sustainability

MDPI

Article

A Method for Constructing Geographical Knowledge Graph from Multisource Data

Xuan Guo [1], Haizhong Qian [1], Fang Wu [2] and Junnan Liu [2,*]

[1] Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China; gx930707@163.com (X.G.); haizhongqian@163.com (H.Q.)
[2] School of Earth Science and Technology, Zhengzhou University, Zhengzhou 450001, China; dtis123@163.com
* Correspondence: lpjunnan.love@163.com; Tel.: +86-0371-6368-5366

Abstract: Global problems all occur at a particular location on or near the Earth's surface. Sitting at the junction of artificial intelligence (AI) and big data, knowledge graphs (KGs) organize, interlink, and create semantic knowledge, thus attracting much attention worldwide. Although the existing KGs are constructed from internet encyclopedias and contain abundant knowledge, they lack exact coordinates and geographical relationships. In light of this, a geographical knowledge graph (GeoKG) construction method based on multisource data is proposed, consisting of a modeling schema layer and a filling data layer. This method has two advantages: (1) the knowledge can be extracted from geographic datasets; (2) the knowledge on multisource data can be represented and integrated. Firstly, the schema layer is designed to represent geographical knowledge. Then, the methods of extraction and integration from multisource data are designed to fill the data layer, and a storage method is developed to associate semantics with geospatial knowledge. Finally, the GeoKG is verified through linkage rate, semantic relationship rate, and application cases. The experiments indicate that the method could automatically extract and integrate knowledge from multisource data. Additionally, our GeoKG has a higher success rate of linking web pages with geographic datasets, and its exact coordinates have increased to 100%. This paper could bridge the distance between a Geographic Information System and a KG, thus facilitating more geospatial applications.

Keywords: knowledge graph; geographical knowledge graph; knowledge extraction; geographic dataset; internet encyclopedias

Citation: Guo, X.; Qian, H.; Wu, F.; Liu, J. A Method for Constructing Geographical Knowledge Graph from Multisource Data. *Sustainability* **2021**, *13*, 10602. https://doi.org/10.3390/su131910602

Academic Editor: Anna Visvizi

Received: 28 July 2021
Accepted: 22 September 2021
Published: 24 September 2021

1. Introduction

In the 2020s, the world has been experiencing the most significant challenges regarding natural disasters and worldwide epidemics. It is clear that these global problems are geospatial—they all occur at a particular location on or near the Earth's surface [1]. At the junction of artificial intelligence (AI) and big data, geographical artificial intelligence has attracted much attention worldwide, and plays an essential role in science and technologies [2,3]. As the backbone of AI, knowledge graphs (KGs) have shown their powerful capabilities in different kinds of intelligent applications, including data retrieval, integration, analysis, etc., [4]. Geographical knowledge graphs (GeoKGs), a kind of domain KG, can organize, interlink, and infer geospatial knowledge; hence they offer excellent opportunities to solve many problems in real life [5].

Geographical knowledge is a higher level of geospatial information, represented by ontology and the semantic web [6]. For example, the knowledge "The Yellow River is the longest river in China" is described as <Yellow River, Longest River, China>, in which Yellow River and China are entities, and Longest River is a relationship. Although most of these techniques focus on representing geographical knowledge, none involve representing knowledge in multisource data. Moreover, the existing KGs are seldom constructed from geographic datasets, thus posing several challenges to the construction of complete geographical knowledge [7]. The first challenge is that most of the existing

KGs lack geographical entities and precise coordinates. The second challenge is that it is hard to extract knowledge from multisource data because these entities are organized in different data sources, and their geometry types are complex [8]. The last challenge faced is the diversified spatial relationships between entities, which KG does not consider [9,10].

Using a new GeoKG construction method based on geographic datasets, we expand the scope of usability of geographical data on the internet. Firstly, a system framework is presented to model domain knowledge in a schema layer and extract geographical knowledge in a data layer, thus guiding GeoKG construction.In the schema layer, concepts and their relationships are constructed to represent geographical knowledge. In the data layer, several methods are proposed to extract knowledge, including extracting geographical entities, transforming attributes into triples, extracting concepts and attributes from encyclopedias, and integrating knowledge from multisource data. Finally, the linkage rate and the semantic relationship rate are analyzed, and some application cases are exhibited to verify the ability to obtain related knowledge. The specific contributions are the following:

- A GeoKG construction framework is proposed, which extracts knowledge from geographic datasets, and completes these knowledge sets using internet encyclopedias. At the same time, the framework can become a reference for other KGs involved in the same space;
- A schema layer for our GeoKG is designed, by which geographic datasets can be formalized to represent geographical knowledge, thus restricting RDF triples in the data layer;
- A geographical knowledge extraction method is proposed, by which the entities and attributes belonging to multiple layers, features, and geometries are composed, thus constructing coordinates and spatial relationships for GeoKG.

The remainder of the paper is structured as follows. Section 2 reviews related work regarding KGs and GeoKGs. In Section 3, the system framework for constructing GeoKGs is proposed. Section 4 exploits the linkage rate and semantic relationship rate, and then demonstrates application cases. Finally, the discussion, conclusions, and future work are discussed in Sections 5 and 6.

2. Geographical Knowledge Graph-Related Literature

In recent years, researchers have paid much attention to GeoKG, the fundamental techniques of which are knowledge representation and KG construction. As such, several research areas and backgrounds related to GeoKG are reviewed.

2.1. Introduction to Knowledge Graph

The KG was proposed by the Google Knowledge Graph project to devise a more intelligent search engine. It consists of concepts, entities, literature and relationships, and focuses on extracting and fusing knowledge from online encyclopedias. Furthermore, it enables the semantic search to understand the query intention better, thus providing more concise results. Taking the search sentence of "the length of the Yellow River" as an example, Google can return a knowledge card, and provide an accurate answer of 5464 km based on KG. As shown in Figure 1, when we search for "Yellow River" in the Google search engine, the related web pages will be presented on the left side, and the attributes (such as length, area, headstream, picture of the river, etc.) will be shown on the right side.

