an open access journal

RESEARCH ARTICLE

Understanding knowledge role transitions: A perspective of knowledge codification

Jinqing Yang[1], Wei Lu[2,3], Yong Huang[2,3], Qikai Cheng[2,3], Li Zhang[2,3], and Shengzhi Huang[2,3]

[1]School of Information Management, Central China Normal University, Wuhan, China
[2]School of Information Management, Wuhan University, Wuhan, China
[3]Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, ChinJinqing Yang

Keywords: bibliometrics, formal knowledge, knowledge codification, knowledge transition

Citation: Yang, J., Lu, W., Huang, Y., Cheng, Q., Zhang, L., & Huang, S. (2022). Understanding knowledge role transitions: A perspective of knowledge codification. *Quantitative Science Studies*, *3*(4), 1133–1155. https://doi.org/10.1162/qss_a_00221

DOI:
https://doi.org/10.1162/qss_a_00221

Peer Review:
https://publons.com/publon/10.1162/qss_a_00221

Supporting Information:
https://doi.org/10.1162/qss_a_00221

Received: 18 November 2021
Accepted: 21 August 2022

Corresponding Author:
Yong Huang
yonghuang1991@whu.edu.cn

Handling Editor:
Ludo Waltman

The MIT Press

ABSTRACT

Informal knowledge constantly transitions into formal domain knowledge in the dynamic knowledge base. This article focuses on an integrative understanding of the knowledge role transition from the perspective of knowledge codification. The transition process is characterized by several dynamics involving a variety of bibliometric entities, such as authors, keywords, institutions, and venues. We thereby designed a series of temporal and cumulative indicators to respectively explore transition possibility (*whether new knowledge could be transitioned into formal knowledge*) and transition pace (*how long it would take*). By analyzing the large-scale metadata of publications that contain informal knowledge and formal knowledge in the PubMed database, we find that multidimensional variables are essential to comprehensively understand knowledge role transition. More significantly, early funding support is more important for improving transition pace; journal impact has a positive correlation with the transition possibility but a negative correlation with transition pace; and weaker knowledge relatedness raises the transition possibility, whereas stronger knowledge relatedness improves the transition pace.

1. INTRODUCTION

Knowledge is a driving force for technological and economic change, notably scientific knowledge (Mistry & Berardi, 2016; Naghavi & Walsh, 2011). Scientific knowledge is a wide and abstract concept and one way to classify knowledge is as tacit or explicit (Ahmadyousefi, Choobchian et al., 2020). As tacit knowledge can only be understood by people who have the same personal experience, the codification of knowledge is considered as the essential method for translating tacit knowledge into economically viable innovations (Lissoni, 2001). Specifically, knowledge codification not only alters the tacit forms of knowledge but is a process of knowledge creation. When new knowledge is codified, new concepts and terminology are introduced, which inherently involves the further creation of knowledge (Cohendet & Meyer-Krahmer, 2001). Codified knowledge is explicit and formal. It comes in a variety of forms (Faber, 2011; Pór & Molloy, 2000; Su & Lee, 2010), such as words and numbers, scientific procedures, or universal principles. For example, keywords are usually considered fine-grained knowledge entities (Yang, Li & Huang, 2018). New knowledge was transformed into formal knowledge after the operation of knowledge codification. There is therefore a need to explore and understand better knowledge transformation as a form for effecting the creation of knowledge.

In the context of knowledge codification, formal knowledge refers to knowledge that is codified. Scientific knowledge can be intuitively defined as formal knowledge when adopted into the domain knowledge base (Hjørland & Albrechtsen, 1995; Möller, Sintek et al., 2008; Wang, Hamilton & Bither, 2005). Correspondingly, informal knowledge refers to the knowledge entities that were not codified or adopted into the knowledge base. Knowledge role transition is the transformation from informal knowledge to formal knowledge in the process of knowledge codification. Formal knowledge is usually described by an ontological approach, which is the formal, explicit specification of a shared conceptualization of a domain. The machine interpretability of formal knowledge (*codified knowledge*) has been increasingly critical to improving the performance of information retrieval and image understanding (Möller, et al., 2008; Wildemuth, 2004) and organizing solutions to other real problems that exist in a specific research domain (Cardoso, Da Silveira, & Pruski, 2020; Tsatsaronis, Varlamis, 2013). Thus, understanding different transition patterns of formal knowledge is critical to tracking credible and valuable scientific knowledge and promoting innovation in science and technology.

Formal knowledge was born from the role transition of informal knowledge in an exogenous and endogenous-driven process. As the dynamic cooperation of research elements involved could lead to the role transition of scientific knowledge, we attempt to understand the patterns of knowledge role transition from multiple dimensions (i.e., publication, author, institution, funding, descriptors (keywords), and venue (journals)). The three primary contributions of this paper are as follows:

1. We attempt to understand the role transition in the process of knowledge growth from the perspective of knowledge codification.
2. We investigate influence factors of the transition possibility from informal knowledge to formal knowledge.
3. We design a series of temporal indicators to characterize the transition pace from informal knowledge to formal knowledge.

The remainder of this paper is structured as follows. Section 2 contains twofold surveys on analysis dimension selection for understanding knowledge transition and the knowledge growth process from a knowledge codification perspective. Section 3 provides concept definitions and proposes variables based on the metadata of scholarly publications. Section 4 introduces the empirical data and constructs the formal knowledge and informal knowledge matched-pair samples. Section 5 presents the match-paired analysis of formal knowledge and informal knowledge and dynamic correlation analysis between various variables and transition time. Section 6 comprehensively discusses the findings and implications of this study. Section 7 concludes the overview of our work.

2. RELATED WORK

2.1. Analysis Dimension Selection for Understanding Knowledge Transition

The growth process is always bounded and usually follows an S-shaped or sigmoid curve in many studies including knowledge growth (Shimogawa, Shinno & Saito, 2012). The process of knowledge growth has also been described as a sequence of life stages: either three main stages (embryonic, early, and recognized) or four S-curve stages (birth, growth, maturity, and senility) (Tu & Seng, 2012; van den Oord & van Witteloostuijn, 2018). Thus, the growth process can also be crudely considered a transition from one stage to another. Knowledge transition could be characterized across several dimensions, such as the number of actors involved (e.g., scientists, institutions), funding support, and knowledge outputs produced (e.g., publications, patents). Most importantly, these dimensions are likely to coevolve and possibly cause different effects over different stages of knowledge growth.

Researchers have devoted much time and effort to understanding the transition process of scientific knowledge for discovering reliable and valuable knowledge. Emerging technology and topic discoveries are representative studies of detecting promising scientific knowledge through understanding the role transition patterns of technologies and topics. To be more specific, Tu and Seng (2012) calculated the publication number corresponding to each keyword to measure the power of transitioning into emerging topics. Some researchers have explored the transition patterns of scientific technologies and research topics from various dimensions according to the scholarly metadata of publications. Rotolo, Hicks, and Martin (2015) summarized scholarly metadata (i.e., authors, institutions, funding, keywords) to crystallize the five attributes of emerging technology: radical novelty; relatively fast growth; coherence; prominent impact; and uncertainty and ambiguity. Soon afterward, Carley, Newman et al. (2018) developed four attributes by metadata analytics to characterize emerging technology: novelty, persistence, community, and growth. Referring to emerging technology, Wang (2018) proposed the four attributes (i.e., radical novelty, relatively fast growth, coherence, and scientific impact) of emerging research topics based on scholarly metadata analytics. Iqbal, Qadir et al. (2019) conducted metadata analysis to identify research topics with high impact from the publication number as well as citation count dimensions. Weis and Jacobson (2021) detected the early warning signal for impactful research by analyzing high-dimensional relationships among metadata of the scientific literature from papers, authors, and journals.

Scientific knowledge transition is affected by various factors. Thus, it is essential for observing the growth process from different dimensions of scholarly publications, which helps us receive a comprehensive understanding of knowledge role transition. Scholarly metadata was associated with *publication authors*, *institutions*, *funding*, *citations*, *venues*, *keywords*, etc., which benefits characteristic extraction. This study will take advantage of the large-scale metadata information and the intertwining relationships in scholarly publications to understand the role transition of scientific knowledge.

2.2. Understanding the Knowledge Growth Process from the Knowledge Codification Perspective

Knowledge codification is a process of reducing and converting tacit knowledge into words, numbers or universal principles that serve to reconstitute new knowledge (Cohendet, & Meyer-Krahmer, 2001). As knowledge owns some elements that can be codified and transferred (Sutton, 2001), it was transferred from one side to the other through knowledge codification (Liu, Ray & Whinston, 2010). If new knowledge is codified, it can as well be transmitted via verbal communication as through a written knowledge repository (Guechtouli & Kasmi, 2014). Knowledge codification is one of the important technologies and strategies in knowledge transmission and management (Maio, Fenza et al., 2010).

