A Study on Big Knowledge and Its Engineering Issues

Ruqian Lu, Xiaolong Jin, Songmao Zhang, Meikang Qiu, *Senior Member, IEEE*, and Xindong Wu, *Fellow, IEEE*

Abstract—After entering the big data era, a new term of 'big knowledge' has been coined to deal with challenges in mining a mass of knowledge from big data. While researchers used to explore the basic characteristics of big data, we have not seen any studies on the general and essential properties of big knowledge. To fill this gap, this paper studies the concepts of big knowledge, big-knowledge system, and big-knowledge engineering. Ten massiveness characteristics for big knowledge and big-knowledge systems, including massive concepts, connectedness, clean data resources, cases, confidence, capabilities, cumulativeness, concerns, consistency, and completeness, are defined and explored. Based on these characteristics, a comprehensive investigation is conducted on some large-scale knowledge engineering projects, including the Fifth Comprehensive Traffic Survey in Shanghai, the China's Xia-Shang-Zhou Chronology Project, the Troy and Trojan War Project, and the International Human Genome Project, as well as the online free encyclopedia Wikipedia. We also investigate the recent research efforts on knowledge graphs, where they are analyzed to determine which ones can be considered as big knowledge and big-knowledge systems. Further, a definition of big-knowledge engineering and its life cycle paradigm is presented. All of these projects are accordingly checked to determine whether they belong to big-knowledge engineering projects. Finally, the perspectives of big knowledge research are discussed.

Index Terms—Big data, knowledge engineering, big data knowledge engineering, big knowledge, massiveness characteristics, big-knowledge system, big-knowledge engineering, life cycle

1 Introduction: Big Knowledge — The Post Big Data Era

As it was mentioned by Gartner's Research Vice-President, Svetlana Sicular, that if one claims he has 'a big problem' then it is often that what he really has is 'a big data problem' [1]. Big data problems include "technology capabilities to store and process unstructured data; to link data of various types, origins and rates of change; and to perform comprehensive analysis, which became possible for many, rather than for selected few" [1]. Sicular then claimed that "Don't expect inexpensive solutions, but expect cost-effective and appropriate answers to your problems" [1].

For most people, the first thing they want to do with big data is to mine knowledge from it. The challenges come from both its front end (i.e., data mining) and back end (i.e., data analytics). Regarding knowledge based systems, it has been recognized that the knowledge one wants is no more than just expert knowledge, no more than just knowledge legacy, but has to be mined from big data consisting of information traces left by human's social and private activities, which are immensely large. These are called autonomous data sources (e.g., the Web) in [2]. On the other hand, big data analytics requires the full set of data, instead of only a sampled subset. This makes lots of traditional statistical techniques out of use. Not only its growth speed is very fast, but also its data structures and semantics are subject to steady change during the growing process. There are often very complicated connections among the data elements and subsets, such that only local processing is far being enough to get clear insights of the whole. All these difficulties mentioned above led to the proposals of 3V [3], 4V or even nV ($n > 4$) properties of big data [4], [5].

The challenges regarding the back end are the questions of how to turn the massive and complex data into knowledge, into what kind of knowledge and how to use it in real applications, which is the ultimate goal of data mining. Thus, Wu et al. proposed the concept of Big data Knowledge Engineering (BigKE) [2]. Technically, it involves attaching semantics to big data and transforming them into fragmented knowledge (i.e., knowledge pieces mined from big data). This is the first stage, *machine learning*. The second stage is to fuse the set of fragmented knowledge into structured knowledge, referred to as *non-linear fusion*. The reason of using this terminology is its graph representation as will be shown below. The last stage, called *demand-driven navigation*, is to adapt the learned knowledge to specific applications.

- *R. Lu and S. Zhang are with the Institute of Mathematics, AMSS, Key Lab of MADIS and Key Lab of IIP, Chinese Academy of Sciences, Beijing 100190, China. E-mail: {ruqian, smzhang}@math.ac.cn.*
- *X. Jin is with the CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100864, China, and the School of Computer and Control Engineering, University of CAS, Beijing 100190, China. E-mail: jinxiaolong@ict.ac.cn.*
- *M. Qiu is with the Department of Electrical Engineering, Columbia University, New York, NY 10027. E-mail: qiumeikang@yahoo.com.*
- *X. Wu is with the Research Institute of Big Knowledge, Hefei University of Technology, Hefei, 230009, China, and the School of Computing and Informatics, University of Louisiana at Lafayette, LA 70504. E-mail: xwu@hfut.edu.cn.*

Manuscript received 24 Jan. 2018; revised 12 Apr. 2018; accepted 11 Aug. 2018. Date of publication 23 Aug. 2018; date of current version 2 Aug. 2019. (Corresponding author: Ruqian Lu.)
Recommended for acceptance by G. Chen.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2866863

1041-4347 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

The construction and application of immense knowledge bases have always been a dream of information scientists and computer engineers. An early proposal made by Vannevar Bush in 1945 was the Memex Initiative [6], [7]. Jim Gray declared in his Turing Award Lecture that establishing a personal and a world memex should be a long term goal of computer science [8]. Even DARPA has shown its interest in Bush's idea and started a memex program in early 2014. People generally recognize that the later popularized technology of hypertext, multimedia and even World Wide Web have been predicted by it [7], [9], [10]. Another initiative came from Edward Feigenbaum who proposed to build very large knowledge bases to implement full artificial intelligence. He even proposed a variant of the Turing Test, called Feigenbaum Test, based on such knowledge bases [11]. His student, Douglas Lenat, initiated the Cyc project to construct a huge commonsense knowledge base in 1984 and reached a state of completion in 1994, followed by an open source project, OpenCyc [12]. Its current version is OpenCyc4 [12], [13]. Other examples of very large knowledge bases are recently arising knowledge graphs, to be discussed in Section 4. Similar works have also been done in China. Cao proposed China's National Knowledge Infrastructure (CNKI) in 1995 and the idea of building a knowledge highway [14], [15], [16], [17], collected more than 3 million facts in its knowledge base.

The term 'big knowledge (BK)' appeared in the second decade of this century, after the big hype of big data in the first decade. From big data to big data mining [18], [19] to BigKE [20] to BK [2] to BK mining [21], people's idea flow converges to the same target step by step. For example, Murphy introduced Google's Knowledge Graph (KG) in 2013 [22] with the slogan 'from strings to things'. Russell published his appeal 'Turning Big Data into Big Knowledge' in [23]. Aaron, the marketing director of Pernix Data, called for 'Turn Big Data into Big Knowledge with Infrastructure Analytics' in 2015 [24]. Liebowitz asked "How to extract big knowledge from big data?" [25]. However, they all just mentioned BK without providing a specific definition, and they were probably only discussing large volumes and complex connections of knowledge. In [22], Murphy claimed that KGs are BK without discussing BK in general. Recently, Hershkovitz pointed out [26] that BK is not simply the set of insights obtained by using various analytic tools to analyze big data; Rather, BK is the product distilled from these insights. In [2], [21], BK was described in form of fragmented knowledge from Heterogeneous, Autonomous information sources for Complex and Evolving relationships (HACE), in addition to domain expertise from human experts.

In summary, current BK explorers make advance in the following aspects: 1) distinguishing BK from conventional knowledge; 2) proposing to turn big data into BK; 3) appealing to make industrial profit from BK; and 4) recognizing complexity and hardness of benefiting BK. However, there are also limitations in these BK discussions: 1) No detailed analysis on the common characteristics of BK has been made in a similar way as nV and HACE have been proposed for big data; 2) No special strategies for processing BK have been provided. On the contrary, most of these explorers concern BK only as a technique of knowledge management; 3) A survey on current BK research has not been found; 4) The fact that mining big data does not necessarily generate BK is neglected by many researchers; 5) Many people talk about big data analytics, but very few care for BK analytics, which must be very different from the former; 6) No attention has been paid to the naturel extension of the BK concept, i.e., big-knowledge systems (BK-S), and its engineering issues.

Our interest is to explore BK in its most general sense. Therefore, the major contribution of this paper is the introduction to ten massiveness characteristics (MC) of BK, where C represents also the first letter of the ten common properties, i.e., concepts, connectedness, clean data resources, cases, confidence, capabilities, cumulativeness, concerns, consistency and completeness. Among them, the first five characterize BK in general, while the sixth one is additional for BK-S. The seventh and eighth MCs regarding massive cumulativeness and concerns, respectively, reflect the properties of advanced BK-S like some well-known knowledge graphs, while the last two MCs involving consistency and completeness propose powerful functionalities of future BK-Ss. These characteristics are evidence that BK is not only distinguished by its quantity. The second major part of this paper is to test these BK criteria on five different large-scale knowledge engineering cases, including two Chinese projects and three world ones. They are carefully compared and checked against the BK characteristics.

The remainder of this paper is organized as follows. Section 2 defines six massiveness characteristics for BK and BK-S. Section 3 discusses five large-scale knowledge engineering projects and shows that only two of them produce BK. Section 4 discusses knowledge graphs from BK-S view points. It provides at the same time two new characteristics for advanced KGs. Section 5 introduces BK Engineering (BK-E) with its life cycle paradigm. Section 6 proposes a multi-world model for future BK-S with two theoretically elaborate characteristics. Finally, Section 7 concludes the paper by summarizing its contributions and providing some perspectives of future research.

2 Big Knowledge and Big-Knowledge System

Experts of data analytics have depicted the essence of big data both in a positive way and a negative way. The former is to give characteristics that all big data possess, e.g., 3V, 4V, 5V, and HACE, while the latter says that big data is something that is not something. We found that one can do the similar thing for BK. In this section, we will first present positive characteristics of BK.

Definition 1 (Big Knowledge). *Big Knowledge (BK) is a massive set of structured knowledge elements, where a knowledge element may be a concept, an entity, a datum, a rule or any other computer operable information element. The most popular properties of BK are the following five MC characteristics.*

Characteristic 1: Massive Concepts (MC1). BK is necessarily massive. However, while data are countable, knowledge as an abstract concept is not countable. For example, we cannot say 'I have three knowledge'. We can only use the number of knowledge elements contained in BK to define its quantity. Concept is the most important one among all knowledge element types. Its number should be massive. It is required that each concept is subject to separate processing by the BK search and inference functions. Without concepts there would be no knowledge at all.