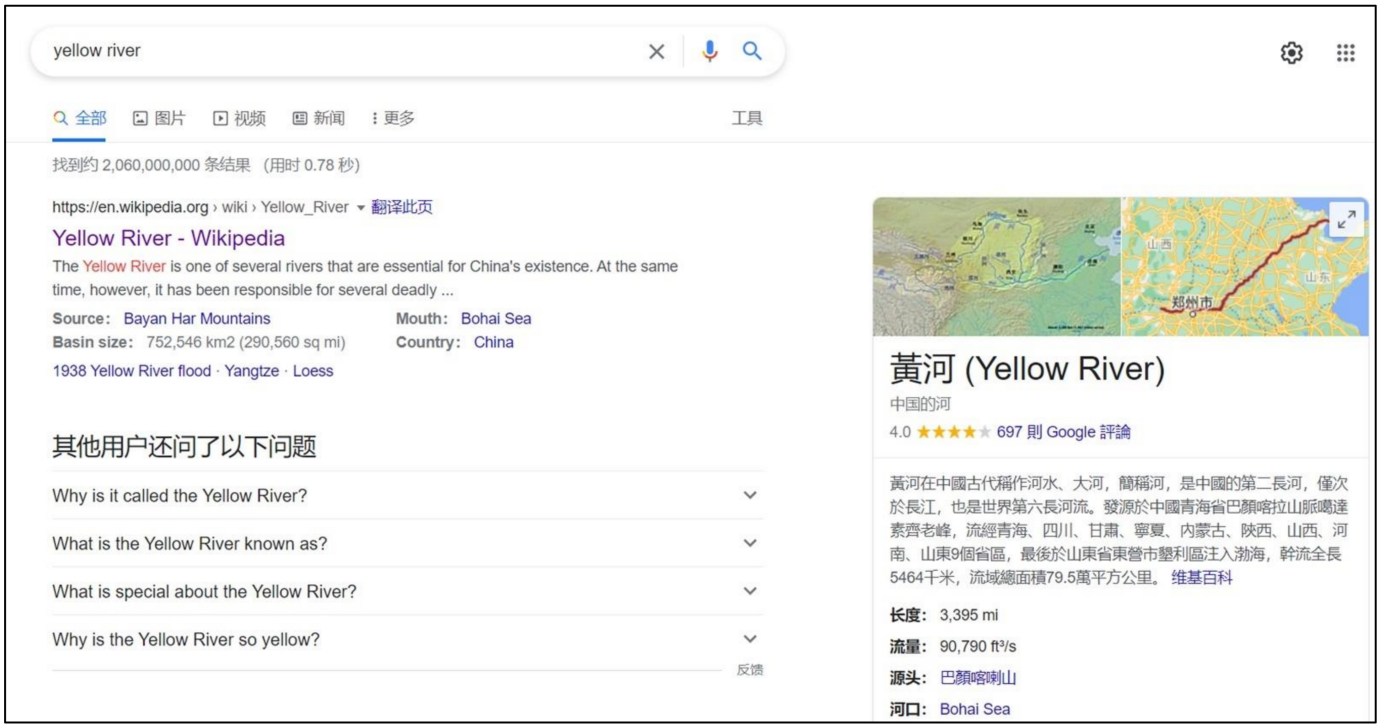

Figure 1. The search results for "Yellow River" in Google search engine.

Nowadays, KGs have become prevalent, and there are some famous KGs. In CN-DBPedia [11], for example, three Chinese internet encyclopedias (i.e., Baidu Baike, Chinese Wikipedia, and Hudong Baike) are used to extract knowledge. Over the past 30 years, many researchers have carried out related work. Hook summarized six application aspects of KGs [12] and Zeyua introduced KGs to explore the scientific literature distribution [13].

2.2. Techniques of Geographical Knowledge Graph

2.2.1. Geographical Knowledge Representation

Geographical knowledge representation can be considered the core idea in GeoKGs. In terms of the representation model, Zheng et al. [14] proposed a model based on spatiotemporal processes, while Kacprzyk et al. [15] represented a method employing chains of contexts and patterns of appropriate user behavior in visual analysis. To better represent these fields of knowledge, Mehdi Mekni [16] proposed a virtual geographic environment using a topologic graph of geographic datasets. Similarly, Laurini [17] presented a conceptual framework to manage geographic entities, relationships, and rules. Unlike prior work, Jiang et al. [5] divided geographical knowledge into factual knowledge and process knowledge to describe external characteristics and spatial transformation.

In the last few decades, the use of the semantic web and ontology in knowledge representation have developed considerably [18]. This has fostered a promising way to connect spatial data with KG, thus augmenting the application of geographic datasets [19]. Hence, many geographic datasets have been published in the form of Linked Data, some of which play a prominent role in the Linked Open Data cloud (https://lod-cloud.net/, accessed on 21 September 2021). More governmental agencies and large-scale data infrastructures run Linked Data initiatives, such as e-Government and open data communities in Europe [20]. Furthermore, Varanka and Usery [21] proposed that the geographic data released in RDF can be treated as knowledge; hence, most of the existing techniques focus on ontologies and rules. For instance, Janowicz [22] modeled semantic knowledge in geographic datasets, and Hofer et al. [23] formalized geographic operators. Additionally, Gould and Mackaness [24] used ontologies to formalize generalized cartography knowledge to facilitate the sharing, expansion, and reuse of mapping knowledge.

2.2.2. Geographical Knowledge Graph Construction

As an expanded KG, the GeoKG is a structured semantic knowledge base, which represents rich geographical knowledge in triples [2]. It is regarded as a promising tool to deal with many technical geographic challenges, such as named entity recognition, toponym disambiguation, and spatial reasoning [5]. In GeoKG, RDF triples are used to describe knowledge, and their visualization relies on a "node–edge" graph (Figure 2). In detail, geographical concepts are represented by nodes, and edges demonstrate relationships in data properties (i.e., the relationship between entities and attributes) and object properties (i.e., the relationship between concepts and entities). As illustrated in Figure 2, the triples <Yellow River, is-a, River> and <Yellow River, inside, China> indicate object property relationships between concepts and entities, and triple <Yellow River, length, "5464 km"> represents a data property. Therefore, GeoKG could link different datasets based on RDF triples, thus enriching geographical knowledge [25].