In the field of library information science, thesauri were the typical ontology codifications, so that the concepts or terms included in a thesaurus could be considered as formal knowledge. The process of knowledge growth can be characterized from the knowledge codification perspective. The relevant studies here are those that examined thesaurus extension where new concepts and terms were adopted into a thesaurus. Fabian, Wächter, and Schroeder (2012) attempted to identify knowledge growth patterns by using their sibling relationship to predict the knowledge base extension. Tsatsaronis, Varlamis et al. (2013) proposed temporal variables to understand the dynamic process of concept transformation of the MeSH (Medical Subject Headings) ontology. Furthermore, Cardoso, Pruski, and Da Silveira (2018) introduced external sources of knowledge (i.e., PubMed and UMLS) to support biomedical ontology evolution by identifying outdated knowledge entities and the required types of change for the domain to evolve. Additionally, researchers universally recognized that the transformation of knowledge entities in the knowledge ecosystem is analogous to that of biological units in the ecosystem (Sice, Thirkle & Ogwu, 2018).

This section has elaborated the knowledge growth process from a knowledge codification perspective. For knowledge entities, researchers focus on their growth process to understand the factors of knowledge transition based on thesauri. In the process of knowledge growth, if the knowledge entity is adopted into thesauri, it transitions into formal knowledge. That can provide realistic scenarios for investigating knowledge role transitions.

3. METHODS

3.1. Concept Definition Under Knowledge Codification

The research objective of this paper is to investigate the role of transition patterns from informal knowledge to formal knowledge in terms of transition possibility and transition pace. In the field of library information science, the thesaurus is composed of codified knowledge. We defined the relevant concepts based on a domain thesaurus to improve readability, as shown in Figure 1.

For example, knowledge A and C have the same time intervals, but they have different growth trajectories, which indicates that knowledge entities have different transition possibilities. "Balloon embolectomy" (A) and "neuraminidase genes" (C) both appeared in abstracts in 1976 for the first time. "Balloon embolectomy" was codified and adopted in a domain thesaurus in 2012, whereas "neuraminidase genes" was not, meaning that "balloon embolectomy" was transitioned into formal knowledge but "neuraminidase genes" was informal knowledge at that time. In addition, knowledge B takes a shorter transition time to become formal knowledge compared to knowledge A, which implies that the pace of knowledge transition is different. Transformation time refers to the length of time that knowledge entities take from their first appearance to their adoption into the thesaurus. For example, "forensic toxicology" (B) also first appeared in abstracts in 1976 and was codified and adopted into the domain

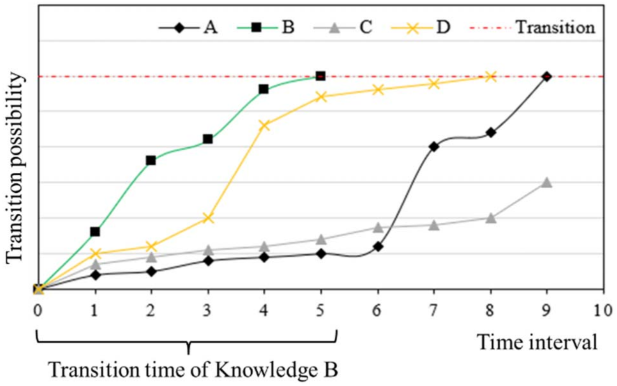

Figure 1. An illustration of the knowledge role transition.

thesaurus in 2006, making its transition time 30 years (2006–1976). Thus, "forensic toxicology" (B) has a shorter transition time than "balloon embolectomy" (A).

1. Formal knowledge: This refers to knowledge that is codified and adopted into a domain thesaurus, such as A, B, and D in Figure 1.
2. Informal knowledge: This is derived from uncodified knowledge entities that are not adopted into a domain thesaurus currently, such as C in Figure 1.
3. Knowledge role transition: In the process of knowledge growth, informal knowledge is codified and adopted into a domain thesaurus. This means that informal knowledge transitions into formal knowledge. This fact is considered as knowledge role transition.
4. Transition possibility: This refers to whether informal knowledge could transition into formal knowledge; this action may be driven by exogenous and endogenous factors.
5. Transition pace: This refers to how long the role transition of knowledge takes, which can be measured through the adoption time in the records of a domain thesaurus.

3.2. Measuring Indicators

To depict or characterize the growth process of new knowledge, we have selected various characteristics by taking advantage of various metadata information and the intertwining relationships in scholarly publications (i.e., authors, keywords/descriptors, citations, institutions, journals, funding) (Salatino, 2019), as shown in Figure 2.

Specifically, according to the metadata information, the number of actors involved, funding, and knowledge outputs produced were utilized in order to characterize the growth patterns of knowledge entities (Carley et al., 2018; Rotolo et al., 2015). In Figure 2, the six circle nodes (except knowledge entities) are the six dimensions of analyzing the influence factors of transition possibility and transition pace. Thus, we propose the corresponding variables: scholarly publication, author, institution, funding, descriptors (keywords), and venue (journals), as shown in Table 1.

In the following content, we elaborate on the rationale behind the set of proposed variables.

3.2.1. Publication dimension

The number of scholarly publications is the more intuitive signal of knowledge output (Wang, 2018). Because annotation terms always appear one at a time in the keyword list, the number of publications is a measuring indicator of knowledge growth (Tu & Seng, 2012). In addition, the citation relationship not only provides information on the impact of scholarly publications but also measures the impact of research components such as authors, institutions, journals,

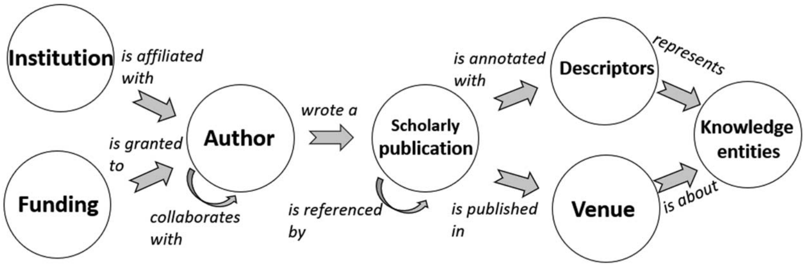

Figure 2. The scholarly metadata and their relationships.

Table 1. The variables corresponding to different dimensions of metadata information

Dimension	Variable	Abbreviation	Description
Publication	*Cumulative publication number*	*Cumulative # pubs*	Number of publications in the span of transition time.
	Yearly average of publication number	*Annual avg. pubs*	Yearly average of publication number in the span of transition time.
	Cumulative citation count	*Cumulative # citations*	Count of accumulative citation in the span of transition time.
	Yearly average of citation count	*Annual avg. citations*	Yearly average of citation count in the span of transition time.
Author	*Cumulative author count*	*Cumulative # authors*	Count of authors in the span of transition time.
	Author average impact	*Author avg. impact*	Average *h*-index of authors adopting descriptors in the span of transition time.
	Yearly average of author count	*Annual avg. authors*	Yearly average of author count in the span of transition time.
	Average author counts per publication	*Avg. authors per pub*	Average author counts per publication in the span of transition time.
Institution	*Cumulative institution number*	*Cumulative # institutions*	Number of institutions in the span of transition time.
	Yearly average of institution number	*Annual avg. institutions*	Yearly average of institution number in the span of transition time.
Funding	*Cumulative funding number*	*Cumulative # funding*	Number of funding awards in the span of transition time.
	Yearly average of funding number	*Annual avg. funding*	Yearly average of funding award number in the span of transition time.
Journal	*Journal cumulative number*	*Cumulative # journals*	Number of journals published by descriptors in the span of transition time.
	Yearly average of journal number	*Annual avg. journals*	Yearly average of journal number in the span of transition time.
	Journal average impact	*Journal avg. impact*	Average impact of journals in which descriptors published in the span of transition time.
Descriptor	*Knowledge relatedness*	*Knowledge relatedness*	Relatedness degree with other descriptors in the span of transition time.
	Yearly average of knowledge relatedness	*Annual avg. knowledge relatedness*	Yearly relatedness degree with other descriptors in the span of transition time.

and keywords. (Waltman, 2016). The number of citations in scholarly publications could provide an indication of attention inside the academic domain. Thus, the publication dimension could provide four indicators (i.e., *cumulative # pubs*, *annual avg. pubs*, *cumulative # citations*, and *annual avg. citations*).

3.2.2. Author dimension

Authors of scholarly publications are important to explore in scientific metrics, such as author collaboration (Ebrahimi, Asemi et al., 2021; Guan, Yan & Zhang, 2017; Kaur & Mahajan, 2015) and author impact (Amjad, Rehmat et al., 2020; Dunaiski, Geldenhuys & Visser, 2018). Authors with high impact would lead the development of a discipline. Researchers contribute to the updating and growth of formal knowledge in the form of research results (Sun & Latora, 2020). The author count has been commonly adopted to compute the author community (Rotolo et al., 2015; Lu, Huang et al., 2021). In addition, we also obtained the measure of the author community by an average author count per publication and chose the common *h*-index method (Hirsch, 2005) to compute author impact; the details of the calculation approach are given in Appendix B in the Supplementary material. Thus, we utilized *cumulative # authors*, *annual avg. authors*, *avg. authors per pub*, and *author avg. impact* to calculate the community size and impact of the authors.