It is difficult to provide an absolute lower bound of the quantity of BK. However, we may try to give a relative one.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

The British Encyclopedia contains 228,274 topics or concepts with 474,675 subentries or subconcepts [27]. The English lexical database, WordNet, contains 155,287 words (i.e., instances) organized in 117,659 synsets (i.e., concepts) [28]. The Big Chinese Dictionary contains 250,000 entries or concepts [29]. If we consider them as examples of BK, we may set the borderline at 100,000 concepts.

Characteristic 2: Massive Connectedness (MC2). Connectedness means the degree of how the knowledge elements are connected. It may be a connection in a neural system, a relation in logic, or a fact in terms of a triple, (subject, predicate, object). Without connections there would be no reasoning at all. Given some BK, it is not only the number of its connections that matters. The harmonic distribution of connections among its knowledge elements is more important for its good structuration.

For a quantitative description we distinguish concept-concept relation from instance-instance relation. OpenCyc (version 2, 3, 4) has (47,000, 177,000, 239,000) concepts and (306,000, 1,505,000, 2,093,000) facts (i.e., relations) [12], [13]. The latter is roughly 8 times as many as the former. This prompts us to take 1M as a lower bound of the number of concept-concept connections. For the number of instance-instance relations, see Table 5 in Section 4.

For the harmony of connection distribution, we propose two measures:

Measure 1. The average rate of node pair connection, $m/(n \times (n-1))$, where n is the number of concepts and m is the number of concept pairs bridged by a connection path.

Measure 2. The average rate of local node pair connection, $K(i)/(k(i) \times (k(i)-1)/2)$ averaged over all nodes, where $k(i)$ is the number of adjacent edges of node i, while $K(i)$ is the number of edges between its neighboring nodes.

Later on, we will see that Measure 1 applies better to Wikipedias with reference links among articles as connections (see MC2 of Section 3.5), while Measure 2 fits better knowledge graphs with predicates of entities (i.e., facts) as connections (see the specification of Section 4).

Characteristic 3: Massive Clean Data Resources (MC3). We differentiate between data sources and clean data resources. During the upgrading process from big data to BK, the original data sources are experiencing a process of distillation. In this process the raw data will be cleaned, sieved, selected and possibly transformed to an appropriate form. All these intermediate representations after cleaning should be conserved and maintained as clean data resources and are very useful in the following senses: data recovery in case of data loss or damage, data provenance during data processing, data re-mining if new information is needed, and data reuse in other applications.

Note that BK must come from big data which can be either raw data or any existing knowledge sources (e.g., books, papers, newspapers, videos, and even archaeological findings). The latter is called knowledge legacy in this paper. They often exist in the form of coarse knowledge blocks and should be first e-transformed in a computer readable form and then decomposed into a concept and/or instance form of fine-grained representation.

Note also that cleaned data resources are not necessary in computer readable form. They may be countable or non-countable, explicit or implicit. In archaeology, an ancient grave before scientific examination is just a data source. Only after careful excavation and thorough investigation it becomes clean data resources. We call it resources because usually there is valuable information provided by the grave implicitly, which becomes available only after long time following-up investigations and researches.

The clean data resources are usually large. For example, in case of KG it may be in form of data reservoir where the KG, OpenKN, has 1.2B Web excerpts as clean data resources supporting its knowledge base with 20M facts [30], [31]; or in form of data partition where the learning system, NELL, has over 50M candidate beliefs, among which 3.56M are of high confidence [32] while the remaining 46.44M candidate beliefs are clean data resources; or in form of data sharing where DBpedia has only 4.58M entities in its knowledge base while the volume of its clean data resources are much larger. There are more than 0.24B links pointing to external knowledge sources, including external Web pages and edited knowledge sources like English Wikipedia and YAGO2 [33], [34]. However, sometimes clean data resources may not be massive. For example, from the ancient oracle words on tortoise shells, people have collected 5,000 different ancient Chinese characters [35]. Among them only 2,000 characters have been recognized. The other 3,000 characters (together with files written in them) remain clean data resources. At last, due to the huge difference of clean data resource/knowledge ratios in different cases, we tend to set the average lower bound of MC3 as 10 in terms of the ratio of BK's clean data resource volume to its own volume.

Characteristic 4: Massive Cases (MC4). Massive cases have a twofold meaning. It first indicates that most of the concepts and relations have many instances. It also suggests that most of the knowledge components have had many applications. Regarding the first aspect, Table 5 shows that most KGs have an instance number around 10M and a fact number between 0.1B and 1B. Regarding the second aspect, we note that, for example, Wikipedia has 18B page views and nearly 0.5B visitors each month [33]. For open source knowledge on the Web, one can estimate its degree of application by the number of downloads (e.g., Microsoft claimed that its 'world-wide telescope' has been downloaded at least 10M times [36]).

Characteristic 5: Massive Confidence (MC5). Massive confidence of BK means that most of the BK's knowledge elements have a high index of confidence. Quantitatively it can be expressed with the following statement: There are two positive real numbers $m, n \leq 100$ with $|100 - \min(m, n)| < \delta$ where δ is a small number, such that the confidence scores of no less than $m\%$ of the knowledge elements are no less than $n\%$.

For a quantitative measure, we note that 16 percent of Knowledge Vault's facts reach a confidence degree of 0.9. Although YAGO2 reports that by a random inspection its confidence degree reaches 0.9 with probability 95 percent, the general confidence of KG based on automated knowledge acquisition is low. In general, we recommend $n = m = 80$. For a simpler formulation of quantitative MC5, see Specification 4 of Section 4.

In practice, these MCs are not equally important. We accept imperfect BK with the following specification:

Specification 1. The necessary characteristics of BK are MC1, MC2 and MC5, whilst MC1-5 add up to its sufficient characteristics.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

This (positive) specification helps us differentiate BK from big data. However, we need also a negative one, just as big data was given a negative definition in [18], where it states that big data are "datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze". Different from big data that are easy to get but difficult to analyze and utilize, BK is difficult to construct and maintain but once constructed it is easy to apply. This is due to its immense size of components and elements, rich content of structured information, delicate and complicated structure of organization, high trustworthy of statements, usefulness for an extra-large set of applications, etc., which are beyond the construction and management abilities of traditional knowledge engineering approaches.

Given the above typical features of BK, and considering that the ultimate goal of BK is to serve the society as well as possible, a system of BK, referred to as Big-Knowledge System (BK-S), should thus include advanced algorithms, techniques and tools for solving vast kinds of problems in a prescribed area and provide very user-friendly knowledge service, all of which outperform those in traditional knowledge engineering approaches. That is:

Characteristic 6: Massive Capabilities (MC6). This characteristic requires two aspects of BK abilities, namely, 1) professional abilities for solving a vast set of domain specific problems (e.g., the success of IHGP that will be presented in Section 3 provides a broad prospect of solving many gene related problems of mankind); 2) user friendly abilities for providing high quality knowledge services (e.g., thanks to IHGP, now the third generation sequencing technique is popular and a price of 1000$ per person for genome sequencing becomes possible).

Upon MC6, we can define BK-S as follows:

Definition 2 (Big-Knowledge System). *A BK-S is a system [37] consisting of knowledge elements and function elements, where the former satisfies MC1-5, whilst the latter implements the abilities in MC6.*

In other words, all six characteristics together form a complete definition of BK-S. For a possible quantitative measure of MC6, please see Specification 5 and Table 7 in Section 4.

3 Large-Scale Knowledge Engineering Projects

In this section, we will introduce and discuss five large-scale knowledge engineering projects. Each project collects a huge amount of data and/or knowledge and produces from them a large-scale framework of knowledge which is applied in different contexts including city traffic, archaeology discovery, human health oriented life science exploration and online knowledge encyclopedia. We will analyze the knowledge these projects produce and see that only two of them deserve to be called big knowledge.

3.1 The Fifth Comprehensive Traffic Survey in Shanghai (CTSS)

In 2014, the Shanghai Municipal Government carried out a large project to thoroughly investigate the state of Shanghai's traffic [38]. This is the fifth one in Shanghai, which had a population of more than 24M at that time. This survey manually collected 250,000 questionnaires from about 300,000 persons. It produced not only a general report and a number of sub-reports corresponding to different traffic domains of Shanghai, but also sub-reports on the traffic of its districts and counties. It also produced a plan for improving the traffic in the next five years.

Has this project produced big knowledge? We use the five MC criteria to check the results obtained in this project. First we list the positive results:

MC3: Massive Clean Data Resources Yes. In this project, different kinds of data are collected automatically using various methods and techniques, including remote (i.e., avionic or satellite) sensed data, mobile communication data, car plate data, major toll point data, GPS data of vehicles, truck transportation data, traffic card swiping data, etc. The data collected reveal that there are 18 million active mobiles with 6-7 communications per mobile each day. That means more than 0.5 billion mobile communication records are collected per day. The survey involved 75,000 families, 65,000 various vehicles (including tramcars, buses, taxies, ferries, metro, bicycles, motorcycles, etc.), 11 communication hubs, 80 railway stops, 90 bus routes and 130 intersections. It particularly investigated peak and off-peak traffic hours, traffic jams in urban areas, passenger capacity of rail transit, quantity demand of private cars, etc.

MC5: Massive Confidence Weakly yes. Traffic data collected in this project are mainly from two types of channels. The first type includes governmental departments, public institutions, commercial corporations, etc. Those data are usually of high confidence. Traffic data of the other type are autonomously generated by the residents of Shanghai. Although they are in principle reliable, their trustworthiness and reliability may be not completely perfect. Moreover, what one wants is the confidence of knowledge which should be abstracted from these data through a reasoning process. The final confidence of obtained knowledge depends also on the confidence of the reasoning process, which is not known to the authors.

The others are negative results, including:

MC1: Massive Concepts No. There is only a small set of traffic concepts related to Shanghai.

MC2: Massive Connectedness No. This is a conclusion of the above statement since few concepts cannot have massive connections. Moreover, there is no massive connectedness regarding entities either, since although CTSS does have massive data, it does not have massive entities because any kind of entities, including vehicles, habitants etc., is considered as just a huge number of indistinguishable copies of the same model.

