in a Networked World: Annual Conference of the Canadian Association for Information Science ... London, Ontario, June 2-4 2005*, ed. Liwen Vaughan. <http://www.caais-caisi.ca/2005proceedings.htm>.

Upon acceptance of a manuscript for publication, authors must provide a digital photo and a one-paragraph biographical sketch (fewer than 100 words). The photograph should be scanned with a minimum resolution of 600 dpi and saved as a .jpg file.

© Ergon – ein Verlag in der Nomos Verlagsgesellschaft, Baden-Baden 2019. All Rights reserved.

KO is published by Ergon.

Annual subscription 2019:

- Print + online (8 issues/ann.; unlimited access for your Campus via Nomos eLibrary) € 359,00/ann.
- Prices do not include postage and packing
- Cancellation policy: Termination within 3 months' notice to the end of the calendar year

Scope

The more scientific data is generated in the impetuous present times, the more ordering energy needs to be expended to control these data in a retrievable fashion. With the abundance of knowledge now available the questions of new solutions to the ordering problem and thus of improved classification systems, methods and procedures have acquired unforeseen significance. For many years now they have been the focus of interest of information scientists the world over.

Until recently, the special literature relevant to classification was published in piecemeal fashion, scattered over the numerous technical journals serving the experts of the various fields such as:

philosophy and science of science
science policy and science organization
mathematics, statistics and computer science
library and information science
archivistics and museology
journalism and communication science
industrial products and commodity science
terminology, lexicography and linguistics

Beginning in 1974, KNOWLEDGE ORGANIZATION (formerly INTERNATIONAL CLASSIFICATION) has been serving as a common platform for the discussion of both theoretical background questions and practical application problems in many areas of concern. In each issue experts from many countries comment on questions of an adequate structuring and construction of ordering systems and on the problems of their use in opening the information contents of new literature, of data collections and survey, of tabular works and of other objects of scientific interest. Their contributions have been concerned with

- (1) clarifying the theoretical foundations (general ordering theory/science, theoretical bases of classification, data analysis and reduction)
- (2) describing practical operations connected with indexing/classification, as well as applications of classification systems and thesauri, manual and machine indexing
- (3) tracing the history of classification knowledge and methodology
- (4) discussing questions of education and training in classification
- (5) concerning themselves with the problems of terminology in general and with respect to special fields.

Aims

Thus, KNOWLEDGE ORGANIZATION is a forum for all those interested in the organization of knowledge on a universal or a domain-specific scale, using concept-analytical or concept-synthetic approaches, as well as quantitative and qualitative methodologies. KNOWLEDGE ORGANIZATION also addresses the intellectual and automatic compilation and use of classification systems and thesauri in all fields of knowledge, with special attention being given to the problems of terminology.

KNOWLEDGE ORGANIZATION publishes original articles, reports on conferences and similar communications, as well as book reviews, letters to the editor, and an extensive annotated bibliography of recent classification and indexing literature.

KNOWLEDGE ORGANIZATION should therefore be available at every university and research library of every country, at every information center, at colleges and schools of library and information science, in the hands of everybody interested in the fields mentioned above and thus also at every office for updating information on any topic related to the problems of order in our information-flooded times.

KNOWLEDGE ORGANIZATION was founded in 1973 by an international group of scholars with a consulting board of editors representing the world's regions, the special classification fields, and the subject areas involved. From 1974-1980 it was published by K.G. Saur Verlag, München. Back issues of 1978-1992 are available from ERGON-Verlag, too.

As of 1989, KNOWLEDGE ORGANIZATION has become the official organ of the INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION (ISKO) and is included for every ISKO-member, personal or institutional in the membership fee.

Annual subscription 2019: Print + online (8 issues/ann.; unlimited access for your Campus via Nomos eLibrary) € 359,00/ann. Prices do not include postage and packing. Cancellation policy: Termination within 3 months' notice to the end of the calendar year

Ergon – ein Verlag in der Nomos Verlagsgesellschaft mbH, Waldseestraße 3-5, D-76530 Baden-Baden, Tel. +49 (0)7221-21 04-667, Fax +49 (0)7221-21 04-27, Sparkasse Baden-Baden Gaggenau, IBAN: DE05 6625 0030 0005 0022 66, BIC: SOLADES1BAD

Founded under the title International Classification in 1974 by Dr. Ingetraut Dahlberg, the founding president of ISKO. Dr. Dahlberg served as the journal's editor from 1974 to 1997, and as its publisher (Indeks Verlag of Frankfurt) from 1981 to 1997.

The contents of the journal are indexed and abstracted in *Social Sciences Citation Index*, *Web of Science*, *Information Science Abstracts*, *INSPEC*, *Library and Information Science Abstracts (LISA)*, *Library, Information Science & Technology Abstracts (EBSCO)*, *Library Literature and Information Science (Wilson)*, *PASCAL*, *Referativnyi Zhurnal Informatika*, and *Sociological Abstracts*.

Werbung