3.2.3. Institution dimension

As per the relationships shown in Figure 2, a research institution symbolizes a large research community, containing many talents, equipment, and other research resources, which invariably influences academic development, and an academic publication means the crystallization of an institution's wisdom (Kahn, 2011). The reputation of an institution also influences the growth of a research topic (Hottenrott, Rose & Lawson, 2021). Academic institutions are also essential objects in scientometrics research (Ellegaard & Wallin, 2015; Yegros-Yegros, Capponi & Frenken, 2021). In our study, we calculated the number of institutions to depict the growth situation of new knowledge, which was divided into two indicators: *cumulative # institutions* and *annual avg. institutions*.

3.2.4. Funding dimension

Generally, relatively large investments indicate that a prominent impact is expected (Álvarez-Bornstein & Bordons, 2021). The amount of funding could also cast light on the development prospects of new knowledge. Thus, early indications of knowledge growth may be revealed from the analysis of funding data. Although the coverage of funding data remains limited (Hopkins & Siepel, 2013), we extracted the usage information of funding data as reported by authors in the acknowledgments section of scholarly publications. The funding support signifies the development and energy of new knowledge through the second-order relationship between funding and knowledge outputs produced. We adopted *cumulative # funding* and *annual avg. funding* in understanding knowledge role transition.

3.2.5. Journal dimension

Journal information was also utilized to measure the importance and development potential of research topics (Moed, 2010). Mainly, the quantity and impact of scholarly publications have a positive correlation with the journal impact factor (Dinesh, 2017). Peset, Garzón-Farinós et al. (2020) found that journal impact has a significant effect on the survival time of author keywords, which is an important perspective to investigate knowledge growth. Thus, we speculate that journal impact reflects the reliability and recognition of new knowledge in the specific domain. In this study, the comprehensive journal impact factor, *cumulative # journals*, and *annual avg. journals* were considered as the indicators of journal venue. The journal impact factor indicates the domain recognition of the journal, and the journal number signifies the popularity of knowledge entities.

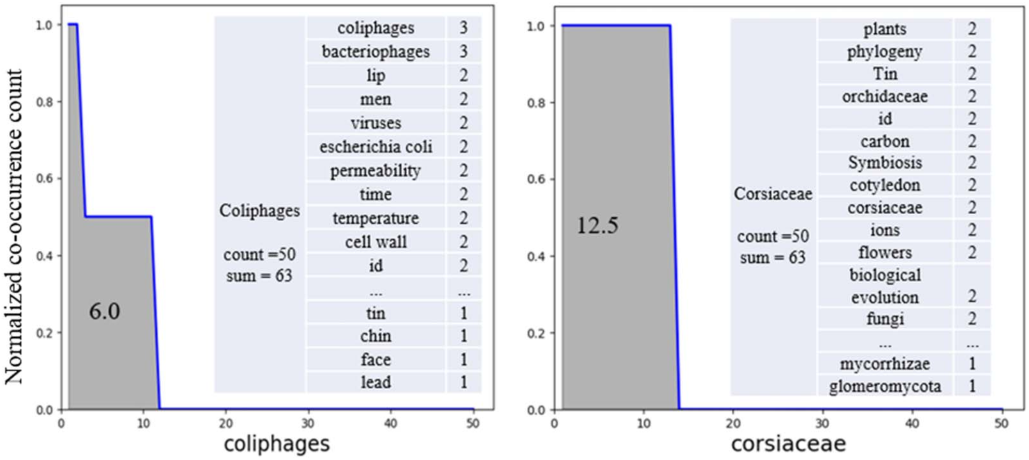

Figure 3. Explanation of the comprehensive calculation of the semantic specificity.

3.2.6. Descriptor dimension

Knowledge relatedness is a key indicator to represent semantic specificity (Breschi, Lissoni & Malerba, 2003). We attempt to measure the semantic specificity of one knowledge entity through the distribution of co-occurrence counts with other knowledge entities. The idea is similar to the Gini coefficient (Gini, 1997) which was utilized to demonstrate a degree of inequality of distribution in bibliometric studies (Cockriel & McDonald, 2018; Leydesdorff, Wagner & Bornmann, 2019; Nuti, Ranasinghe et al., 2015). Referring to the Gini coefficient, we calculated the integral area of the Lorenz curve to express knowledge relatedness. If a new knowledge entity jointly occurs with others more often, it has lower semantic specificity in the specific domain. For example, the "*coliphages*" and "*corsiaceae*" cases with the same count have different semantic specificity, as per the comparison examples shown in Figure 3.

In Figure 3, the *x*-axis represents the order of other knowledge entities, and the *y*-axis corresponds to the normalized co-occurrence count of knowledge entity pairs. The specific calculation process is shown in Appendix B in the Supplementary material.

Based on the above analysis, we can divide the variables into cumulative and temporal variables to explore the two research problems of transition possibility and transition pace. The analysis framework for the knowledge role transition is shown in Figure 4.

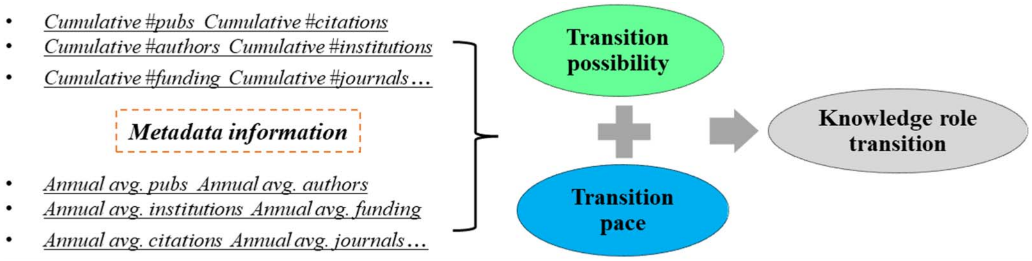
Cumulative #pubs Cumulative #citations
Cumulative #authors Cumulative #institutions
Cumulative #funding Cumulative #journals…
Metadata information
Annual avg. pubs Annual avg. authors
Annual avg. institutions Annual avg. funding
Annual avg. citations Annual avg. journals…
Transition possibility
Transition pace
Knowledge role transition

Figure 4. Analysis framework of knowledge role transition.

4. DATA

In this study, the whole PubMed XML data set, which is an essential literature resource for the medicine domain, was parsed. The 30,376,130 scientific publications were collected up to 2019 from the PubMed data set. The metadata information of these scientific publications has enriched characteristics for exploring knowledge transition. A scholarly publication is associated with various metadata, such as *author*, *institution*, *citation*, *journal*, and *keywords*. The essential data acquisition and processing are shown in Figure 5.

4.1. Data Collection

The acquisition of multidimension data is performed around the publications in PubMed, which includes an amount of metadata information. First, because the citation relationship of PubMed is incomplete, we obtained the citation data of Web of Science (WoS) to make up for missing citation relationships in PubMed (Xu, Kim et al., 2020). Second, the journal information is collected from the SJR website (Scimago Journal & Country Rank), which includes SJR for evaluating journal impact (Guerrero-Bote & Moya-Anegón, 2012), *h*-index, Cites/Doc. (2 years), etc.

Finally, as a domain knowledge base, the MeSH thesaurus was parsed to gain descriptors that were adopted by domain experts. We randomly selected a version of the MeSH thesaurus in a recent 5-year period. MeSH includes three items of data: descriptors (subject heading), qualifiers, and supplementary concept records. Among these, descriptors are divided into 16 trees and, as of 2015, they number 27,885 descriptors. We took MeSH descriptors as formal knowledge.

4.2. Data Reprocessing

More importantly, the twofold data items need to be preprocessed L: author name disambiguation and knowledge match. Author name disambiguation is to resolve the problem of author consistency. Knowledge match is to establish the relationship between knowledge entities and papers.

4.2.1. Author name disambiguation

Author name disambiguation is a general method for identifying unique authors in some studies. According to our investigation, Author-ity (Torvik & Smalheiser, 2009) and Semantic Scholar (Ammar, Groeneveld et al., 2018) are two high-quality data sets. Keeping in mind that

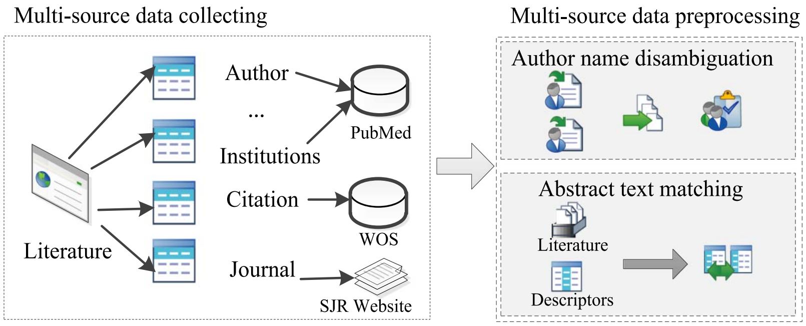

Figure 5. The main process of data collecting and reprocessing; WoS (Web of Science), SJR (Scimago Journal & Country Rank).

the Author-ity data set has a higher F1 score (98.16%) than the Semantic Scholar data set (Xu et al., 2020), this process was to select the author's unique ID from the Author-ity data set as the primary unique identifier according to the proven strategy. However, it is limited by the time range, which only contains PubMed papers before 2009. Thus, authors after 2009 were supplemented by using the author name disambiguation results of Semantic Scholar.