MC4: Massive Cases No. Each of the few concepts has only limited featured instances. For example, Shanghai does have a population of more than 20M. But all habitants in this project are considered as equal copies of the same model and thus their individuality is neglected.

Since the necessary characteristics MC1 and MC2 are lacking, the knowledge collected by this project is not BK.

3.2 The China's Xia-Shang-Zhou Chronology (XSZC) Project

The XSZC project [39], [40], proposed and organized by the State Commission of Science and Technology of China, aimed to decide the chronology of a period covering more than 1200 years of China's ancient history. China has been known as a country with five thousand year's civilization. But according to the 'Chronology of 12 Dukes' chapter in the classic book "The Historical Records" [41], authored by

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
China's Ancient History without Chronology

Epoch	Exact Years	Information Source	Reliability Level
Three emperors and five sovereigns	Ca. 3000 BC-? BC	Legend books	Legend level
Xia Dynasty	? BC-? BC	+Historical stories and archaeological ruins	Controversy level
Shang Dynasty	? BC-? BC	+Oracle records w/o chronology	Dynasties level w/o chronology
Early Zhou Dynasty	? BC-841 BC	+Bronzeware inscription with few records	Few chronological data

Qian Sima, a well-known Chinese historian of the Han dynasty and the father of Chinese historiography, China's royally recorded chronological history only started from the year 841 BC. China's ancient history as a country before 841 BC can be found in other books, documents, biographies, even inscriptions on stones or bronze wares, cattle bones or tortoise shells. However, the chronological data were often missing or imprecise. Table 1 shows what people knew about China's ancient history before this project.

The XSZC project started in 1996 and its major part was finished in 2000. It includes 9 topics and 44 subtopics. 70 scientists of history, archaeology, philology, paleography, geography, astronomy and chronology joined this research. The major part of its results, i.e., the Xia-Shang-Zhou Chronology Table published in 2000 [42], was the most scientific one among all existing Chinese ancient chronology tables until that time. It extends the recorded history of Chinese ancient civilization for more than 1200 years, including the starting and ending years of all dynasties and all kings' regimes during that period. Note that archaeological findings provide the evidence about the Chinese neolithic culture until the 62nd century BC but without chronology [43]. Table 2 presents the main conclusions of the project.

Next, let's investigate whether the knowledge produced by this project is BK. We also start from the positive parts.

MC3: Massive Clean Data Resources Yes. The historical information contained in archaeological findings is often implicit and not countable. Exactly deciding ancient historical events depends on trustworthy archaeological findings and their appropriate interpretation. It also includes historical documents, books, oracle inscriptions and bronzeware inscriptions (see Fig. 1). More than 10,000 bronzeware pieces and 50,000-60,000 cattle bone/tortoise shell pieces carrying inscription in ancient Chinese characters have been found. All these served as clean data resources of the XSZC project. Besides, it should also include ancient historical texts. Just for an example, the number of Chinese ancient astronomical books counts about 400.

MC5: Massive Confidence Controversially yes. That means there is massive evidence confirming the validness of the project's results. However, there are different interpretations regarding these results. The massive evidence includes: discovery of relics of the three dynasties, for example Erlitou, an early capital of Xia with its palace ruins near the city Luoyang, was found, which confirms the existence of the Xia dynasty and the transition time between Xia and Shang dynasties; written archaeological records (e.g., the 50,000-60,000 pieces of oracle mentioned above) and the many inscriptions on bronzewares containing records on big ancient events; astronomical records (e.g., experts decided 899 BC as the first year of Zhou Yi King because of a total solar eclipse in that year); chemical techniques [40] (e.g., the carbon 14 (c14) technique may refine the time slots of historical events with precision of 95 percent , i.e., the average error is about $\pm$300 years); physical methods [40] (e.g., by using improved accelerator mass spectrometer (AMS) c14 technique, an error rate of 0.4-0.5 percent may be reached, i.e., $\pm$32-40 years of error). Thus the last two methods both fit the MC5 standard.

However, there are controversial opinions from both domestic and international experts. Some experts asserted that the so-called Xia dynasty was only invented by people of the Zhou dynasty because there were no written records about the Xia dynasty in spite of many archaeological findings. But since all opposite points of view are not based on contemporary techniques, we omit the discussion about it.

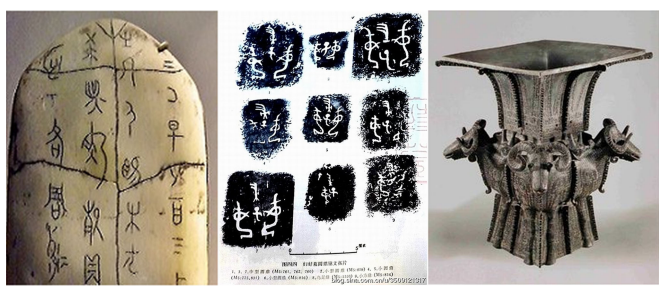
The negative results are:

MC1: Massive Concepts No, but potentially yes. This project involves concepts such as dynasties, kings, enthroning and abdicating etc. Their number is not big. But these concepts will help derive a huge set of events over a period of 1229 years, involving a massive set of other concepts, which will be revealed in future research. Therefore, different from the CTSS project, the XSZC project has a potential of fitting MC1.

MC2: Massive Connectedness No, but potentially yes. This is a conclusion of the above statement since few concepts cannot have massive connections. For the same reason as above we attach a 'potentially yes' to it.

TABLE 2
China's Ancient History after the XSZC Project

Epoch	Exact Years	Roughly Contemporary Events in the World
Three emperors and five sovereigns	3000 BC-2070 BC	Neolithic era - First pyramids in Egypt (2700 BC)
Xia Dynasty (17 kings)	2070 BC-1600 BC	Foundation of Babylon state (1800 BC)
Shang Dynasty (31 kings)	1600 BC-1046 BC	Israel out of Egypt (1445 BC), Trojan War (1193 BC)
Early Zhou Dynasty (10 kings)	1046 BC-841 BC	First Olympic Games (776 BC)



(a) (b) (c)

Fig. 1. (a) Oracle on a tortoise shell [44]. (b) Inscription 'Fu Hao' rubbed from a bronze tripod [45]. (c) Square zun (vessel) with four sheep [46].

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

MC4: Massive Cases No, although potentially yes. This project generated (event, time) connections for all regimes of the 58 kings of the three dynasties. This number is not big. But these data help determine the chronology of many important events over 1229 years. The number of latter's connections will be massive if subsequent research is done.

3.3 The Troy and Trojan War Project (TTWP)

The archaeological research on Troy [47] and Trojan War [48] has lasted for several hundred years. It has many similar characteristics with the XSZC project. First, both of them are well-known world historical events. Second, both of them have attracted heavy studies from lots of world experts. Third, both of them have left significant controversies unsolved among the experts. The researchers had interests in the following questions: Did the city Troy really exist? If yes, when and where did it exist? Did the Trajan War really happen? If yes, when and where did it happen?

The research on these questions was mainly initiated by Homer's epic cycle poems, Iliad and Odyssey. The former describes (the last ten days of) the Trojan War where the Achaeans (Greeks) besieged and destroyed the city Troy. The latter narrates the journey home of Odysseus, one of the war's heroes. The first man who started excavation in Trojan area in 1865 was the Britain Frank Calvert, who found the right place Hisarlik, a small Turkish village. Three years later the rich German businessman, Heinrich Schliemann, started excavation at Hisarlik, which harvested rich ancient remains of Troy. Since then most of the experts believe that the Homeric Troy city was buried under Hisarlik. Following Schliemann, the excavation of Troy was continued again and again until recent (2014) [49].

The results of continuous excavations show that the ruins in Hisarlik incorporate many ancient cities one above another. There are in total nine layers of ruins covering a time cycle starting from 3000 BC until 500 AD [47]. Some layers are even composed of several sublayers. Experts acknowledge now that the seventh layer is most possibly the ruin of the ancient Troy city.

Did TTWP generate BK? Let's check it against the five MCs. Here it is difficult to draw a clear line between positive and negative results since the answer is often partially yes/no.

MC1: Massive Concepts No, but yes in generalized sense. The research of TTWP includes four parts [50]: 1) the city Troy before the Trojan War, where it has been and how was it constructed, developed and destroyed? 2) the Trojan War, did it really happen? If yes, how and where was it? 3) the city Troy after the Trojan War: how did Troy became today's Hisarlik? 4) the influence of Troy and the Trojan War on the neighboring areas [51]. The historians agree that Troy's influence is not only limited in Trojan area, neither only in Asia Minor. It extends to many countries of Asia and Europe. The concepts involved in the first three parts are not massive. But those in the fourth part are massive, where so many concepts about nationalities, geography, states and poleis, conflicts, splitting and unification of political powers, etc. are involved.

MC2: Massive Connectedness No, but yes in generalized sense. This is a conclusion of the statements above.

MC3: Massive Clean Data Resources Yes. These resources come from three data sets. The first one is a great set of legend books, where the Greek epic cycle is the major source. It includes a collection of epic poems, composed in eight groups including 77 books [52]. The second one is the whole treasure dug out from the ruins of Hisarlik and its neighboring area. The third set is the history literature of relevant countries such as Greece, Turkey, Italy and Iran, in addition the ancient Persian and Rome, not yet mentioning the numerous contemporary research literatures.

MC4: Massive Cases No. Although the number of concepts is massive, that of instances of each concept is limited since only instances relating to Trojan events are counted.

MC5: Massive Confidence Partially yes. The existence of the ancient Troy city has been confirmed by archaeological work lasting more than hundreds of years. However, it has never been confirmed that the so-called Trojan War has happened at this city.

3.4 The International Human Genome Project (IHGP)

IHGP is a well-known large-scale knowledge engineering project starting from the last century [53], [54]. Its main part started in 1990 and ended in 2003. Its task is the sequencing of three billions of base pairs of human chromosomes, DNA sequencing for short. Because of the existence of massive individual differences in human DNA provided by volunteers, the sequencing result of this project is just a framework of human's genome. Following works extended the main task to gene determination and its genetic and epigenetic effects.