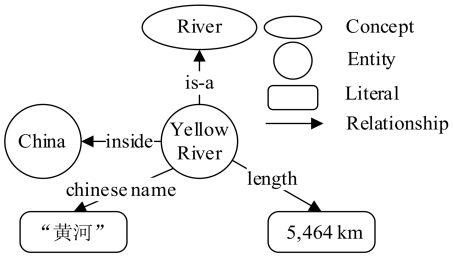

Figure 2. An example of a part of a GeoKG.

Some GeoKGs have been constructed from geographic datasets, such as OSMonto, OSM Semantic Network, Yago2, etc., OSMonto [26] is an ontology for Open Street Map (OSM) tags, and the OSM Semantic Network [27] contains RDF triples extracted from OSM tags on Wiki websites. Although OSMonto and OSM Semantic Network extract a large number of concepts, they do not contain geographical entities or employ common-sense knowledge. In addition to concepts, Yago2 [28] extracts entities from Wikipedia. However, it does not contain a lot of geographical entities or Chinese information, because Wikipedia contains only a few Chinese pages. Furthermore, Liu et al. [29] showed that Linked Data have made considerable progress in publishing, retrieving, and integrating data. Based on Linked Data, LinkedGeoData could map OSM into RDF triples to devise a geographic data browser [30]. In terms of GeoKG construction, Chen et al. [2] presented a crowdsourced geographic knowledge graph that extracted different kinds of entities from OSM and enriched them with human geographic knowledge from Wikidata. Under the "One Belt One Road" initiative, Wu et al. [4] introduced the techniques for constructing a Chinese knowledge graph, which have greatly promoted the development of AI.

In summary, these proposed methods provide abundant semantics. However, general KGs lack geographical knowledge. Moreover, most of them are only constructed from internet encyclopedias, and they cannot extract knowledge from geographic datasets, thus lacking precise coordinates and spatial relationships.

3. GeoKG Construction Techniques

3.1. The System Framework of GeoKG Construction

The general KG is constructed via a "down-top" approach [4]. In contrast, GeoKG is constructed via a "top-down" method, consisting of two stages: designing the schema layer, and extracting geographical knowledge in the data layer (Figure 3). The schema layer is used to construct concepts and relationships. In the data layer, methods of extracting, integrating, and storing geographical knowledge are discussed sequentially. Then, concepts and relationships in the schema layer can be completed by generalizing knowledge in the data layer.

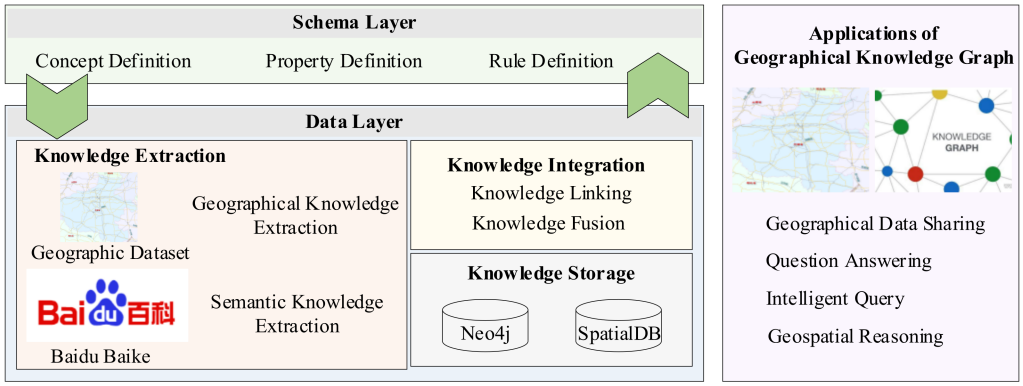

Figure 3. The system framework of geographical knowledge graph construction.

The first stage of GeoKG construction is schema layer modeling. Hu et al. [31] proposed two design patterns to design the schema layer, including content and logical patterns. The content pattern is adopted to formalize relationships and the geographical concepts of entity, feature, geometry, coordinate, and reference system.

In the data layer, knowledge is automatically extracted from geographic datasets and Baidu Baike. Because it is extracted from multisource data, knowledge integration methods of linking and fusing are adopted to integrate equivalent entities and concepts. With current technology, a single database cannot directly store knowledge and geographic datasets. Neo4j is one of the best graph database management systems, and Spatialite is a database engine with a spatial plugin. Both are used to store extracted knowledge to meet application demands.

3.2. Available Data Analysis

The available data sources for constructing GeoKGs include geographic datasets and Baidu Baike.

3.2.1. Geographic Datasets

Geographic datasets are carriers of spatial information that meet the demands of production units and social masses. As a primary type of geographic data, vector data are hierarchical, block-divided, and feature-divided. They consist of two components: one managing spatial data (i.e., geometry) and the other managing thematic data. Vector data represent elements in the form of points, lines, and polygons based on mathematical projection, thus demonstrating locations explicitly. Furthermore, they can easily represent spatial distribution and topological structure because they are stored in a two-dimensional Cartesian coordinate system. Geographic datasets could also be stored in a spatial database in the form of several tables. In each table, rows represent features, columns display attribute values, and geometry columns express coordinates (Figure 4). Therefore, it is easy to operate the spatial database through Structured Query Language (SQL).

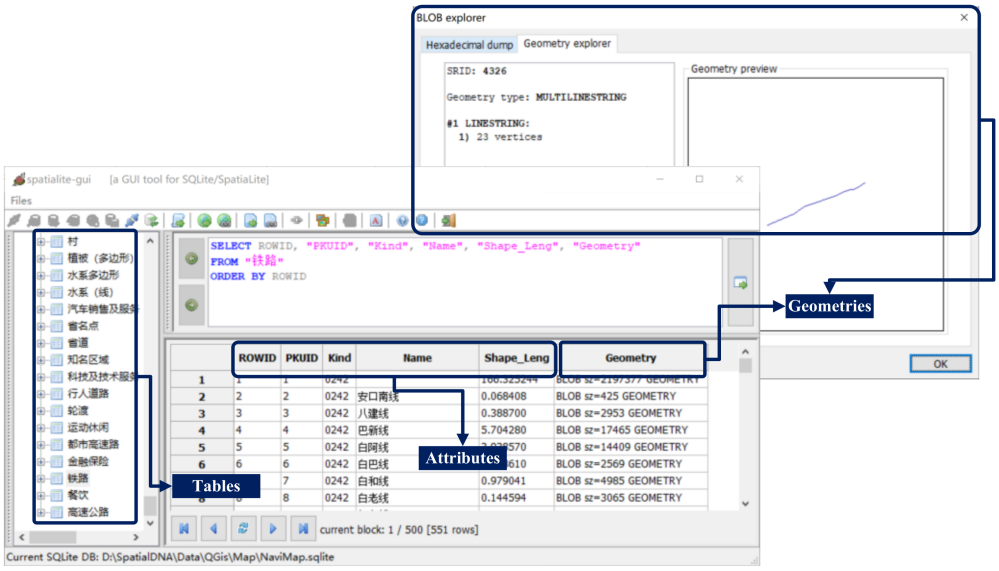

Figure 4. An example of a geographic dataset in a spatial database.