4.2.2. Knowledge match

The abstracts of scientific literature usually emphasize research contributions (Bu, Li et al., 2021). The knowledge that appears in the abstract almost completely describes the core content of a scholarly publication. Knowledge match can build a bridge between fine-grained knowledge and scholarly publication. In this study, the formal knowledge is from the preferred concepts in the MeSH thesaurus, so the abstract of scholarly publication can be annotated by the formal knowledge using sequence matching algorithms. We found 372,899,456 matching records from 30,376,130 publications, which were integrated with the original annotated records in the PubMed database. Therefore, the scholarly publications could be conveniently retrieved by matching and annotating records of formal knowledge.

4.3. Empirical Data Construction

Formal knowledge is usually kept in the form of thesauri based on ontology. Medical Subject Headings (MeSH) was developed by the National Library of Medicine (NLM), which is a controlled vocabulary for indexing scholarly publications in the PubMed database (Liu, Peng et al., 2015). The domain-specific structured ontology describes what occurs in each domain and is usually considered as a domain knowledge base with specific hierarchies to represent concepts and relations (Nayak, Dutta et al., 2019). The transition time of formal knowledge is generally concentrated in the time interval of one to 40 years (see Figure 6), with an average value of 35 years. To collect records of the knowledge transition process, we need to match the formal knowledge with the abstracts of scholarly publications before the transition year. After mapping the existing literature to descriptors, we obtained 17,639 descriptors, which cover 63.3% of the total number of descriptors (27,885).

As some MeSH descriptors have a structure distinct from the usual keywords (i.e., author-keywords; Valderrama-Zurián, García-Zorita et al., 2021), biomedical entities extraction is essential to discover informal knowledge in scholarly publications. We selected BERN (Kim, Lee et al., 2019), which learned the descriptor composition features of MeSH descriptors, to extract biomedical knowledge entities. Keeping in mind that the character length scope of

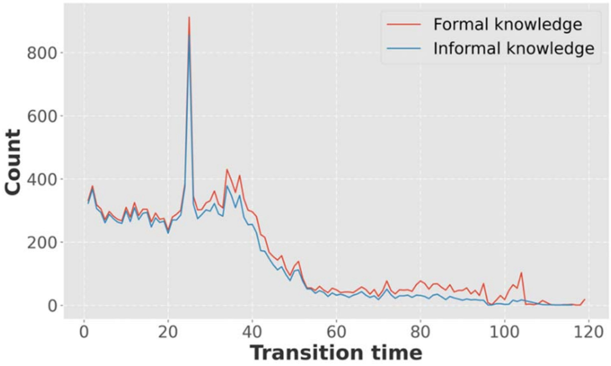

Figure 6. The samples' distribution of formal knowledge and informal knowledge.

knowledge entities from BERN is from 1 to 145, whereas the that of MeSH descriptors is from 2 to 45, we therefore restricted the word length scope of the experimental and control groups. Knowledge entities that were adopted into the MeSH ontology were considered formal knowledge, and other knowledge entities with the same year of debut and word length scope (from 2 to 45) were chosen as informal knowledge to constitute matched-pair samples with formal knowledge. The 17,639 MeSH descriptors were randomly matched with 3,449,589 informal knowledge entities as the control group. The samples' distribution of the formal knowledge and informal knowledge over the transition time is shown in Figure 6. As the MeSH thesaurus adopted more knowledge entities in 1999 than in other years, the spike in Figure 6 contains mainly the 761 knowledge entities adopted in 1999, which have 25 years of data.

5. RESULTS

5.1. Transition Possibility Analysis

5.1.1. Matched-pair statistical analysis

Kolmogorov–Smirnov tests were conducted for the distribution of *cumulative # pubs*, *cumulative # citations*, *cumulative # authors*, *cumulative # institutions*, *cumulative # funding*, *cumulative # journals*, *journal avg. impact*, *author avg. impact*, and *knowledge relatedness* variables between treatment and control groups. The tests revealed that the distribution of these variables does not follow the normal distribution ($p < 0.001$). Thus, the Wilcoxon signature rank test was performed to verify the differences in the distribution of these variables between formal and informal knowledge. Most tests showed that statistically significant differences between formal and informal knowledge are less than the 0.001 level, except the test of *journal avg. impact*, as shown in detail in Table 2.

To observe the significant difference between formal and informal knowledge, we respectively provided the box plots for these variables: *cumulative # pubs*, *cumulative # citations*, *cumulative # authors*, *author avg. impact*, *cumulative # institutions*, *cumulative # funding*, *cumulative # journals*, *journal avg. impact*, and *knowledge relatedness*, shown in Figure 7. The test results imply that *cumulative # pubs*, *cumulative # citations*, *cumulative # authors*, *cumulative # institutions*, *cumulative # funding*, *cumulative # journals*, *author avg. impact*, and *knowledge relatedness* could effectively distinguish formal knowledge from informal knowledge, but *journal avg. impact* is not significantly discriminative for formal knowledge and informal knowledge. Even the statistically significant difference of the *author avg. impact* variable is less than the 0.001 level, and the medians of the treatment and control groups are similar. Thus, a more in-depth comparative analysis is necessary for *cumulative # pubs*, *cumulative # citations*, *cumulative #authors*, *cumulative # institutions*, *cumulative # funding*, *cumulative # journals*, *author avg. impact*, and *knowledge relatedness variables*, and notably *author avg. impact*.

5.1.2. Differentiation analysis over transition time

The test statistics analysis only presents the discriminatory power of the variables from the perspectives of the overall distribution and does not reflect the specific characteristics of each variable. Keeping in mind that different knowledge entities may have different transition time intervals, we calculated the average values of these variables over transition time. This could highlight the performance of these variables in terms of discriminating knowledge entities with different transition time intervals. We visualized the distributions of average values of these variables over transition time, as shown in Figure 8.

Table 2. Wilcoxon signed-rank test results for cumulative variables

Variables	*Cumulative #pubs*	*Cumulative #citations*	*Cumulative #authors*	*Cumulative #institutions*	*Cumulative #funding*	*Cumulative #journals*	*Journal avg. impact*	*Author avg. impact*	*Knowledge relatedness*
Sig.	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.115	<0.001	<0.001
Treatment Median	134	61.14	489.0	106.0	10.0	67.0	1.27	10.6	15.3
Control Median	32	15.18	124.0	23.0	5.0	22.0	1.28	11.3	18.6

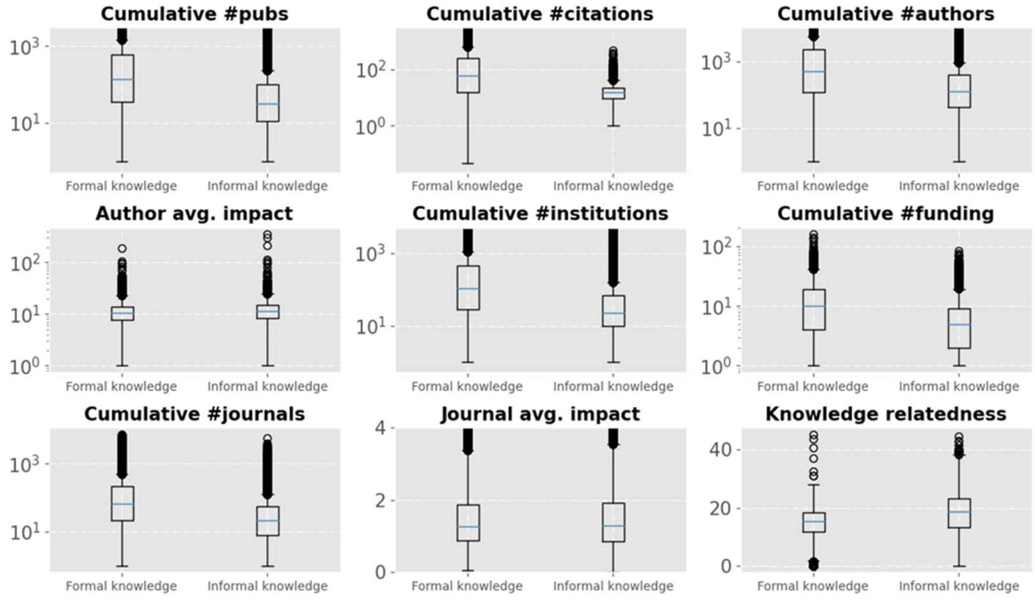

Figure 7. The boxplots of matched-pair analysis between formal and informal knowledge. Each dot represents one knowledge entity that includes formal and informal knowledge.