Next, we examine IHGP according to the above five MC standard.

MC1: Massive Concepts Yes. We just consider three data sets. The first one is the set of genes. IHGP has found that the total number of human genes is about 25,000 [54]. A more exact number can be found in the 'English-Chinese Human Gene Dictionary' [55], which collects 40,172 genes, including 24,858 protein-coding genes and 15,314 non-coding genes (such as rRNA, tRNA, scRNA, snRNA, snoRNA and others). If we consider each gene as a concept, then we have already 40,172 concepts. The second set is the Human Phenotype Ontology (HPO) containing 11,000 items [56]. The third set is the Human Gene Mutation Database (HGMD), which collects 214,158 mutation records [57]. Thus, the three sets in total make 265,330 concepts. This number already fits the MC1 characteristic (at least 100,000 concepts) well.

MC2: Massive Connectedness Yes. Researchers have found comprehensive connections in IHGP. Among them the Single Nucleotide Polymorphism (SNP) database including 3.7M SNP sites is a well-known example [54]. This database keeps a huge set of relations to many biological phenomena. These connections are distributed in a genome-wide way. It was reported in 2015 that the Genome-wide Repository of Associations between human genome SNPs and Phenotypes (GRASP) database v2.0 already contains 658.87M SNP associations collected in 2,082 studies [58]. This number satisfies the MC2 standard well. In fact there are many other meaningful connections to SNP. For example, in March 2014, it was claimed that 3,961 of the 3.7M SNP sites are related to 571 known diseases [59]. Three years later, in 2017, this number raised to 24,218 [60], [61]. Considering the tremendous size of possible gene combinations, the number of diseases caused by multiple genes must be also tremendous.

MC3: Massive Clean Data Resources Yes. The principal clean data resources of IHGP are $3n$ billion base pairs of human chromosome, where n is the number of voluntary chromosome providers. The most convincing fact is again

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
MC2 Harmony of Wikipedia's Connections [68]

	# of Nodes	# of Edges	Harmony (Mesure1)
Wikipedia (En)	11,699,099	519,092,989	0.9995
Wikipedia (Ch)	1,699,387	65,656,260	0.9998
Wikipedia (Ge)	3,469,352	91,131,146	0.9988

the 3.7M SNP sites which are the major source of future gene mutation findings. This already makes IHGP reaching the MC3 standard (10 times as much as the number of concepts). In order to implement the following work of IHGP, the number of voluntary providers must be greatly increased since in this case DNA sequencing must be done for additional large human groups, say, according to different diseases, different nationalities or even habitants of different villages and tribes, different chromosome segments, etc. For example, UK has declared a 'hundred thousand human genome sequencing' project to be completed within 3-5 years [62]. China has completed Major Histocompatibility Complex (MHC) area sequencing for a group of 20,635 humans and constructed a most up-to-date MHC gene database for the Han nationality [63]. Based on this database, research has been done for, say, determining pathogenic genes of psoriasis. These examples remind us that human chromosome sequencing data will be immense.

MC4: Massive Cases Yes. The results of sequencing will have a big impact on world's 7.5B populations. DNA sequencing techniques have been applied to finding pathogenic genes, determining blood relationships, latching down criminals, and many other targets. We even believe that with the further descending prices, for example, if the sequencing cost decreases to less than 100$ per person, people will be happy to pay the fee of sequencing themselves.

MC5: Massive Confidence Yes. In the project's first stage, i.e., until 2001, about 90 percent of the base pairs have been sequenced. The error rate was 0.1 percent with 150,000 gaps remaining. Only 28 percent of the genes were found. Later in 2003, when the complete draft was finished, the sequenced base pairs reached 99 percent of the whole. The error rate was lowered down to 0.01 percent with 400 gaps only. Certainly these figures can be considered as a measure of confidence of the knowledge acquired by this project.

3.5 The Free Online Encyclopedia – Wikipedia (WIKI)

Wikipedia is a well-known online encyclopedia, where all articles in 299 different languages are created and edited by volunteers. It is the largest and most popular general reference work on the Internet. Wikipedia is ranked the fifth-most popular website [33]. More details about Wikipedia can be found at its website [33]. In what follows, we quickly examine whether or not Wikipedia produces BK in terms of the five MC criteria.

First we list the positive aspects.

MC1: Massive Concepts Yes. As of September 30, 2017, Wikipedia possesses more than 46.5M articles in 299 different languages [64], among which English Wikipedia has published 5,512,475 articles, with year-on-year growth of more than 11.8 percent. Each article in Wikipedia presents a topic which can generally be regarded as a concept or an entity. In addition, each article may contain many other related concepts or entities. According to our statistic over 500 randomly selected articles, the rate of conceptual ones is about 24.8 percent. Thus, we estimated the number of English concept articles is about 1.4M. That is to say merely the number of conceptual articles in English Wikipedia fulfills the specification of MC1 (0.1M). Besides, in the past few years, each month there were more than 10,000 new articles. For instance, 12,937 new articles were added to Wikipedia in September 2017. This suggests that the number of concepts/entities in Wikipedia still increases rapidly.

MC2: Massive Connectedness Yes. It is reported that as of September 2009, Wikipedia had more than 309M internal links, 130M interwiki links and 25.5M external links [64]. In September 2017, Wikipedia contains up to 39.0M another kind of connections called redirects [64]. In addition, as aforesaid, DBpedia borrows 580M various links from English Wikipedia [65], [66], [67]. As for the harmony of connections, see Table 3 for recent statistics of English, Chinese and German Wikipedias, where the connections mean internal links among articles.

MC4: Massive Cases Yes. It is clearly reported that as of February 2014, Wikipedia had 18B page views and nearly 500M unique visitors each month [33], which suggests that most of the knowledge components (i.e., concepts and entities as articles of Wikipedia) have had many applications. At least, they were extensively referenced.

MC5: Massive Confidence Yes. As of March 2017, Wikipedia has about 40,000 high-quality articles (featured articles), which cover vital topics [64]. In 2005, the top journal, Nature, reported that the level of accuracy of Wikipedia approached that of Encyclopedia Britannica [69].

There is also a negative aspect.

MC3: Massive Clean Data Resources No. Wikipedia works in a crowdsourcing mode. There is no known database collecting material serving as clean data resources for the articles. The references in each article cannot be viewed as clean data resources, because all referenced articles remain in their original form. They have not been processed or semi-processed in any way to serve as clean data resources. On the other hand, although Wikipedia reserves all traces of revision of each article, these revision traces cannot be considered as clean data resources, either, because they are only useful for provenance. They do not provide any new information for further mining.

Since MC3 is not a necessary condition for BK, by summarizing the above analysis we confirm that the whole set of Wikipedia articles is BK.

To conclude this section, we present a comparison of the knowledge provided by the above five projects in Table 4.

4 KG as Representation of BK-S

As a graph-based knowledge representation, Knowledge Graph is actually not new. Its semantic origin gets back to the idea of semantic network developed since the middle of last century. Many models have been established. To mention a few, we recall the semantic network of Quinlan in 1963 [70], the case grammar of Fillmore in 1968 [71], the CODASYL data model in 1971 [72], the ER model of Chen in 1976 [73], and the conceptual graph of Sowa in 1976 [74]. In 2007, Mentawei opened a new era of semantic network based knowledge representation by publishing its KG, Freebase [75]. But KG became well-known to public only later when Google bought Freebase from Mentawei in 2010 and

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

TABLE 4
Comparison of the Five Projects

	MC1	MC2	MC3	MC4	MC5
CTSS	No	No	Yes	No	Weakly yes
XSZC	No, potentially yes	No, potentially yes	Yes	No, potentially yes	Controversially yes
TTWP	No, but yes in a generalized sense	No, but yes in a generalized sense	Yes	No	Partially yes
IHGP	Yes	Yes	Yes	Yes	Yes
WIKI	Yes	Yes	No	Yes	Yes

published its own KG in 2012 [22]. Since then KG research went into a stage of rapid development.

KGs can be classified into three categories according to their difference in research targets. The first category is search oriented. These KGs are developed for facilitating a search engine with a huge knowledge network. Their nodes are usually entities while edges are facts. They are generally not open to the public. Examples are Google KG [22], Knowledge Vault [76], Facebook Graph [77], OpenKN [30], [31], etc. The second category is open to the public and service oriented. Each of such KGs usually borrows knowledge from other knowledge bases as start. Many new KGs even maintain a great number of links to other knowledge sources. In this way the new KGs grow very fast. Examples are OpenCyc1-4 [12], [13], Probase [78], [79], DBPedia [67], YAGO1-3 [34], [80], [81], CNKI [3], [8], [82], [83], [84]. The third category is of research type. These KGs usually generate a large part of knowledge themselves. Many of them provide results for language learning and natural language processing, such as, WordNet [28], HowNet [85], [86], FrameNet [87], ConceptNet [88], NELL [89] and MuiseNet [90]. All KGs made heavy use of large-scale knowledge acquisition techniques like OpenIE (Open Information Extraction) [91], [92], [93]. Relevant systems include TextRunner [94], ReVerb [93], KnowItAll [91], [92], OLLIE [95] and XHSNet [15], [83], [96].

There have been many works surveying and comparing currently existing KGs. For example, Farber et al. made a comparative survey on DBpedia, Freebase, OpenCyc, Wikidata, and YAGO in [97] and further established 34 data quality criteria for KGs, calculated and analyzed corresponding features of these five KGs in detail [98]. Paulheim analyzed DBpedia, YAGO, Freebase, Wikidata, NELL, OpenCyc, Google Knowledge Graph and Knowledge Vault with their knowledge refinement features [99]. Miran's report presents a survey on OpenIE systems including Textrunner, Reverb and OLLIE [100]. In this section we will give a review on contemporary KGs according to our KB standards.