However, it is inefficient to query data across different tables, and semantics in geographic datasets are weak. Hence, new ways of organizing geographic datasets are required, and internet encyclopedias should be introduced to complete semantics in geographic datasets.

3.2.2. Baidu Baike

Baidu Baike is the most popular internet encyclopedia in China. It has some advantages, such as covering a wide range of fields, allowing users to edit almost all accessible pages, and expressing entities in the form of web pages. On each web page, labels, images, and information boxes are used to describe entity characteristics.

3.3. Schema Layer Modeling

According to Section 3.1, the schema layer is conceptualized and implemented to integrate geographical knowledge.

3.3.1. Concepts Design

1. Features and Geometries

Knowledge sharing and cyclic utilization are the primary functions of the modeling schema layer. GeoSPARQL (http://www.opengis.net/ont/geosparql, accessed on 21 September 2021) ontology is introduced to express geographical knowledge about features and geometries. In the following, the prefixes geo and sf are used to represent the namespaces of GeoSPARQL and simple feature geometries, respectively.

As shown in Figure 5, there are some existing concepts and relationships in GeoSPARQL. To represent geographical knowledge in vector data, the class *SpatialObject* is created as an extended concept, and all the other concepts are inherited from it directly or indirectly. The object property *spatialRelation* is used to connect *SpatialObject*. The concepts *Feature* and *Geometry* are constructed as subclasses of *SpatialObject*, and *Feature* is linked to one or more *Geometry* using the object property *hasGeometry*. Two literals are associated with the concept *Geometry* via the data properties *asWKT* and *EPSG*, which store coordinates in well-known text (WKT) and the spatial reference system of the European Petroleum Survey Group (EPSG). Moreover, the concepts point, curve, surface, and geometry collection are inherited from *Geometry* to represent geometries in geographic datasets.

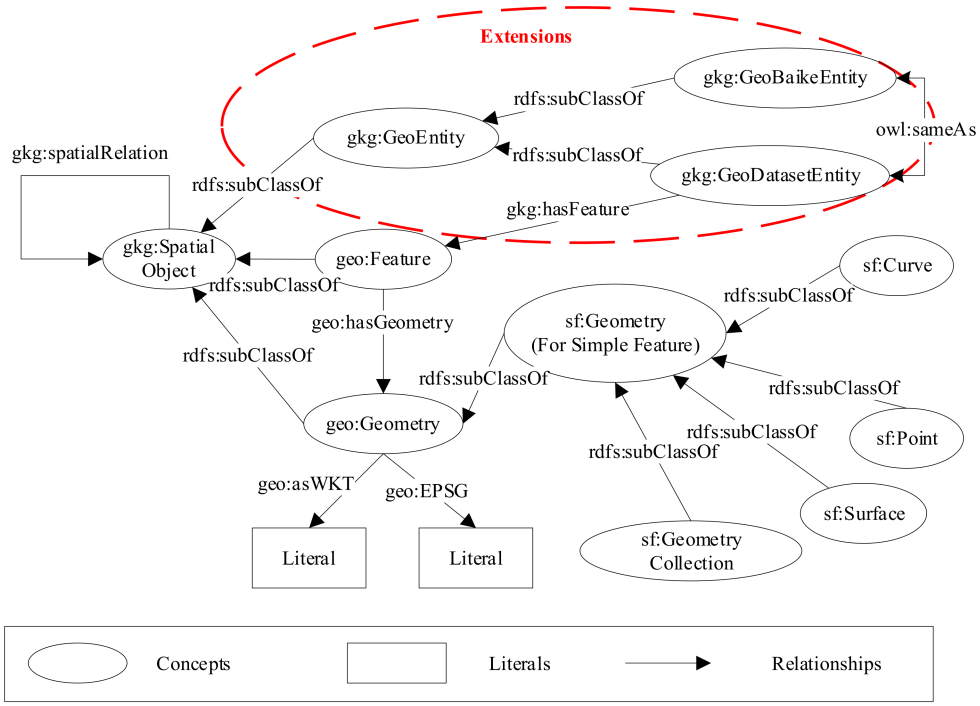

Figure 5. Concepts and relationships in the schema layer.

2. Entities

As shown in Figure 5, a prefix *gkg* is used to limit the knowledge scope, such as the concept *gkg:GeoEntity* and object property *gkg:spatialRelation*. Moreover, two disjointed subclasses, named *gkg:GeoBaikeEntity* and *gkg:GeoDatasetEntity*, are created to represent the geographical entities extracted from Baidu Baike and vector data, respectively. The object property *sameAs* represents the linkage between instances of two concepts, thus integrating geographical entities in multisource data. Furthermore, *GeoDatasetEntity* is connected to one or more *Feature* concept by the object property *hasFeature*. Therefore, an entity can represent its geographical information and semantics simultaneously.

3.3.2. Relationships Design

In addition to concepts, relationships also play a significant role in formalizing the real world. As shown in Figure 6, GeoKG consists of two types of geographical relationships: spatial and semantic relationships. The semantic relationship is divided into data property and object property. The object property consists of *subclassOf*, *equivalentClass*, *is-a*, and *sameAs*. *SubClassOf* and *equivalentClass* formalize parent–child relationships and equivalence between concepts, respectively. The relationship *is-a* associates concepts and instances, and *sameAs* defines the same geographical entities. Moreover, data properties (e.g., name, width, length, EPSG, etc.) are used to describe geographical entity's attributes. In addition to semantic relationships, topological, distance, and orientation relationships are crucial in GeoKG. In the following subsections, each of these spatial relationships will be described in detail.