Figure 8 indicates that *author avg. impact* and *journal avg. impact* do not show the distinguishing effects on formal knowledge and informal knowledge, which implies that the *author avg. impact* and *journal avg. impact* variables cannot determine the transition from informal knowledge to formal knowledge. Moreover, the average values of *cumulative # pubs*, *cumulative # citations*, *cumulative # authors*, *cumulative # institutions*, *cumulative # funding*, and *cumulative # journals* of formal knowledge are not less than those of informal knowledge. This suggests that these variables could be utilized to distinguish between formal knowledge

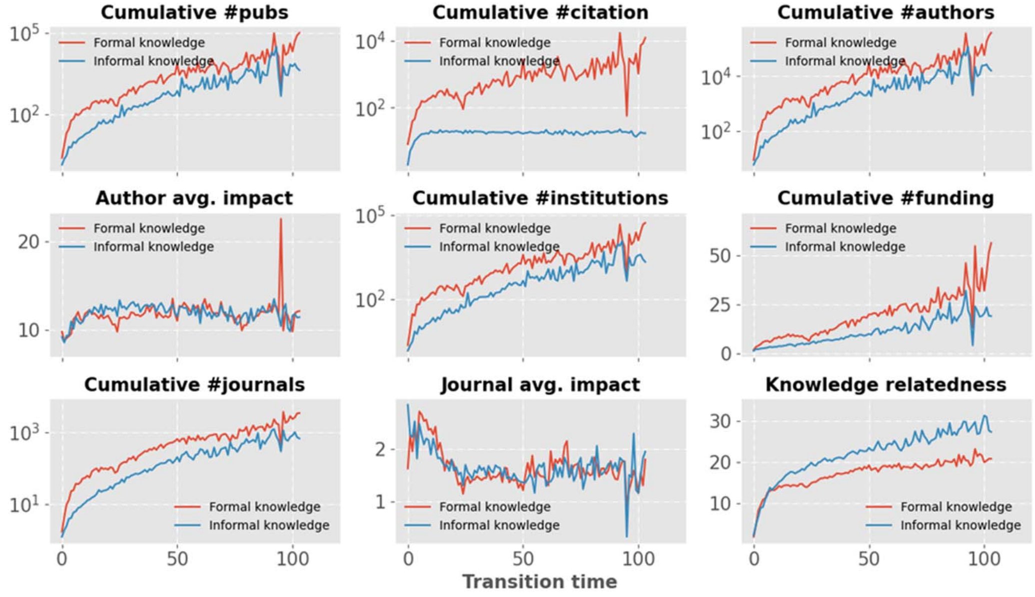

Figure 8. Distribution of average values of cumulative variables over transition time.

and informal knowledge. Interestingly, only the *knowledge relatedness* values of informal knowledge are not less than those of formal knowledge. It could be understood that the smaller the value of *knowledge relatedness*, the more specific the semantics of the knowledge and the more likely it is to be adopted as formal knowledge. What is more, the *cumulative # pubs*, *cumulative # authors*, *cumulative # institutions*, and *cumulative # journals* variables could more effectively distinguish formal knowledge from informal knowledge with a short transition time than that with a long transition time. In contrast, the space between the blue and red curves of *cumulative # citations*, *cumulative # funding*, and *knowledge relatedness* in the bigger transition time intervals seems to be relatively larger than that in the smaller transition time intervals. This means that the *cumulative # citations*, *cumulative # funding*, and *knowledge relatedness* variables have stronger distinguishable effects on knowledge entities with a long transition time, different from the *cumulative # pubs*, *cumulative # authors*, *cumulative # institutions*, and *cumulative # journals* variables.

Furthermore, it is important to note that, in Figure A-1 in Appendix A in the Supplementary material, we provided scatter plots of the knowledge entities in the coordinates of each variable versus transition time to provide a comprehensive response to the differentiation performance of each variable. Combined with the above analysis, we suppose that those multidimensional variables could improve the performance for distinguishing the formal and informal knowledge overall transition times. Likewise, it is essential to understand knowledge growth and transition patterns from multidimensions.

5.2. Transition Pace Analysis

5.2.1. Static correlation analysis

Transition time is the interval of time that elapses from informal knowledge to formal knowledge. Thus, we conducted an analysis of the correlation between temporal variables and transition time to explore the influence factors of transition pace. Because some samples have shorter transition times and the values of cumulative variables increase with the consumption of transition time, the correlation coefficients were calculated respectively based on the first 5 years and 10 years of history data from informal knowledge to formal knowledge, as shown in Table 3.

Intuitively, most values of temporal variables are negatively correlated with transition time, except *journal avg. impact*. It is easy to understand that the more annual publications, citations, contributing authors, institutions, journals, and supporting funds, the stronger the authors' impact, and the faster new knowledge could be adopted into the domain knowledge base. Specifically, *annual avg. funding*, *annual avg. knowledge relatedness*, *avg. authors per pub*, *journal avg. impact*, and *author avg. impact* are not less than 0.4, whereas *annual avg. pubs*, *annual avg. authors*, *annual avg. institutions*, *annual avg. citations*, and *annual avg. journals* are less than 0.3.

Furthermore, we calculated the mean values of *annual avg. knowledge relatedness*, *avg. authors per pub*, *journal avg. impact*, and *author avg. impact* corresponding to each transition time interval. Figure 9(a) shows the distribution of *avg. authors per pub* over transition time: The mean value of *avg. authors per pub* gradually decreases in the smaller time intervals, whereas it has an overall upward trend in the bigger time intervals. This implies that the *avg. authors per pub* variable has different effects on formal knowledge with different transition time intervals. Figure 9(b) indicates that *author avg. impact* has an overall decreasing trend over transition time, which is a robust impact element on transition pace. Figure 9(c) suggests that the values of *annual avg. knowledge relatedness* drop rapidly and then slowly, which is

Table 3. The correlation coefficients between temporal variables and transition time

Temporal variables	*Annual avg. pubs*	*Annual avg. authors*	*Annual avg. institutions*	*Annual avg. funding*	*Annual avg. citations*	*Annual avg. journals*	*Annual avg. knowledge relatedness*	*Avg. authors per pub*	*Journal avg. impact*	*Author avg. impact*
Correlation coefficients of the first 5 years	−0.18***	−0.16***	−0.13***	−0.40***	−0.16***	−0.25***	−0.47***	−0.48***	0.60***	−0.43***
Correlation coefficients of the first 10 years	−0.20***	−0.17***	−0.16***	−0.40***	−0.17***	−0.25***	−0.50***	−0.52***	0.64***	−0.49***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

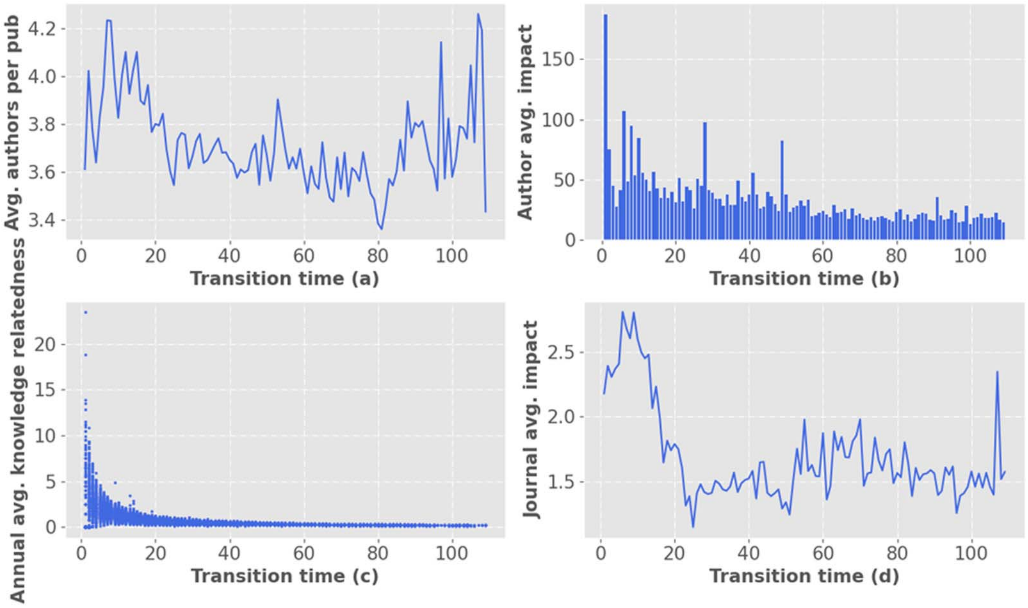

Figure 9. The distribution of variable values of four examples. Each blue dot represents one formal knowledge entity.

more effective for characterizing the transition pace of formal knowledge with short to medium transition times. Figure 9(d) suggests that the values of *journal avg. impact* decrease sharply and then stay relatively stable. In summary, the values of one variable may not always have a strong correlation with transition time in the whole span, even if there are opposite trends in the first and second half transition time intervals. Thus, multidimensional variables are essential to provide a comprehensive description of the pace at which knowledge is adopted.

5.2.2. Dynamic correlation analysis

To further explore the dynamic correlation between temporal variables and transition time, we calculated the correlation coefficients consecutively in each span of history data. As the average value of transition time of all formal knowledge is 35 years, we respectively calculated the correlation coefficients of temporal and cumulative variables in the first 35 spans of history data. In the static correlation analysis, most correlation coefficients, except for *journal avg. impact*, are negative numbers in the first 5 and 10 years of history data. To visualize this, the correlation coefficients of these variables were taken as negative, namely zero minus the correlation coefficients, to obtain positive values. In addition, to highlight the performance of the temporal variables, we also calculated the dynamic correlation of the cumulative variables. The results of the dynamic correlation are a comparative analysis of temporal and cumulative variables, including the *journal avg. impact*, *avg. authors per pub*, and *author avg. impact* variables, are shown in Figure 10.