As for the eligibility of KGs as BK-S, we first investigate the three most essential characteristics of BK, namely, MC1, MC2 and MC5. For this purpose we list their basic data in Table 5. Note that different systems are using different knowledge element representations, where some are using concept/class type (C-type), while others are using entity/object type (E-type). In addition, the statistics of different versions and publication years are mixed together. All these make a precise comparison difficult. To diminish ambiguity, the publication year and reference are attached to each entry of Table 5. We set the lower bound of BK at the head of each column, e.g., 100K is the lower bound for the number of concepts in the second column. This number was calculated with such principle: 1) It is just enough for making no less than half of the systems of the same type in Table 5 pass the limit; 2) At least one of these KGs would fail to pass the limit, if the bound increases for one order higher. By a careful investigation on Table 5, we propose the following quantitative bounds:

Specification 2 (Quantitative). MC1 characteristic for KG: The number of concepts is no less than 100K or the number of entities is no less than 10M.

Specification 3 (Quantitative). MC2 characteristic for KG: The number of facts is no less than 1M and 1B for C-type and E-type KG, respectively. The lower bound of the harmony of connections is set to 80 percent, based on statistics of several major KGs, see Table 6.

Specification 4 (Quantitative). MC5 characteristic for KG: It depends on the way of knowledge collection used by different KGs. Those KG relying on manual knowledge collection and checking can usually reach a standard of nearly 100 percent reliability, such as OpenCyc, while those relying on OpenIE techniques like Probase and Knowledge Vault often report a precision more than 90 percent, which

TABLE 5
The Basic Statistics of Some Well-Known KGs

KG	Concept (100K)	Entity (10M)	Fact (1M / 1B)	Precision (90%)
Freebase (E)	2000 (2016) [101]	44M (2013) [102]	3.1B (2017) [98]	99% (2017) [98][1]
Google Knowledge Graph (E)	N/A	570M (2012) [103]	70B (2016) [104]	88% (2012) [105]
Knowledge Vault (E)	N/A	45M (2014) [76]	1.6B (2014) [76]	90% (2014, ca. 270M estimated) [76]
Probase (C)	5.4M (2016) [78]	N/A	20,757,545 (2016) [78]	92.8% (2012) [79]
DBpedia	N/A	4.58M (2014) [65]	3B (2014) [65]	99% (2017) [98]
OpenKN (E)	10,000 (2014) [30]	20M (2014) [30]	2.2B (2014) [30]	N/A
Xlore	856,146 (2013) [106]	7,854,301 (2013) [106]	N/A	90.48% (F1-score) (2013) [106]
YAGO2 (C)	350,000 (2015) [34]	10M (2015) [34]	120M (2015) [34]	95% (2011) [34]
YAGO3	N/A	45M (2015) [81]	540M (2015) [81]	99% (2017) [98]
Wikidata	N/A	15M (2014) [107]	43M (2014) [107]	99% (2017) [98]
OpenCyc	6,000 (2008) [12]	N/A	60,000 (2008) [12]	nearly 100%
OpenCyc2	47,000 (2009) [12]	N/A	306,000 (2009) [12]	nearly 100%
OpenCyc3 (C)	177,000 (2010) [12]	N/A	1,505,000 (2010) [12]	nearly 100%
OpenCyc4 (C)	239,000 (2012) [12]	N/A	2,093,000 (2012) [12]	nearly 100%

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

TABLE 6
MC2 Harmony of KG's Connections [108]

	# of Nodes	# of Edges	Harmony (Mesure2)
NELL	698,877	1,142,732	0.014
YAGO3	3,392,749	9,861,018	0.943
Probase	16,924,244	33,343,000	0.807
DBpedia	3,966,924	13,820,853	0.978

was usually calculated from a sample set of data. In this paper, we set a rough lower bound of MC5 at 90 percent for both approaches of precision determination and both C-type and E-type KG.

Note that the quantitative bounds of MCs given in Section 2 are general principles for all kinds of BK. These principles should be concretized in any particular situation of KB. In Section 2 we only considered concepts as knowledge elements, while for KG we consider both concepts and entities. The standards for KG must be adapted to these two cases separately. This is why the criteria here look a little bit different from Section 2.

An investigation on Table 5 reveals that from the 14 listed KGs, four KGs of C-type and three KGs of E-type meet or roughly meet the above characteristics. They are marked with (C) or (E) in the first column, respectively.

Besides concrete definitions of MC1, MC2 and MC5 for KG as BK, we still need a definition of MC6 for KG as BK-S. A general requirement of MC6 has been provided in Section 2. However, its quantification is difficult, if not impossible. Here, instead of counting the techniques or algorithms the BK-S has, we decide MC6 by measuring the quality of KG's knowledge service as follows.

Specification 5 (Specialized). MC6 characteristic for KG: A KG is said to meet the massive capability characteristic, if its knowledge service quality is in the first $m\%$ percent of no less than $n\%$ of all service quality aspects, where m (sufficiently small) and n (sufficiently large) are positive numbers no larger than 100.

Note that the above positive numbers m and n should be determined case by case. To have a measurable standard, we borrow the idea of [98] where a list of data quality dimensions have been defined and five KGs have been taken for comparison. We consider their 'data quality' as our knowledge service quality. Among the many dimensions of [98], we select seven criteria: accuracy (objective correctness), trustworthiness (subjective correctness), consistency (no conflict), completeness (of population), timeliness (of updates), accessibility (easily retrievable) and coverage (of domains). The data of Table 7 are selectively fetched and calculated from Tables 1 and 15 in [98]. The values for each criterion of each KG in the table are averages over all sub-criteria. In this case, we set $m\% = 0.334$, $n\% = 0.666$. The values of g-score show how many criteria are met for this KG, while the d-score shows how many KG meet this criterion. The last datum of the right most column shows that there are two KGs out of five, namely, Freebase and Wikidata, which meet the requirement of MC6.

Based on the above discussion on KGs, we propose two extra MCs for advanced BK-S.

Characteristic 7: Massive Cumulativeness (MC7). A BK-S should undergo a continuous and uninterrupted growth and renewing process of its knowledge elements and its service capabilities. People are not only interested in knowledge contained in it, but also in its modification, cumulative evolution and future tendency. People also enjoy its steadily evolving new functions. All techniques and tools implementing these functions should be improved, renewed and reinvented as well with the proceeding of time. This is to say that advanced BK-S should be living and dynamic, rather than dead and static.

Most existing KGs are cumulative. For example, from YAGO2 in 2012 to YAGO3 in 2014, the number of objects (facts) increased from 10M (120M) to 45M (540M). This is a 4.5 times increase within two years. Another cumulative KG is OpenKN [30], [31], consisting of heterogeneous nodes and edges and specialized in grabbing, analyzing, integrating and recommending news from the Web. BK-Ss converging to some finite state are called stable. In the process of its evolution, new information and data arrive steadily. Data renewal promotes knowledge renewal, including the renewal of concepts, connections and other knowledge elements.

Note that among the three categories of KG mentioned above, the first two categories share another characteristic, i.e., the broadness of knowledge types and limitlessness of knowledge volume. This is the following:

Characteristic 8: Massive Concerns (MC8). This kind of BK-Ss, limited to KG in this subsection, is not limited to any special set of knowledge domains. They collect any kind of knowledge. This kind of KGs is particularly suitable for search engine use and public knowledge popularization purpose. Most of the currently popular KGs are of this type, e.g., Google KG [22], Knowledge Vault [76], Probase [78], [79], YAGO1-3 [34], [80], [81], Wikidata [107] and Xlore [106]. Particularly, Xlore is a multilingual domain independent system, integrating English Wiki, Chinese Wiki and other two Wiki-type knowledge services in Chinese, i.e., Baidu-Encyc and Hudong-Encyc.

TABLE 7
Comparison of Knowledge Service Quality of Five KGs Based on the Data in [98]

Dimension	DBpedia	Freebase	OpenCyc	Wikidata	YAGO	d-score
Accuracy	0.996	0.998	1.000	0.998	0.872	5/5
Trustworthiness	0.333	0.833	0.333	0.917	0.417	2/5
Consistency	0.622	0.817	0.666	0.833	0.442	2/5
Completeness	0.746	0.709	0.467	0.758	0.725	4/5
Timeliness	0.167	0.666	0.083	0.666	0.417	0/5
Accessibility	0.928	0.491	0.571	0.773	0.819	3/5
Coverage	0.880	0.920	0.810	0.410	0.820	4/5
g-score	4/7	5/7	2/7	5/7	4/7	2/5

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

5 Big-Knowledge Engineering

Feigenbaum has once defined knowledge engineering as 'the art of designing and constructing expert systems and other knowledge based systems' [109]. But as the 8 MCs have described above, BK and BK-S are so massive and complicated such that the traditional knowledge engineering techniques fail to construct advanced BK-S. More precisely we present the following characterization:

Definition 3 (Big-Knowledge Engineering). *Big-Knowledge Engineering (BK-E) refers to the engineering of BK-S, i.e., the art of using scientific methods to acquire BK, and to design, construct and apply BK-S towards a strongly significant target, following the life cycle defined below.*

Corollary 1. *BK-E distinguishes itself in the following senses, where those marked with star '*' are usually missing in non-BigKE projects [2]:*

- *Big target*: It is of crucial significance to determine the scientific or application target of BK-E at the beginning, in particular the skeletal design of the BK-S to be developed and its long range view of evolution. All projects in Section 3 have big targets.*
- Big data*: Possess a large amount of big data in all possible forms. Utilize advanced data analytics techniques to process, mine and manage them.*
- *Big knowledge*: A super capability of massive knowledge mining, composition, fusing and management.*
- *Big techniques*: Make use of all available (not limited to IT based) techniques and tools to help mine, validate and structure knowledge. See Section 3.2 for a typical example.*
- *Big services*: Develop a powerful set of techniques and tools to satisfy the multiple and variable needs of all potential users of the BK-S generated by it. For example see the discussion on digital earth in Section 6. Potential quantified criteria are the service quality indices showcased in Table 7.*
- Big life cycle*: In general, the process of knowledge acquisition and renewal is infinite. Also the improvement and renewal of management and service capabilities is infinite. Therefore, BK-E is always in action, whenever BK-S is in service. Its quantization depends on that of its models. One of the models is shown below.*

The above definition and corollary about BK-E are only guiding principles. Based on the above analysis various models of BK-E are possible. In what follows, we propose a life cycle model for BK-E as one of its possible paradigms. Since we did not find any current or past knowledge engineering project meeting all the above criteria perfectly, we have to collect examples from different projects across the following presentation.