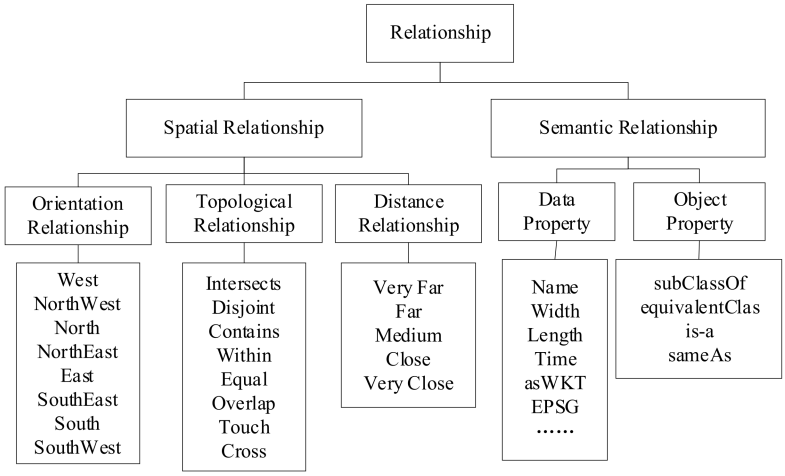

Figure 6. Relationships in the schema layer.

1. Topological Relationship

The topological relationship is invariant under topological transformations, including rotation, scale adjustments, and translation [9]. It is inherited from *gkg:spatialRelation*, thus representing the proximity between geographical entities. As shown in Figure 7, the topological relationships between entities include intersect, disjoint, contain, within, equal, overlap, touch, and cross. In these relationships, disjoint, touch, intersect, and equal are symmetric, while equal, contain, and within are transitive. Contain and within and disjoint and intersect are inverse. Taking "A contains B" as an example, B is entirely inside of A, and neither the interior nor the boundary of B intersects A's exterior.

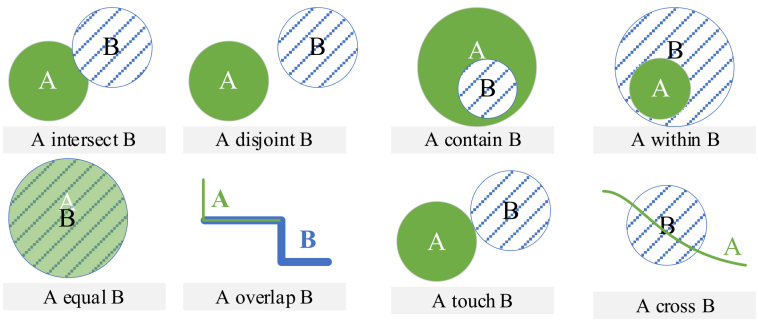

Figure 7. Topological relationships.

2. Distance Relationship

The distance relationship is defined as the minimum distance between two entities, and it is also inherited from *gkg:spatialRelation*. Both qualitative and quantitative distances are used in GeoKG. The quantitative distance is expressed by a data property with a precise value. Additionally, qualitative distance is divided into inner-city and inter-city, and these can be converted through thresholds. At the inter-city scale, the minimum speed of a high-speed train (i.e., 250 km/h in China) is used to calculate thresholds. As shown in Figure 8, running times of 20 min (about 25 km), 1 h (about 250 km), 2 h (about 500 km), 5 h (about 1200 km), and more than 5 h are qualitatively described as very close, close, medium, far and very far, respectively. At the inner-city scale, the distances of 3 km, 8 km, 15 km, and over 15 km are qualitatively considered very close, close, medium, and far, respectively. For example, the distance between Zhengzhou Railway Station and Zhengzhou East Railway Station is 11 km, qualitatively described as medium. Zhengzhou is about 130 km away from Luoyang, and the distance relationship is expressed as close.

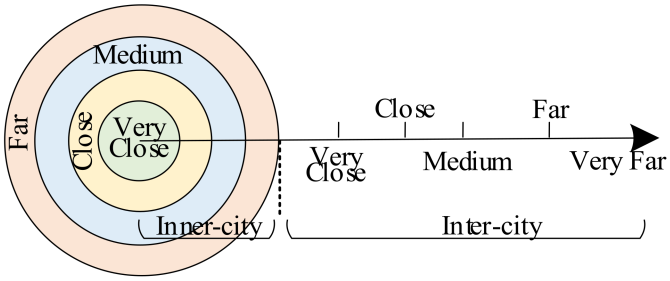

Figure 8. Distance relationships.

3. Orientation Relationship

The orientation relationship, inherited from *gkg:spatialRelation*, is another crucial spatial relationship in GeoKG, including northwest (NW), north (N), northeast (NE), west (W), east (E), southwest (SW), south (S), and southeast (SE).

3.4. Data Layer Construction

Geographical knowledge extraction and integration are used to construct a data layer to represent spatial location and morphological characteristics in GeoKG.

3.4.1. Knowledge Extraction

1. Concept Extraction from Geographic Dataset

Generally, concepts are mainly extracted from geographic datasets to complete the schema layer. Concepts of ground object categories (such as *Expressway* and *TransportationWarehousing*) are created and connected to the schema layer based on layers in the geographic datasets. Then, in each layer, the attribute fields of geographic datasets (such as "Kind") are used to extract the subclass concepts of ground objects. For example, as shown in Figure 9, the field Kind is used to create concepts *RailwayStation* and *BusStation*, which belong to the concept *TransportationWarehousing*. Then, these relationships can be represented in triples—<*TransportationWarehousing*, *is-a*, *GeoDatasetEntity*>, <*BusStation*, *is-a*, *Transportation Warehousing*>, and <*RailwayStation*, *is-a*, *TransportationWarehousing*>.