Comparing these variables, the correlation coefficients of *annual avg. pubs*, *annual avg. authors*, *annual avg. citations*, *annual avg. institutions*, and *annual avg. journals*, respectively are firstly lower and then higher than those of *cumulative # pubs*, *cumulative # authors*, *cumulative # citations*, *cumulative # institutions*, and *cumulative # journals*. However, the correlation coefficients of *annual avg. funding* and *annual avg. knowledge relatedness* are always higher than those of *cumulative # funding* and *knowledge relatedness*. The maximum correlation values of *avg. authors per pub*, *author avg. impact*, *annual avg. funding*, *journal avg. impact*, *annual avg. knowledge relatedness*, and *knowledge relatedness* are all bigger than

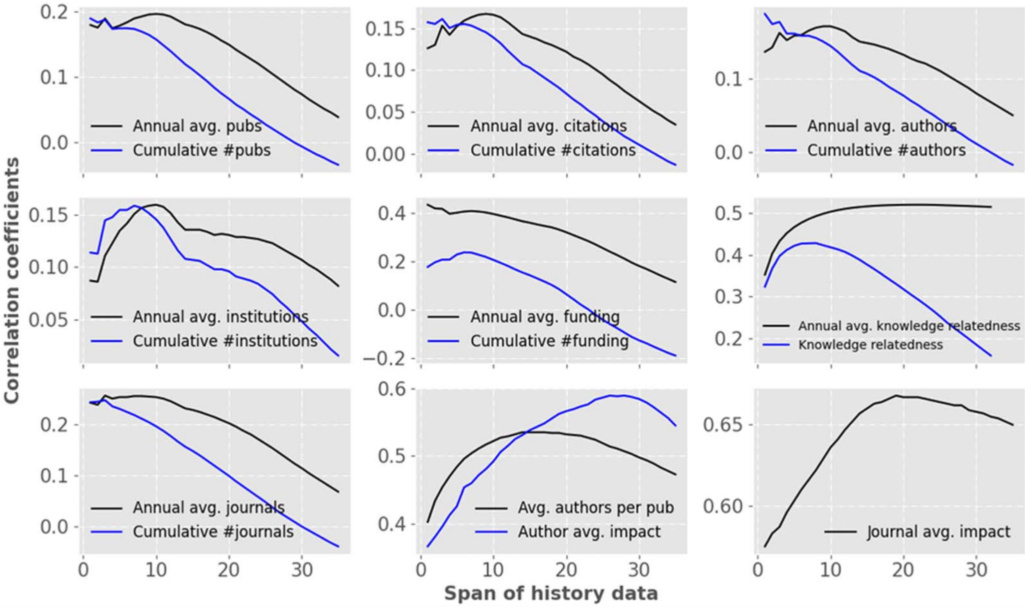

Figure 10. Dynamic correlation values of temporal and cumulative variables.

0.4, respectively 0.53, 0.59, 0.43, 0.61, 0.51, and 0.43. Thus, the characteristics at the publication, citation, and institution levels play a few roles in characterizing the pace of knowledge transition.

At the author level, the values of *avg. authors per pub* and *author avg. impact* are more correlated with transition time: The curves of the two variables both increase at first and then decrease. Specifically, *avg. authors per pub* has a maximum correlation value in the 14th span of history data (0.535), and the maximum correlation value of *author avg. impact* is in the 26th span (0.589). The analysis results indicate that the *avg. authors per pub* and *author avg. impact* variables could be utilized to characterize the pace of knowledge transition. This suggests the higher the average number of authors per paper, the greater the impact of the authors, and the earlier new knowledge could receive attention from domain researchers.

At the funding level, the *annual avg. funding* variable has its maximum correlation value in the year of debut, whereas the curve of *cumulative # funding* reaches its peak in the first six years of history data. The correlation maximum of *cumulative # funding* is 0.237 while that of *annual avg. funding* is 0.433. The *annual avg. funding* variable is more correlated with transition time than *cumulative # funding*. This implies that early funding support is more important for improving the transition pace.

At the descriptor level, *annual avg. knowledge relatedness* increases rapidly and then stays stable, and *knowledge relatedness* also increases and then gradually decreases. The curve of *annual avg. knowledge relatedness* reaches an inflection point in the 10-year span, the correlation value of which is 0.5. Looking at the distribution of *knowledge relatedness* over transition time, the *knowledge relatedness* variable has a correlation maximum of about 0.43 at the 7-year span. The curve of *annual avg. knowledge relatedness* is always above that of *knowledge relatedness*, which suggests that *annual avg. knowledge relatedness* has a stronger correlation with transition time, which reveals the intrinsic and implicit interaction law of knowledge transition. Thus, the relatedness degree with other knowledge is an important variable for characterizing the pace of knowledge transition.

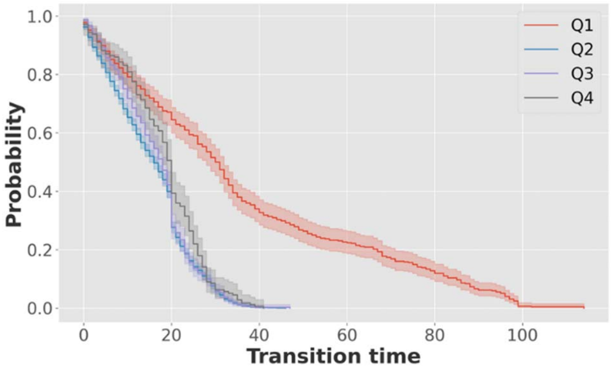

Figure 11. Survival analysis of formal knowledge published in different journals.

Particularly at the journal level, the "*SJR Quartile*" was taken to delineate the levels of journal impact. "*SJR Quartile*" indicates the quartile to which a given journal belongs according to its impact. *Quartile 1* (*Q1*) is the highest impact score and *Quartile 4* (*Q4*) is the lowest. Formal knowledge is mainly derived from the new knowledge that appears in high-impact journals. After statistical analysis, we found that formal knowledge originated from *Q1 journals* (14,359), *Q2 journals* (2,532), *Q3 journals* (154), and *Q4 journals* (18). This indicates that the new knowledge entities that appear in journals with high impact factors are more likely to be transformed into formal knowledge. In terms of transition pace, the journal number gradually decreases while the journal impact gradually increases, as shown in Figure 10. This suggests that formal knowledge with a shorter transition time has a bigger journal number but a weaker journal impact. To further understand the effect of journal impact on transition pace, the transition year is regarded as the "*time of death*" (i.e., the time at which the event occurred). We grouped formal knowledge by the initial journal impact to conduct survival analysis (González, García-Massó et al., 2018). Figure 11 shows that some descriptors from *SJR Q1* have a longer survival time, with their fold at the top, and some from *SJR Q2* have the shortest survival time, with their fold always at the bottom in the short intervals of transition time. This means that informal knowledge with a high journal impact is more likely to transition into formal knowledge, whereas a lower journal impact level than *SJR Q1* improves the pace of transition.

6. DISCUSSION

We aim to understand the knowledge role transition from a perspective of knowledge codification, which influences the speed of knowledge creation: innovation. In this section, we discuss the implications of major findings to current theories, which may inform new research challenges and ideas for future study.

Codified knowledge is beneficial in facilitating the creation, circulation, and reconstitution of knowledge. Knowledge codification is inherently a complex process, which is influenced by a variety of external and internal factors. Knowledge being codified could be understood as its role transition in our work. Theoretically, because each change in scientific knowledge is typically a reaction to scholarly publications, the characteristics of explicit and implicit scholarly publications could largely characterize the transition process of formal knowledge. In terms of practice, formal knowledge is represented as thesauri or ontologies. When a new knowledge entity is adopted into a thesaurus, it transitions into formal knowledge. In previous studies, scholars have explored the evolutionary pattern of scientific knowledge growth by dividing it into life cycle stages (Shimogawa et al., 2012). Accordingly, knowledge role transition also belongs to a phenomenon in the process of knowledge growth and evolution, like the process of awakening sleeping knowledge (Yang, Bu et al., 2022).

To explore more internal and external elements that influence knowledge growth and evolution, multidimension data were collected and different variables were developed by taking advantage of the large-scale metadata information of scholarly publications (Sharma & Khurana, 2021; Wang, 2018; Weis & Jacobson, 2021). These studies found that the metadata information could reveal the evolution of scientific knowledge growth, but the results varied for specific growth phenomena. For instance, the cumulative citation was a better signal for identifying impactful research 5 years after publication (Weis & Jacobson, 2021), whereas the *cumulative #citations* variable is better at distinguishing knowledge entities with a long transition time in our work.

The *transition possibility* is defined to describe whether informal knowledge could transition into formal knowledge. This phenomenon could be regarded as a state change in the process of knowledge growth and evolution. We find that the cumulative variables from metadata information have a better effect on revealing this phenomenon. Specifically, *cumulative # pubs*, *cumulative # authors*, and *cumulative # journals* could better distinguish between formal knowledge and informal knowledge with a short transition time, whereas *cumulative # citations*, *cumulative # funding*, and *knowledge relatedness* are better at distinguishing knowledge entities with a long transition time. As for *cumulative # pubs*, *cumulative # authors*, and *cumulative # journals*, their numbers were fixed as soon as the literature was published, so these indicators therefore tend to differentiate between knowledge entities with short transition times. *Cumulative # citations*, *cumulative # funding*, and *knowledge relatedness* indicators have a significant delayed effect (Mariani, Medo, & Zhang, 2016): It generally takes a few years for them to emerge with an edge. The *cumulative # funding* variable could differentiate formal knowledge from informal knowledge: Formal knowledge has a higher median value of *cumulative # funding* than informal knowledge in Table 2. This finding indicates that funded knowledge has a higher potential for development than nonfunded knowledge, which is consistent with the results of the latest study (Mosleh, Roshani, & Coccia, 2022). Further, our study explores the factors influencing the growth and evolution of scientific knowledge from a more microscopic perspective and finds that metadata variables have their scope of applicability.