The life cycle of BK-E is:

1) *Analysis Stage*: Determine the scientific goal and service target of BK and BK-S, e.g., the delicate planning and organization of the five projects discussed in Section 2.
2) *Design Stage*: Determine a framework of BK and the structure of BK-S. The first thing is to have a BK management system like the KGMS in [110], which could be a prototype of this kind. For knowledge representation, see the examples of ontology structure [13] or Semantic Web representation [22].
3) *Tool Development Stage*: Develop techniques and tools for knowledge acquisition, system construction and user service, e.g., implementing machine learning algorithms like the OpenIE method [92] and/or knowledge embedding techniques [111].
4) *Knowledge Collection Stage*: Use all possible techniques to acquire and test data and knowledge fragments from all possible sources, build clean data resources, e.g., extracting information from other sources and/or searching the Web for acquiring data [76].
5) *Integration Stage*: Construct the BK by decomposing, mining, transforming and fusing all knowledge into the framework. Combine BK with tools developed in the third stage to build a BK-S, e.g., probabilistic knowledge fusion [76] or multilingual source integration [67].
6) *Evaluation Stage*: Test the BK according to MC1-MC5, or also other MCs if required. Assign reliability values to knowledge elements, e.g., confidence calculation [32] or knowledge completion [112].
7) *Validation Stage*: Test the BK-S according to MC6, or also other MCs if required. Check availability and correctness of knowledge management and service functions, e.g., Google KG's search assistant [22] and collaborative Web service recommendation [113].
8) *Application Stage*: Apply the BK-S for practical cases. Keep the system running and fork a thread in parallel with feedback information to some early stages if necessary, e.g., OpenKN for daily news filtration and recommendation [30] and NELL for never ending language learning [89].

TABLE 8
Comparison of the Five Projects in Terms of the BK-E Framework

	Big Target	Big Data	Big Knowledge	Big Tech.	Big Serv.	Big Life Cyc.
CTSS	No	Yes	No	Yes	Yes	No
XSZC	Yes	Yes	No	Yes	No	No
TTWP	Yes	Yes	No	Yes	No	No
IHGP	Yes	Yes	Yes	Yes	Yes	Yes
WIKI	Yes	Yes	Yes	No	Yes	Yes

Based on the above discussion on BK-E, let's recheck the five projects, CTSS, XSZC, TTWP, IHGP and WIKI, in Section 2. The results are presented in Table 8, where these projects are checked with the six criteria. Simply speaking, the only criterion satisfied by all projects is big data. The only project satisfying all criteria is IHGP. The criteria satisfied by least projects are BK and big life cycle, which are only satisfied by half of the projects. The project satisfying least criteria is XSZC. The reader may have noticed that this table does not have an entry checking explicitly whether these projects generate BK-S. This is because we measure MC6 with quality of knowledge service (see Table 8). Hence, 'big knowledge' + 'big service' is equal to 'BK-S' implicitly. This implies that both IHGP and WIKI have generated BK-S. Note also that Table 8 does not include KGs, because there is big difference between different KGs.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

6 A Multi-World Model of BK

In this section, we propose promising properties of BK, which do not exist in any current BK-Ss but should be possessed by future BK-Ss. The reasoning power of a BK-S depends on its two properties, namely, consistency and completeness. In the previous sections, we have only required knowledge confidence, but not yet consistency, nor completeness. Usually it is very expensive, if not impossible, to check these two properties of a big system. BKs such as some KGs have therefore blurred the contradiction of facts with fuzzy or probabilistic measures. It follows among others from our negative definition that BKs size is beyond the ability of traditional logic processing. In fact, Lenat was aware of this problem when constructing the Cyc knowledge base [12], [13] and has proposed efficient ways to solve it. All propositions of Cyc have been classified in different micro theories, each of which is guaranteed to be consistent. However, the trans-micro-theory consistency is not assured. We call this methodology as local consistency. We propose that the reasoning of BK is also based on local consistency, but with different meaning.

For this purpose, we introduce the idea of possible worlds, which is well-known in modal logic semantics. For a BK-S, any knowledge element may belong to different possible worlds and have different values in them. The set of all knowledge elements are covered by a set of possible worlds. Each possible world has a confidence degree, meaning the degree of uniformly confidence of all knowledge elements in this possible world. There are many ways of calculating the confidence degree of a possible world. Thus, we can introduce a new standard for consistency.

Characteristic 9: Massive Consistency (MC9). By massive consistency, we mean that the whole BK can be covered by a set of possible worlds, where each possible world is logically consistent and at least one possible world satisfies MC1, MC2 and MC5. Note that there may be no global consistency in the BK as a whole. A knowledge element may have value true with high confidence in some possible worlds, while having value false in some other possible worlds equally with high confidence. For instance, the conclusion that the Xia dynasty has started in 2070 BC has a high confidence in the possible world of the chronology table (i.e., Table 2), while it is totally false in the possible world of opponents who deny the existence of the Xia dynasty. It also has a low confidence in the third possible world, i.e., opinions claiming that the Xia dynasty has existed but more than 10 centuries earlier than stated in the table. As a consequence, when we say consistency maintenance in the following, we mean local maintenance.

Together with consistency, we should also introduce completeness with three-fold meanings. The first one is purely logical. It says that for each possible world there should be an axiom system such that each true proposition of this possible world can be derived from it; The second meaning is practical. It says that each concept has at least one entity and each entity should have values for each attribute of the class it belongs to; The third meaning allows the existence of a complete set of possible worlds regarding different cognitions toward the same topic as we have seen in the chronology project.

Among the three meanings mentioned above, the first one has been studied by Lenat [13]. It is logically perfect, but may be cumbersome to implement in BK. The second one has been accepted by most current KGs. But it fits entity description better and is less appropriate for concepts that usually need literal explanations. The third one is a unique property of our BK architecture. It does not throw away any knowledge that may not have been confirmed totally at present, but may be proved true in the future. Some important past knowledge may be worth retained in BK, even if it has already been proven false. For example, people may wish to learn that there was Ptolemy's geocentric theory before Galileo's heliocentric theory.

Another aspect of knowledge completeness is the different view angles observing the same object. For example, the Wikipedia text on digital earth tells us never to be satisfied with only one view angle. Its authors suggested studying digital earth from various multiple views [114]: Many digital earths meeting different people's needs; Many digital earths meeting different application's needs; Many digital earths reflecting different cosmological times; Many digital earths supported by different software packets.

This is what we mean by completeness regarding knowledge's multiple view angles. Thus, we have

Characteristic 10: Massive Completeness (MC10). It is possible to assign (possibly in a one-to-many way) knowledge elements of BK to different possible worlds such that the multiset union of these possible worlds covers the BK completely. There is a massive subset of these possible worlds fulfilling (may be different but at least one of) the completeness criteria.

7 Conclusions

The major contributions of this paper can be summarized as follows:

- It presents the first definition and thorough study on basic features of BK in form of its massiveness characteristics. With it we tried to break the state of BK as a 'bare name without interpretation'.
- It proposes, for the first time, a BK model featured with ten MCs, which are both qualitative and quantitative, as a contrast against those nV-defined big data models. With it we tried to eliminate any misunderstanding of BK as something that is important just because it is of a big volume.
- It extends the study of BK to that of BK-S and BK-E, defining their models and characteristics, together with a model of BK-E life cycle, thus forming a complete framework of BK characterization.
- It makes a comprehensive investigation on several large-scale knowledge engineering projects, both social scientific and naturel scientific, by examining the knowledge they produced and the development paradigm they followed, both qualitatively and quantitatively.
- It provides a first sketch of BK-E development standard, which can serve as a reference for the organizers of future BK-E projects, no matter they are government departments, professional organizations or big enterprises.

BK is a new field of artificial intelligence and knowledge engineering. It opens up another window of the information world to both experts and non-experts obsessed with big data. It deserves long range research efforts in the future. For promising research objectives, we propose more precise (domain or problem dependent) characterization of BK, more

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

efficient representation paradigm of BK, upgrading existing large-scale knowledge based systems to make them fit the BK-S standard, transforming ongoing large-scale knowledge engineering projects to BK-E projects and focusing big data mining efforts more on BK mining as the first targets of developing BK research. Possible technical issues include, but are not limited to, classification, semantics, programming languages, distributed processing, inexact reasoning and business applications of BK, as well as BK analytics, BK mining, BK visualization, BK management, BK infrastructures, BK clouds, BK industrial standards, and others.

Acknowledgments

The authors would like to thank Mr. Xikun Huang and Mr. Guilong Chen for providing experimental data on harmonic connections of MC2, and Prof. Cungen Cao, Ms. Shuhan Zhang, and Prof. Fei Wang for their kind help. The core content about BK definition and KG analysis has been given by the first author as invited talks at numerous conferences. This work has been supported by the National Key Research and Development Program of China under grant 2016YFB1000902, NSFC Projects (61472412, 61572473, 61621003, 91746209, and 61772501), the US National Science Foundation (NSF) under grants 1763620 and 1652107, the Program for Yangtze River Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education (China) under grant IRT17R32 (Phase 2), Beijing Science and Technology Project on Machine Learning based Stomatology, and Tsinghua-Tencent-AMSSCJoint Project on WWW Knowledge Structure and its Application.

References

[1] S. Sicular, "Gartner's big data definition consists of three parts, not to be confused with three"v"s." 2018. [Online]. Available: https://www.forbes.com/ sites/gartnergroup/2013/03/27/

[2] X. Wu, H. Chen, G. Wu, J. Liu, Q. Zheng, X. He, A. Zhou, Z.-Q. Zhao, B. Wei, M. Gao, et al., "Knowledge engineering with big data," *IEEE Intell. Syst.*, vol. 30, no. 5, pp. 46–55, Sep./Oct. 2015.

[3] "Big data, bigger opportunities: Investing in information and analytics." 2018. [Online]. Available: https://www.slideshare.net/jmclough/big-data-2013powerpoint

[4] I. Bhandar, "Big data innovation summit," Express Scripts, Boston, Tech. Rep., 2013.

[5] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, 2015, Art. no. 59C64.