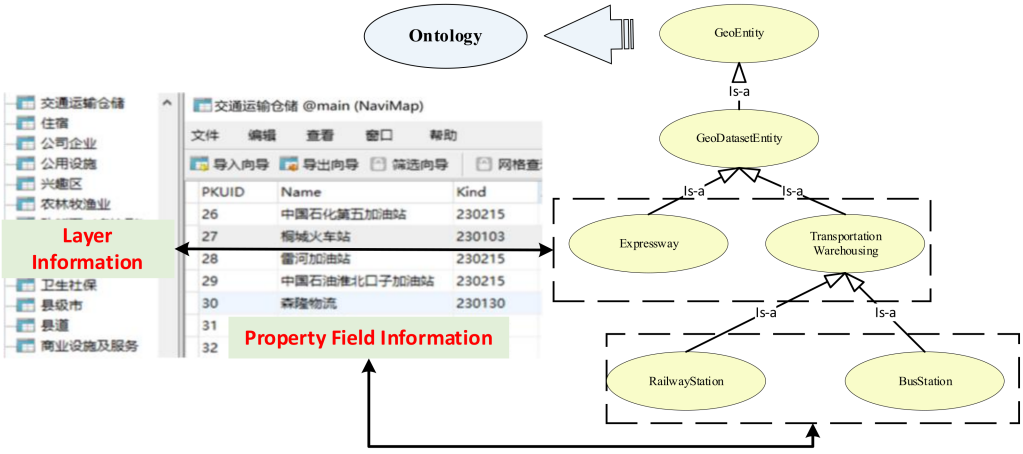

Figure 9. The relationships extracted from geographic datasets.

2. Entity Extraction from Geographic Dataset

Aiming at dividing geographical entities into multiple features and geometries, geographical entity extraction rules are designed based on layers and attribute fields. The spatial database includes a list of tables, each of which contains many rows (i.e., geographic features). These rows are composed of fields; property fields express attributes, and geometry fields represent spatial location.

The technical challenge of entity extraction lies in combining geometries. When an entity is only composed of a point, it can be presented in the WKT format of POINT, whose basic unit is a pair of longitude and latitude. The entity geometry format will be POLYGON if the points form a closed-loop containing a list of points; otherwise, it will be LINESTRING. Furthermore, MULTIPOINT, MULTILINE, and MULTIPOLYGON are used to construct entities whose basic units are POINT, LINE, and POLYGON. When the entity is composed of multiple geometry types, its geometric format must be a collection of geometric types.

When combining geometries from layers, the correspondence between entity names and property fields is used to form pairs. As shown in Figure 10, the correspondence between the layer Expressway and its attribute field ID is used to build pairs, creating a triple <Name, has, Layer-Feature IDs>. Then, the fields Geometry, ID, and Name are connected to concepts *GeoDatasetEntity*, *Feature*, and *Geometry*, while other attributes are mapped to data properties. Finally, coordinates and spatial reference systems are also transformed to WKT and EPSG code in the data layer. Besides, entity name and its administrative region are used to identify the same name entities in different areas, thus distinguishing the same name entity in the data layer.

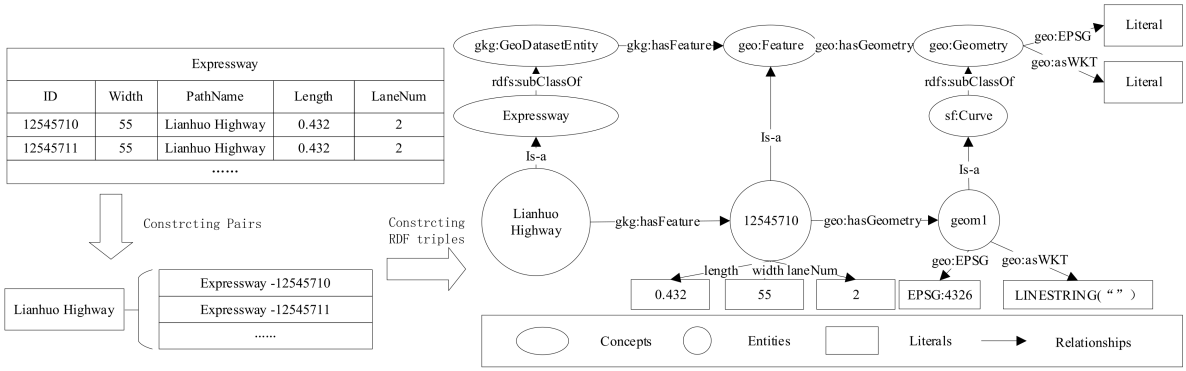

Figure 10. Knowledge extraction from geographic datasets.

3. Knowledge Extraction from Encyclopedias

For geographical entities, spatial information and semantics are the main areas of concern. Knowledge in Baidu Baike is extracted by opening an encyclopedia entry based on an entity name and then locating elements using XPath. As shown in Table 1, the rules are designed for extracting titles, synonyms, information boxes, and overview pictures, thus completing entity semantics.

Table 1. Extraction rules for Baidu Baike.

Information Type	XPath
Overview pictures	div.summary-pic > a > img
Synonym	span > span.viewTip-fromTitle
Entry title	dd.lemmaWgt-lemmaTitle-title > h1
Information box	dl > dt.basicInfo-item.name dl > dd.basicInfo-item.value

The "attribute–value" pairs (including attribute name and value) in the information box are extracted. For an attribute value existing in the extracted entities, an object property is built from the attribute name. Additionally, if the attribute value does not exist in extracted entities, the data property is designed to describe entity semantics. In addition to completing the knowledge in the data layer, the schema layer is also completed based on concepts and relationships extracted from Baidu Baike. As shown in Figure 11, *Zhengzhou* is extracted as an entity, the attribute value *7446 km*^{2} is represented as the literal, and the

attribute name *Area* is built as a data property. Additionally, *Zhengzhou East Railway Station* is built as an entity.

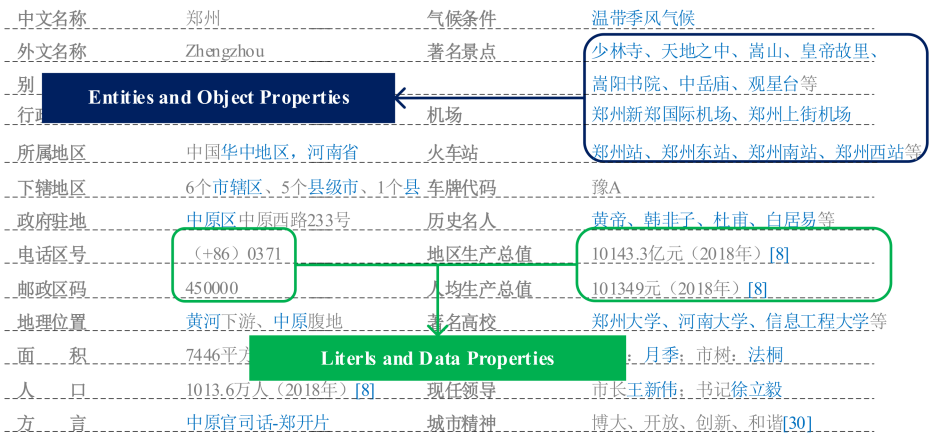

Figure 11. Semantics extraction from Baidu Baike.