In terms of transition pace, temporal variables are better suited to describe the pace of knowledge transition. The *author avg. impact* variable is more correlated with transition time. Namely, the correlation coefficient of the first 10 years is 0.49 and the maximum is 0.59 at the 0.0001 level. High-impact authors have a leadership effect that attracts more followers (Bu, Ding et al., 2018), and knowledge entities that gain more attention are more likely to be codified, which contributes to the transition from informal knowledge to formal knowledge. In addition, it is important to note that only the *journal avg. impact* variable has a positive correlation with transition time, whereas the others are negatively correlated with transition time. We further find that some knowledge entities from *SJR Q1* take a longer time to be formal knowledge and some from *SJR Q2* need a shorter time, which is consistent with the results that *SJR Q2* shows a longer average survival time than those from *SJR Q1* (Peset et al., 2020). The two results both indicate that the knowledge entities or keywords from the *SJR Q2* journal are of greater concern to researchers. Most surprising is that the smaller knowledge relatedness value improves the possibility of transition, whereas the bigger *knowledge relatedness* or *annual avg. knowledge relatedness* value improves the transition pace of formal knowledge. According to the calculation approach of knowledge relatedness in Appendix B in the Supplementary material, the broader the semantics of a knowledge entity, the more other knowledge entities co-occur with it, and the more balanced the distribution of co-occurrence.

The more convergent the semantics the more likely it is to be codified, and the broader the semantics the faster the role transition. In the future, the balance point of transition pace and transition possibility is an interesting research problem for scientometrics.

7. CONCLUSION

Knowledge role transition is a highly complex process that is influenced by a variety of external and internal factors. By analyzing the large-scale metadata of publications in PubMed, we found that cumulative variables (i.e., *cumulative # pubs*, *cumulative # authors*, *cumulative # institutions*, and *cumulative # journals*) tended to predict formal knowledge with short transition times, whereas *cumulative # citations*, *cumulative # funding*, and *knowledge relatedness* distinguish those with long transition times. The temporal variables (i.e., *avg. authors per pub*, *author avg. impact*, *annual avg. funding*, *journal avg. impact*, and *annual avg. knowledge relatedness*) are more correlated with transition time. Specifically, early funding support is more important for improving the transition pace, notably in the year of debut. Journal impact has a positive correlation with the transition possibility but a negative correlation with transition pace. The weaker knowledge relatedness raises the transition possibility, whereas the stronger knowledge relatedness improves the transition pace.

This study has significant theoretical and practical implications regarding knowledge codification and role transition. In theoretical terms, it helps to better understand the implicit and dynamic patterns of knowledge codification. In practical terms, the findings concerning knowledge role transition patterns will help maintenance experts to update the terms or concepts of thesauri by recommending credible and valuable new domain knowledge entities. The thesaurus is also an important indexing tool in information retrieval system: The new domain knowledge detection benefits the automatic annotation of scientific literature, which improves the performance of the academic retrieval system. More importantly, understanding knowledge role transition allows us to learn from the past to improve the ability to detect knowledge innovation in the future. Overall, these findings are of great significance for domain knowledge management and early detection of credible and valuable knowledge.

However, there are some potential limitations in our work. First, because we focus on biomedical publications, these findings may not generalize to other disciplines. Second, to reduce the complexity of the calculation, we selected preferred concepts of MeSH to represent knowledge entities. Third, we have only taken a descriptive statistical analysis of the indicators set in the process of knowledge role transition but not an in-depth analysis of the underlying mechanisms. In future work, we should semantically encode knowledge entities and investigate the patterns of knowledge role transition in other disciplines.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers and the journal editors for their constructive suggestions on this paper. This research has been supported by Major Projects of the National Social Science Foundation of China (No. 17ZDA292).

AUTHOR CONTRIBUTIONS

Jinqing Yang: Conceptualization, Methodology, Writing—Original draft, Writing—Review & editing. Wei Lu: Conceptualization, Methodology, Supervision. Yong Huang: Conceptualization, Supervision, Writing—Review & editing. Qikai Cheng: Data curation, Formal analysis. Zhang Li: Data curation. Shengzhi Huang: Data curation.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work is supported by Major Projects of the National Social Science Foundation of China (No. 17ZDA292).

DATA AVAILABILITY

The underlying data is available in Figshare (Yang, 2022).

REFERENCES

Ahmadyousefi, R., Choobchian, S., Chizari, M., & Azadi, H. (2020). The role of knowledge management in the development of drought crisis management programmes. *Knowledge Management Research & Practice*, *20*(2), 177–190. https://doi.org/10.1080/14778238.2020.1832871

Álvarez-Bornstein, B., & Bordons, M. (2021). Is funding related to higher research impact? Exploring its relationship and the mediating role of collaboration in several disciplines. *Journal of Informetrics*, *15*(1), 101102. https://doi.org/10.1016/j.joi.2020.101102

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., … Etzioni, O. (2018). Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the NAACH-HLT 3* (pp. 84–91). https://doi.org/10.18653/v1/N18-3011

Amjad, T., Rehmat, Y., Daud, A., & Abbasi, R. A. (2020). Scientific impact of an author and role of self-citations. *Scientometrics*, *122*(2), 915–932. https://doi.org/10.1007/s11192-019-03334-2

Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, *32*(1), 69–87. https://doi.org/10.1016/S0048-7333(02)00004-5

Bu, Y., Li, M., Gu, W., & Huang, W. B. (2021). Topic diversity: A discipline scheme-free diversity measurement for journals. *Journal of the Association for Information Science and Technology*, *72*(5), 523–539. https://doi.org/10.1002/asi.24433

Bu, Y., Ding, Y., Liang, X., & Murray, D. S. (2018). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology*, *69*(3), 438–448. https://doi.org/10.1002/asi.23966

Cardoso, S. D., Da Silveira, M., & Pruski, C. (2020). Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies. *Knowledge-Based Systems*, *194*, 105508. https://doi.org/10.1016/j.knosys.2020.105508

Cardoso, S. D., Pruski, C., & Da Silveira, M. (2018). Supporting biomedical ontology evolution by identifying outdated concepts and the required type of change. *Journal of Biomedical Informatics*, *87*, 1–11. https://doi.org/10.1016/j.jbi.2018.08.013, PubMed: 30205172

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics*, *115*(1), 35–49. https://doi.org/10.1007/s11192-018-2654-5

Cockriel, W. M., & McDonald, J. B. (2018). The influence of dispersion on journal impact measures. *Scientometrics*, *116*(1), 609–622. https://doi.org/10.1007/s11192-018-2755-1

Cohendet, P., & Meyer-Krahmer, F. (2001). The theoretical and policy implications of knowledge codification. *Research Policy*, *30*(9), 1563–1591. https://doi.org/10.1016/S0048-7333(01)00168-8

Dinesh, K. S. (2017). Ranking of arts and humanities journals published in India: A scientometric analysis. *Pearl: A Journal of Library and Information Science*, *11*(2), 155–158. https://doi.org/10.5958/0975-6922.2017.00021.3

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). Author ranking evaluation at scale. *Journal of Informetrics*, *12*(3), 679–702. https://doi.org/10.1016/j.joi.2018.06.004

Ebrahimi, F., Asemi, A., Shabani, A., & Nezarat, A. (2021). Developing a prediction model for author collaboration in bioinformatics research using graph mining techniques and big data applications. *International Journal of Information Science and Management*, *19*(2), 1–18. https://doi.org/10.21203/rs.3.rs-113236/v1

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105*(3), 1809–1831. https://doi.org/10.1007/s11192-015-1645-z, PubMed: 26594073

Faber, P. (2011). The dynamics of specialized knowledge representation: Simulational reconstruction or the perception–action interface. *Terminology*, *17*(1), 9–29. https://doi.org/10.1075/term.17.1.02fab

Fabian, G., Wächter, T., & Schroeder, M. (2012). Extending ontologies by finding siblings using set expansion techniques. *Bioinformatics*, *28*(12), i292–i300. https://doi.org/10.1093/bioinformatics/bts215, PubMed: 22689774

Gini, C. (1997). Concentration and dependency ratios. *Rivista di Politica Economica*, *87*, 769–792.

González, L. M., García-Massó, X., Pardo-Ibañez, A., Peset, F., & Devís-Devís, J. (2018). An author keyword analysis for mapping Sport Sciences. *PLOS ONE*, *13*(8), e0201435. https://doi.org/10.1371/journal.pone.0201435, PubMed: 30067822

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics*, *11*(2), 407–422. https://doi.org/10.1016/j.joi.2017.02.007

Guechtouli, W., & Kasmi, A. (2014). Knowledge transfer dynamics: How to model knowledge in the first place? *La Revue des Sciences Commerciales*, *20*(2), 7–27.