[6] "Memex." 2018. [Online]. Available: https://en.wikipedia.org/wiki/Memex

[7] "Memex aims to create a new paradigm for domain-specific search." 2014. [Online]. Avaibale: https://www.darpa.mil/news-events/2014-02-09

[8] J. Gray, "What next? a dozen information-technology research goals," Microsoft Research, Tech. Rep. MSR-TR-99–50, Redmond, 1999.

[9] P. R. Cohen, R. Schrag, E. Jones, A. Pease, A. Lin, B. Starr, D. Gunning, and M. Burke, "The DARPA high-performance knowledge bases project," *AI Mag.*, vol. 19, no. 4, 1998, Art. no. 25.

[10] K. Zetter, "DARPA is developing a search engine for the dark web," https://www.wired.com/2015/02/darpa-memex-dark-web/, Feb. 10, 2015.

[11] E. A. Feigenbaum, "Some challenges and grand challenges for computational intelligence," *J. ACM*, vol. 50, no. 1, pp. 32–40, 2003.

[12] "OpenCyc." [Online]. Available: http://www.baike.com/wiki/OpenCyc

[13] D. B. Lenat and R. V. Guha, *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing, 1989.

[14] C. Cao, "Significance of the national knowledge infrastructure," *Bulletin Chinese Acad. Sci.*, vol. 16, no. 4, pp. 255–259, 2001.

[15] C. Cao, S. Wang, and L. Jiang, "A practical approach to extracting names of geographical entities and their relations from the web," in *Proc. 7th Int. Conf. Knowl. Sci. Eng. Manage.*, 2014, pp. 210–221.

[16] Q. Feng, C. Cao, Y. Sui, Y. Zheng, and Q. Qin, "Masaq: A multi-agent system for answering questions based on an encyclopedic knowledge base," *in Proceedings of the 2nd International Workshop on Declarative Agent Lang. and Technol. (DALT'04)*, 2004, pp. 109–124.

[17] L. Liu, S. Zhang, L. Diao, and C. Cao, "An iterative method of extracting Chinese ISA relations for ontology learning," *J. Comput.*, vol. 5, no. 6, 2010, Art. no. 871.

[18] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011. [Online]. Available: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation, McKinsey Global Institute.

[19] D. Che, M. Safran, and Z. Peng, "From big data to big data mining: Challenges, issues, and opportunities," in *Proc. 18th Int. Conf. Database Syst. Adv. Appl.*, 2013, vol. 7827, pp. 1–15.

[20] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[21] X. Wu, J. He, R. Lu, and N. Zheng, "From big data to big knowledge: HACE + BigKE," *Acta Automatica Sinica*, vol. 42, no. 7, pp. 965–982, 2016.

[22] K. Murphy, "From big data to big knowledge," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 1917–1918.

[23] A. Russell, "Turning big data into big knowledge." 2018. [Online]. Available: http://socialscience.ucdavis.edu/iss-journal/features/turning-big-data-into-big-knowledge

[24] J. Aaron, "Turning big data into big knowledge." 2018. [Online]. Available: https://thestack.com/big-data/2015/12/07/turning-big-data-into-big-knowledge/

[25] J. Liebowitz, "How to extract big knowledge from big data?" 2018. [Online]. Available: https://www.sas.com/en_us/insights/articles/big-data/big-knowledge-big-data.html

[26] 2016. [Online]. Available: https://www.linkedin.com/pulse/analyst-toolkit-from-big-data-knowledge-dr-shay-hershkovitz

[27] M. Sader and A. Lewis, Eds., *Encyclopedias, Atlases and Dictionaries*. New York, NY, USA: RR Browker, 1995.

[28] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," *Int. J. Lexicography*, vol. 3, no. 4, pp. 235–C244, 1990.

[29] Z. Luo, "Big Chinese Dictionary," Shanghai Lexicographical Publishing House, Shanghai, 2016.

[30] Y. Jia, Y. Wang, X. Cheng, X. Jin, and J. Guo, "OpenKN: An open knowledge computational engine for network big data," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, Aug. 2014, pp. 657–664.

[31] X. Jin, "Big data knowledge graph and its applications," Inst. Comput. Technol., Chinese Acad. Sci., Beijing, China, Tech. Rep. TR-2016-07-06, 2016.

[32] "Read the Web: Research project at Carnegie Mellon University," 2017. [Online]. Available: http://rtw.ml.cmu.edu/rtw/

[33] "Wikipedia," 2017. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia

[34] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum, "YAGO2: Exploring and querying world knowledge in time, space, context, and many languages," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 229–232.

[35] C. Eder, Chinese museum needs your help decoding ancient script, https://www.bookstr.com/chinese-museum-needs-your-help-decoding-ancient-script, Jul. 25, 2017.

[36] M. Martin, "It's not your grandpa's planetarium anymore." 2010 [Online]. Available: http://www.technewsworld.com/story/71409.html

[37] "System." 2018. [Online]. Availanle: https://en.wikipedia.org/wiki/System

[38] "The 5th comprehensive investigation report on Shanghai's traffic." 2018. [Online]. Available: https://wenku.baidu.com/view/11794001b9f3f-90f77c61b4b.html?from=search

[39] "Xia-Shang-Zhou chronology proj." 2018. [Online]. Available: https://baike.baidu.com/

[40] N. Yue, *Archaeology Project in China: The Whole Story of Xia-Shang-Zhou Chronology Project*. Hainan Publishing Co., Haikou, 2007.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

[41] Q. Sima, *Historical Records*. South San Francisco, CA, USA: China Book Press, 2007.

[42] "Xia-Shang-Zhou chronological table." [Online]. Available: https://baike.baidu.com/item/

[43] S. Allan, *The formation of Chinese Civilization: An Archaeological Perspective*, K. C. Chang and P. Xu, Eds. New Haven, CT, USA: Yale University Press and New World Press, 2005.

[44] "Chinese ancient oracles." 2018. [Online]. Available: https://f11.baidu.com/it/u=1894150616,1400725470&fm=72

[45] "Rubbings of chinese ancient oracle of female general Fu Hao." 2018. [Online]. Available: https://image.baidu.com/search/index?tn=baiduimage&ipn=r&ct=201326592&cl=2&lm=-1&st=-1&fm=result&fr=&sf=1&fmq=1489300190737

[46] "'Zun' in square form with four sheep." 2018. [Online]. Available: https://baike.baidu.com/item/

[47] "Troy" 2018. [Online]. Available: https://en.wikipedia.org/wiki/Troy

[48] "Trojan war." 2018. [Online]. Available: https://en.wikipedia.org/wiki/Trojan_War

[49] Canakkale, "New term excavations start at city of Troy with turkish team," Dogan News Agency, 2014. [Online]. Available: hurriyetdailynews.com

[50] T. Bryce, *The Trojans and Their Neighbours*. New York, NY, USA: Taylor & Francis, 2005.

[51] M. Askin, *Troy : With Legends, Facts, and New Developments*. Keskin Color, Istanbul, 2005.

[52] "Epic cycle." 2018. [Online]. Available: https://en.wikipedia.org/wiki/Epic_Cycle

[53] "Human genome project." 2018. [Online]. Available: https://en.wikipedia.org/wiki/Hum an_Genome_Project

[54] "Overview of human genome project," 2016. [Online]. Available: https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

[55] "English-Chinese human gene dictionary." 2018. [Online]. Available: https://book.dou ban.com/subject/6023540/

[56] "Human phenotype ontology." 2018. [Online]. Available: http://human-phenotype-ontology.github.io/.

[57] "Human gene mutation database." 2018. [Online]. Available: http://www.hgmd.cf.ac.uk/ac/index.php

[58] 2018. [Online]. Available: http://apps.nhlbi.nih.gov/Grasp/Overview.aspx.

[59] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, and L. Hindorff, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, pp. D1001–D1006, 2014.

[60] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson, "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic Acids Rese.*, vol. 45, pp. D896–D901, 2017.

[61] X. Sun, M. Gong, and S. Zhang, "Clinical practice of precision medicine (in Chinese)," *Sci. Technol. Rev.*, vol. 35, no. 16, pp. 13–20, 2017.

[62] "Strategy for UK life sciences," 2012. [Online]. Available: https://www.gov.uk/government/publications/strategy-for-uk-life-sciences-one-year-on

[63] "China constructed world largest Han nationality MHC database," 2018. [Online]. Available: http://www.huaxiahealthcare.com/union.php?class id=36&view-id=1515

[64] "Wikipedia statistics," Sep. 2017. [Online]. Available: https://stats.wikimedia.org/EN/Sitemap.htm

[65] "DBpedia version 2014 released." 2014. [Online]. Available: http://wiki.dbpedia.org/news/dbpedia-version-2014-released

[66] "DBpedia." 2018. [Online]. Available: https://en.wikipedia.org/wiki/DBpedia

[67] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., "DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[68] G. Chen, "On the harmonic connectivity of wikipedia articles," Inst. Comput. Technology, Chinese Acad. Sci., Beijing, China, Tech. Rep. TR-2018-02-09, 2018.

[69] "Reliability of Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Reliability_of_Wikipedia

[70] R. Quillan, *A Notation for Representing Conceptual Information: An Application to Semantics and Mechanical English Paraphrasing*. Systems Development Corporation, Santa Monica, 1963.

[71] C. J. Fillmore, "The case for case," In E. Bach & R. T. Harms (Eds.), *Univ. Linguistic Theory*, pp. 1–89, New York, NY: Holt, Rinehart, and Winston, 1968.

[72] "CODASYL." [Online]. Available: https://en.wikipedia.org/wiki/CODASYL

[73] P. P.-S. Chen, "The entity-relationship model – toward a unified view of data," *ACM Trans. Database Syst.*, vol. 1, no. 1, pp. 9–36, Mar. 1976.

[74] J. F. Sowa, "Conceptual graphs for a data base interface," *IBM J. Res. Develop.*, vol. 20, no. 4, 1976, pp. 336–357.

[75] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.

[76] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 601–610.

[77] [Online]. Available: https://developers.facebook.com/docs/graph-api.

[78] "Probase." [Online]. Available: https://www.microsoft.com/en-us/research/project/probase/

[79] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 481–492.

[80] F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A core of semantic knowledge unifying wikipedia and wordnet," in *Proc. 16th Int. World Wide Web Conf.*, 2007, pp. 697–706.