3.4.2. Knowledge Integration

From the above steps, geographical knowledge is extracted from multisource data (i.e., Baidu Baike and geographic datasets). Thus, it is necessary to integrate this knowledge in two ways: knowledge linking and knowledge fusion.

Knowledge linking aims to discover equivalence relationships between entities. The Baidu Baike website address is built with the entity name, and a specified web page about the entity can thus be acquired. Then, the relationship *sameAs* is created to link these entities. As shown in Figure 12, the entities are extracted from geographic datasets and Baidu Baike, and they are deposited into concepts *gkg:GeoBaikeEntity* and *gkg:GeoDatasetEntity*, respectively. Then, the relationship *sameAs* is built to represent the equivalent property.

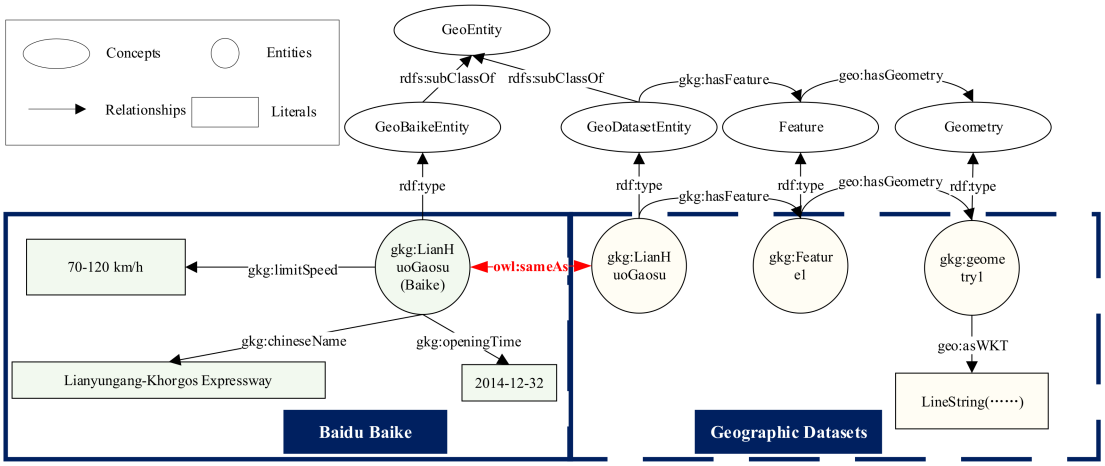

Figure 12. Linking geographical entities with the same name.

Differently from knowledge linking, geographical knowledge is fused in terms of attribute fields and values. In terms of attribute fields, fields with the same meaning but different names are unified based on statistics. As shown in Figure 13, the attribute fields *Line Length* and *Mileage* are both used to describe the entity length for line geometry, and *Line Length* is used as a final relationship in the RDF triple (i.e., <Lianluo Highway, *Line Length*, "4395 km">). In addition to attribute fields, the knowledge extracted from geographic datasets is considered the attribute value used to replace knowledge extracted from Baidu Baike, because of itss accurate geographical coordinates. Although knowledge

fusion strategies are a bit simple, they comprise an approach to acquire more accurate geographical knowledge.

Figure 13. Different expressions in Baidu Baike information box.

3.4.3. Knowledge Storage

Knowledge storage involves saving the acquired knowledge. In a relational database, storing RDF triples is redundant, and the JOIN operations demand more time. Similarly, graph databases cannot support the spatial index and real-time extraction of spatial relationships. Therefore, a single database cannot meet the actual needs. Spatialite is a spatial database with many advantages, such as small size, fast storage, high retrieval speed, and low cost. It is used to store geographical data and some structured semantics. As shown in Table 2, there are four tables in our database. The table GeoEntit stores geographical entities, including generated ID, name, added time, the collection of feature IDs and its corresponding layer name, Baidu Baike information, and geometry WKT. The table GeoField_Baike stores statistics about attribute fields extracted from Baidu Baike, including ID, name, frequency, and geographical IDs. The table GeoRelation stores relationships, including relationship ID, name, and frequency. The table re_Geo_Geo stores the RDF triples extracted from geographic datasets, including the triple ID, geographic entity IDs, and relationship ID. To combine knowledge acquired from geographic datasets and Baidu Baike, a graph database (Neo4j) is used to store relationships. In detail, the nodes store concepts, entities, and attribute values, while edges represent relationships.

Table 2. The attributes of tables in a relational database.

Table Name	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
GeoEntity	Geo_ID	Name	Time	Feature_IDs [1]	WebInfo	WKT
GeoField_Baike	BaikeField_ID	Name	Frequency	Geo_IDs	—-	—-
GeoRelation	Relation_ID	Name	Frequency	—-	—-	—-
ge_Geo_Geo	RDF_ID	Geo_ID	Relation_ID	Geo_ID	—-	—-

[1] Field "Feature_IDs" stores sets with feature IDs and their layer names.

4. Experiments and Evaluation

Transportation, warehousing, roads and administration (http://navinfo.com/digitalmap, accessed on 21 September 2021) are used as experimental data sources. Linkage rate, semantic relationship rate, and application cases are demonstrated to exploit the constructed GeoKG.

4.1. Linkage Rate

As shown in Table 3, the constructed GeoKG contains over 126,000 entities. There are three entity types: point-type, line-type, and polygon-type entities. More than 15,000 geographical entities are linked to Baidu Baike pages, accounting for 12.17% (i.e., over 15,000 *sameAs* relationships are constructed in the data layer). Additionally, there is an interesting phenomenon whereby the linkage rate varies significantly between geometry types. Polygon-

type entities have a high linkage rate of 100%, while the linkage rates of point-type and line-type entities are as low as 10.22% and 24.23%, respectively.