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, *6*(4), 674–688. https://doi.org/10.1016/j.joi.2012.07.001

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102, PubMed: 16275915

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American*

Society for Information Science, 46(6), 400–425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hopkins, M. M., & Siepel, J. (2013). Just how difficult can it be counting up R&D funding for emerging technologies (and is tech mining with proxy measures going to be any better)? *Technology Analysis & Strategic Management*, *25*(6), 655–685. https://doi.org/10.1080/09537325.2013.801950

Hottenrott, H., Rose, M. E., & Lawson, C. (2021). The rise of multiple institutional affiliations in academia. *Journal of the Association for Information Science and Technology*, *72*(8), 1039–1058. https://doi.org/10.1002/asi.24472

Iqbal, W., Qadir, J., Tyson, G., Mian, A. N., Hassan, S. U., & Crowcroft, J. (2019). A bibliometric analysis of publications in computer networking research. *Scientometrics*, *119*(2), 1121–1155. https://doi.org/10.1007/s11192-019-03086-z

Kahn, M. (2011). A bibliometric analysis of South Africa's scientific outputs—Some trends and implications. *South African Journal of Science*, *107*(1), 1–6. https://doi.org/10.4102/sajs.v107i1/2.406

Kaur, H., & Mahajan, P. (2015). Collaboration in medical research: A case study of India. *Scientometrics*, *105*(1), 683–690. https://doi.org/10.1007/s11192-015-1691-6

Kim, D., Lee, J., So, C. H., Jeon, H., Jeong, M., … Kang, J. (2019). A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, *7*, 73729–73740. https://doi.org/10.1109/ACCESS.2019.2920708

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, *13*(1), 255–269. https://doi.org/10.1016/j.joi.2018.12.006

Lissoni, F. (2001). Knowledge codification and the geography of innovation: The case of Brescia mechanical cluster. *Research Policy*, *30*(9), 1479–1500. https://doi.org/10.1016/S0048-7333(01)00163-9

Liu, D., Ray, G., & Whinston, A. B. (2010). The interaction between knowledge codification and knowledge-sharing networks. *Information Systems Research*, *21*(4), 892–906. https://doi.org/10.1287/isre.1080.0217

Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., & Zhu, S. (2015). MeSHLabeler: Improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*, *31*(12), i339–i347. https://doi.org/10.1093/bioinformatics/btv237, PubMed: 26072501

Lu, W., Huang, S., Yang, J., Bu, Y., Cheng, Q., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management*, *58*(4), 102594. https://doi.org/10.1016/j.ipm.2021.102594

Maio, C. D., Fenza, G., Loia, V., & Senatore, S. (2010). Knowledge structuring to support facet-based ontology visualization. *International Journal of Intelligent Systems*, *25*(12), 1249–1264. https://doi.org/10.1002/int.20451

Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, *10*(4), 1207–1223. https://doi.org/10.1016/j.joi.2016.10.005

Mistry, J., & Berardi, A. (2016). Bridging indigenous and scientific knowledge. *Science*, *352*(6291), 1274–1275. https://doi.org/10.1126/science.aaf1160, PubMed: 27284180

Mosleh, M., Roshani, S., & Coccia, M. (2022). Scientific laws of research funding to support citations and diffusion of knowledge in life science. *Scientometrics*, *127*(4), 1931–1951. https://doi.org/10.1007/s11192-022-04300-1, PubMed: 35283543

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002

Möller, M., Sintek, M., Buitelaar, P., Mukherjee, S., Zhou, X. S., & Freund, J. (2008). Medical image understanding through the integration of cross-modal object recognition with formal domain knowledge. *Proceedings of the First International Conference on Health Informatics* (pp. 134–141).

Nayak, G., Dutta, S., Ajwani, D., Nicholson, P., & Sala, A. (2019). Automated assessment of knowledge hierarchy evolution: Comparing directed acyclic graphs. *Information Retrieval Journal*, *22*(3), 256–284. https://doi.org/10.1007/s10791-018-9345-y

Naghavi, M., & Walsh, D. (2011). Learn from Ireland's knowledge economy. *Nature*, *476*(7361), 399. https://doi.org/10.1038/476399b, PubMed: 21866142

Nuti, S. V., Ranasinghe, I., Murugiah, K., Shojaee, A., Li, S. X., & Krumholz, H. M. (2015). Association between journal citation distribution and impact factor: A novel application of the Gini coefficient. *Journal of the American College of Cardiology*, *65*(16), 1711–1712. https://doi.org/10.1016/j.jacc.2014.12.071, PubMed: 25908079

Peset, F., Garzón-Farinós, F., González, L. M., García-Massó, X., Ferrer-Sapena, A., … Sánchez-Pérez, E. A. (2020). Survival analysis of author keywords: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology*, *71*(4), 462–473. https://doi.org/10.1002/asi.24248

Pór, G., & Molloy, J. (2000). Nurturing systemic wisdom through knowledge ecology. *The Systems Thinker*, *11*(8), 1–5.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Salatino, A. (2019). *Early detection of research trends*. Milton Keynes: Open University.

Sharma, K., & Khurana, P. (2021). Growth and dynamics of Econophysics: A bibliometric and network analysis. *Scientometrics*, *126*(5), 4417–4436. https://doi.org/10.1007/s11192-021-03884-4

Shimogawa, S., Shinno, M., & Saito, H. (2012). Structure of S-shaped growth in innovation diffusion. *Physical Review E*, *85*(5), 056121. https://doi.org/10.1103/PhysRevE.85.056121, PubMed: 23004835

Sice, P. V., Thirkle, S. A., & Ogwu, S. A. (2018). MIKE: Management, information and knowledge ecology. *International Journal of Systems and Society*, *5*(1), 13–27. https://doi.org/10.4018/IJSS.2018010102

Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in *Technology Foresight*. *Scientometrics*, *85*(1), 65–79. https://doi.org/10.1007/s11192-010-0259-8

Sun, Y., & Latora, V. (2020). The evolution of knowledge within and across fields in modern physics. *Scientific Reports*, *10*(1), 12097. https://doi.org/10.1038/s41598-020-68774-w, PubMed: 32694516

Sutton, D. C. (2001). What is knowledge and can it be managed? *European Journal of Information Systems*, *10*(2), 80–88. https://doi.org/10.1057/palgrave.ejis.3000397

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3), 11. https://doi.org/10.1145/1552303.1552304, PubMed: 20072710

Tsatsaronis, G., Varlamis, I., Kanhabua, N., & Nørvåg, K. (2013). Temporal classifiers for predicting the expansion of medical subject headings. *International Conference on Intelligent Text Processing* *and Computational Linguistics* (pp. 98–113). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-37247-6_9

Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management*, *48*(2), 303–325. https://doi.org/10.1016/j.ipm.2011.07.006

Valderrama-Zurián, J. C., García-Zorita, C., Marugán-Lázaro, S., & Sanz-Casado, E. (2021). Comparison of MeSH terms and Key-Words Plus terms for more accurate classification in medical research fields. A case study in cannabis research. *Information Processing & Management*, *58*(5), 102658. https://doi.org/10.1016/j.ipm.2021.102658

van den Oord, A., & van Witteloostuijn, A. (2018). A multi-level model of emerging technology: An empirical study of the evolution of biotechnology from 1976 to 2003. *PLOS ONE*, *13*(5), e0197024. https://doi.org/10.1371/journal.pone.0197024, PubMed: 29795575

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology*, *69*(2), 290–304. https://doi.org/10.1002/asi.23930

Wang, X., Hamilton, H. J., & Bither, Y. (2005). *An ontology-based approach to data cleaning*. Regina, Canada: Department of Computer Science, University of Regina.

Weis, J. W., & Jacobson, J. M. (2021). Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology*, *39*, 1300–1307. https://doi.org/10.1038/s41587-021-00907-6, PubMed: 34002098

Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246–258. https://doi.org/10.1002/asi.10367

Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., ... Ding, Y. (2020). Building a PubMed knowledge graph. *Scientific Data*, *7*(1), 205. https://doi.org/10.1038/s41597-020-0543-2, PubMed: 32591513

Yang, J. (2022). Understanding knowledge role transitions: A perspective of knowledge codification [Data set]. *Figshare*. https://doi.org/10.6084/m9.figshare.21387936.v1

Yang, J., Bu, Y., Lu, W., Huang, Y., Hu, J., Huang, S., & Zhang, L. (2022). Identifying keyword sleeping beauties: A perspective on the knowledge diffusion process. *Journal of Informetrics*, *16*(1), 101239. https://doi.org/10.1016/j.joi.2021.101239

Yang, L., Li, K., & Huang, H. (2018). A new network model for extracting text keywords. *Scientometrics*, *116*(1), 339–361. https://doi.org/10.1007/s11192-018-2743-5

Yegros-Yegros, A., Capponi, G., & Frenken, K. (2021). A spatial-institutional analysis of researchers with multiple affiliations. *PLOS ONE*, *16*(6), e0253462. https://doi.org/10.1371/journal.pone.0253462, PubMed: 34185774