[81] F. Mahdisoltani, J. Biega, and F. Suchanek, "YAGO3: A knowledge base from multilingual wikipedias," in *Proc. 7th Biennial Conf. Innovative Data Syst. Res.*, 2015.

[82] C. Cao, Q. Feng, Y. Gao, F. Gu, J. Si, Y. Sui, W. Tian, H. Wang, L. Wang, Q. Zeng, and X. Zhou, "Progress in the development of national knowledge infrastructure," *J. Comput. Sci. Technol.*, vol. 17, no. 5, pp. 523–534, Aug. 2002.

[83] C. Cao, Y. Sun, Y. Sui, and Q. Zeng, "On representation of mathematical knowledge in CNKI," *J. Softw.*, vol. 17, no. 8, pp. 1731–1742, 2006.

[84] I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C. M. Meyer, and C. Wirth, "Uby: A large-scale unified lexical-semantic resource based on lmf," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 580–590.

[85] Z. Dong, Q. Dong, and C. Hao, *Hownet and the Computation of Meaning*. River Edge, NJ, USA: World Scientific, 2006.

[86] Z. Dong, Q. Dong, and C. Hao, "Hownet and its computation of meaning," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 53–56.

[87] "FrameNet." 2018. [Online]. Available: https://en.wikipedia.org/wiki/FrameNet

[88] R. Speer and C. Havasi, "Representing general relational knowledge in ConceptNet 5," in *Proc. 8th Int. Conf. Lang. Resources Eval.*, 2012, pp. 3679–3686.

[89] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, Jr., and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 1306–1313.

[90] Y. Huang, "Muisenet: A web-based english knowledge graph," Inst. Comput. Technol., Chinese Acad. Sci., Beijing, China, Tech. Rep. TR-2016-06-04, 2016.

[91] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-scale information extraction in knowitall: (preliminary results)," in *Proc. 13th Int. Conf. World Wide Web*, 2004, pp. 100–110.

[92] O. Etzioni, A. Fader, J. Christensen, and S. Soderland, "Open information extraction: The second generation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 3–10.

[93] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1535–1545.

[94] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2007, pp. 2670–2676.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

[95] M. Banko, "Open information extraction for the web," Ph.D. dissertation, Department of Computer Science and Engineering Univ. Washington, Seattle, WA, USA, 2009.
[96] C. Zhang, C. Cao, Y. Sui, and X. Wu, "A Chinese time ontology for the Semantic Web," *Knowl.-Based Syst.*, vol. 24, no. 7, pp. 1057–1074, 2011.
[97] M. Farber, B. Ell, C. Menne, and A. Rettinger, "A comparative survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO," *Semantic Web*, 2015. [Online]. Available: http://www.semantic-web-journal.net/system/files/swj1141.pdf
[98] M. Farberr, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO," *Semantic Web*, vol. 9, no. 1, pp. 77–129, 2018.
[99] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2017.
[100] S. Miran, "A survey of open information extraction systems," University of Maryland, College Park, MD, USA, Tech. Rep. Project Report CS 290D, 2014.
[101] "Microsoft concept graph for short text understanding." 2016. [Online]. Available: https://concept.research.microsoft.com/Home/Introduction
[102] "Explore Freebase data," 2013. [Online]. Available: http://www.freebase.com
[103] A. Akesson, "Google my business profiles start ranking in non-branded searches." 2018. [Online]. Available: https://www.venndigital.co.uk/blog/google-my-business-profiles-start-ranking-in-non-branded-searches-78887/
[104] J. Jarvis, "Google knowledge graph has more than 70 billion facts," Oct. 2016. [Online]. Available: https://twitter.com/jeffjarvis/status/783338071316135936
[105] 2018. [Online]. Available: https://searchengineland.com/googles-knowledge-graph-errors-126098
[106] Z. Wang, J. Li, Z. Wang, S. Li, M. Li, D. Zhang, Y. Shi, Y. Liu, P. Zhang, and J. Tang, "XLore: A large-scale English-Chinese bilingual knowledge graph," in *Proc. 12th Int. Semantic Web Conf. Posters Demonstrations Track*, vol. 1035, 2013, pp. 121–124.
[107] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.
[108] X. Huang, "On the harmonic connectivity of knowledge graphs," Acad. Math. Syst. Sci., Chinese Acad. Sci., Beijing, China, Tech. Rep. TR-2018-02-07, 2018.
[109] E. Feigenbaum and P. McCorduck, *The fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. Reading, MA, USA: Addison-Wesley/Longman, 1983.
[110] L. Bellomarini, G. Gottlob, A. Pieris, and E. Sallinger, "Swift logic for big data and knowledge graphs," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1–10.
[111] Q. Wang, B. Wang, and L. Guo, "Knowledge base completion using embeddings and rules," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 1859–1865.
[112] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 990–998.
[113] J. Cao, Z. Wu, Y. Wang, and Y. Zhuang, "Hybrid collaborative filtering algorithm for bidirectional web service recommendation," *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 607–627, 2013.
[114] M. F. Goodchild, H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, M. Ehlers, J. van Genderen, D. Jackson, et al., "Next-generation digital earth," *Proc. Nat. Acad. Sci. United States America*, vol. 109, no. 28, pp. 11 088–11 094, 2012.

Ruqian Lu received the diploma degree in mathematics from Jena University, Germany, in 1959. He is a fellow of the Chinese Academy of Sciences, professor with the Institute of Mathematics, Academy of Mathematics and Systems Science. He is holding a concurrent professorship with the Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include artificial intelligence, knowledge engineering, knowledge based software engineering, programming languages and formal semantics, quantum logic, and quantum computation. He has published more than 200 papers and authored a dozen books. He has received China's National Second Class and CAS's First Class Prize for Progress in Science and Technology (twice), the Hua Luogeng Mathematics Prize from China's Mathematics Society, and the Lifelong Achievement Prize from the China Computer Federation (CCF).

Xiaolong Jin received the PhD degree in computer science from Hong Kong Baptist University, in 2005. He is currently a professor with the CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences (CAS). His research interests include knowledge graph, knowledge engineering, social computing, social networks, etc. He has coauthored two monographs published by Springer and the Tsinghua University Press, and published more than 150 papers in prestigious international journals, including the *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, the *ACM Transactions on the Web*, the *ACM Transactions on Intelligent Systems and Technology*, the *IEEE Transactions on Wireless Communications*, the *IEEE Transactions on Parallel and Distributed Systems*, and conferences, including AAAI, IJCAI, CIKM, WWW, GLOBECOM, ICC, and AINA. He received the Best (Student) Paper Awards of IEEE ICBK 2017, IEEE CIT-2015, CCF Big Data 2015, IEEE AINA 2007, and IEEE ICAMT 2003.

Songmao Zhang received the PhD degree from the Institute of Mathematics, Chinese Academy of Sciences (CAS), in 1992. She has been a full-time professor with the Academy of Mathematics and Systems Science, CAS, since 2007. In addition, she was a visiting scholar in research institutions and universities in the US, Australia, Germany, France, and United Kingdom. Within the area of artificial intelligence, her research interests include ontology matching, knowledge representation and reasoning in the Semantic Web, AI-based automatic animation, data mining, and natural language understanding. She has been the principal investigator of many national projects, including those from NSF of China, National 863 High-Technology Foundation of China, and from the Chinese Academy of Sciences.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.

Meikang Qiu received the BE and ME degrees from Shanghai Jiao Tong University and received PhD degree of computer science from the University of Texas at Dallas. Currently, he is a faculty member at Columbia University. He is an IEEE Senior member and an ACM Senior member. He is the chair of the IEEE Smart Computing Technical Committee. His research interests include cyber security, big data analysis, cloud computing, smarting computing, intelligent data, embedded systems, etc. A lot of novel results have been produced and most of them have already been reported to the research community through high-quality journal and conference papers. He has published four books, and 400 peer-reviewed journal and conference papers (including 200+ journal articles, 200+ conference papers, and 70 + IEEE/ACM Transactions papers). His paper published in the *IEEE Transactions on Computers* about privacy protection for smart phones was selected as a highly cited paper in 2017. His paper about embedded system security published in the *Journal of Computer and System Science* (Elsevier) was recognized as a highly cited paper in both 2016 and 2017. His paper about data allocation for hybrid memory published in the *IEEE Transactions on Computers* was selected as a hot paper (1 in 1000 papers) in 2017. His paper on Tele-health system won the IEEE System Journal 2018 Best Paper Award. He also won the ACM Transactions on Design Automation of Electrical Systems (TODAES) 2011 Best Paper Award. He has won another 10+ Conference Best Paper Awards in recent years. Currently, he is an associate editor of 10+ international journals, including the *IEEE Transactions on Computers* and *IEEE Transactions on Cloud Computing*. He has served as leading guest editor for an *IEEE Transactions on Dependable and Secure Computing* (*TDSC*), special issue on social network security. He is the general chair/program chair of a dozen of IEEE/ACM international conferences, such as IEEE TrustCom, IEEE BigDataSecurity, IEEE CSCloud, and IEEE HPCC. He won Navy Summer Faculty Award in 2012 and Air Force Summer Faculty Award in 2009. His research is supported by US government such as NSF, NSA, Air Force, Navy and companies such as GE, Nokia, TCL, and Cavium.

Xindong Wu received the bachelor's and master's degrees in computer science from the Hefei University of Technology, China, and the PhD degree in artificial intelligence from the University of Edinburgh, United Kindom. He is a Yangtze River scholar with the Research Institute of Big Knowledge at the Hefei University of Technology, China, and a professor with the School of Computing and Informatics, University of Louisiana at Lafayette. His research interests include data mining, knowledge engineering, and the World Wide Web. He was the editor-in-chief of the *IEEE Transactions on Knowledge and Data Engineering (TKDE)* between 2005 and 2008. He served as a program committee chair/co-chair for ICDM 03 (the 2003 IEEE International Conference on Data Mining), KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), and CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management, and ICBK 2017 (the 8th IEEE International Conference on Big Knowledge). He is a fellow of the IEEE and the AAAS.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on September 27,2022 at 14:58:39 UTC from IEEE Xplore. Restrictions apply.