Table 3. Scale and linkage rate of GeoKG.

	Bridge	Bus Station	Airport	Railway Station	Expressway Service	Railway	Nation Road	Expressway	Province	City	Overall
Features	180,168	12,250	579	9538	7115	549	797,929	623,285	32	371	1631,816
Entities	87,481	11,881	378	9050	2321	548	10,213	4365	32	371	126,640
Linkage	3306	717	208	7006	113	347	2720	598	32	371	15,418
Linkage Rate [1]/%	3.78	6.03	55.03	77.41	4.87	63.32	26.63	13.70	100.00	100.00	**12.17**
Point-Type Linkage Rate/%	10.22			Line-Type Linkage Rate/%		24.23		Polygon-Type Linkage Rate/%		100	

[1] Linkage rate is the number of geographical entities between geographic datasets and Baidu Baike linked to the number of entities extracted from geographic datasets.

4.2. Semantic Relationship Rate

The geographical entities extracted from geographic datasets are full of exact coordinates, represented by EPSG and WKT. Moreover, the semantics are enriched because of their linkage with Baidu Baike. As shown in Table 4, the semantic relationship rate of the Chinese name is 100% because of the opening of Baidu Baike pages based on entity names. Although the geographical location rate reaches more than 88%, there are only 215 entities with precise geographic coordinates, accounting for 1.39%. The missed coordinates can be completed using the geographic datasets, thus increasing the rate to 100%.

Table 4. Statistics of semantic relationship rate.

Order	Relationship Type	Count	Coverage Rate/%
1	Chinese name	15,418	**100.00**
2	Geographical location	13,633	**88.42**
3	Foreign name	6998	45.39
4	Station level	4212	27.32
5	Regional management	4046	26.24
6	Main route	2947	19.11
7	Start date of construction	2931	19.01
8	Date of coming into service	2192	14.22
9	Postal code	1964	12.74
… …	… …	… …	… …
37	Geographic coordinates	**215**	**1.39**

4.3. Application Cases Based on GeoKG

Figure 14 demonstrates the retrieval process based on GeoKG. The first step is to click on the map, thus identifying the nearest entity on the map. Then, entities are retrieved via their semantic and spatial relationships in the databases. Finally, information about these entities can be shown on the map or in a graph. The application cases based on GeoKGs are as follows.

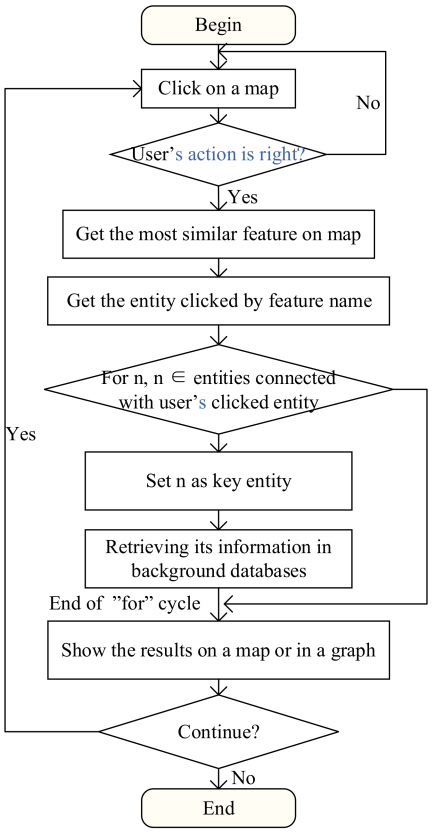

Figure 14. Retrieval process.

4.3.1. Processing One Layer

By processing the railway layer in geographic datasets, semantics and exact geographic coordinates can be obtained. Taking Longhai Railway as an example, the knowledge card will be shown on the right side, containing its overview, pictures and semantics (Figure 15). At the same time, the spatial information is displayed on the map.

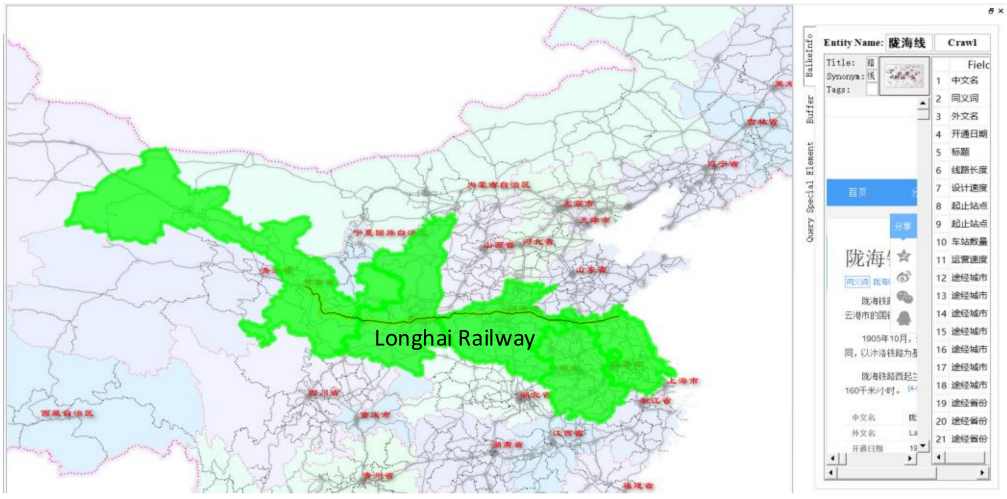

Figure 15. An application case after processing the administration layer.

4.3.2. Processing Multiple Layers

In addition to knowledge about the clicked-on entity, information about the past administration can be obtained after processing the polygonal province layer in geographic

datasets. As shown in Figure 15, the entity Longhai Railway (i.e., the black parts) and its past areas (i.e., the green parts) are represented on the map.

The relationship between point and line is hard to judge directly because of the deviation between point and line. Hence, the point–line relationship is acquired by GeoKG. In Figure 16, the railway stations in Longhai Railway are shown on the left side, and the details of Zhengzhou are represented on the right side (including a detailed map and a graph).

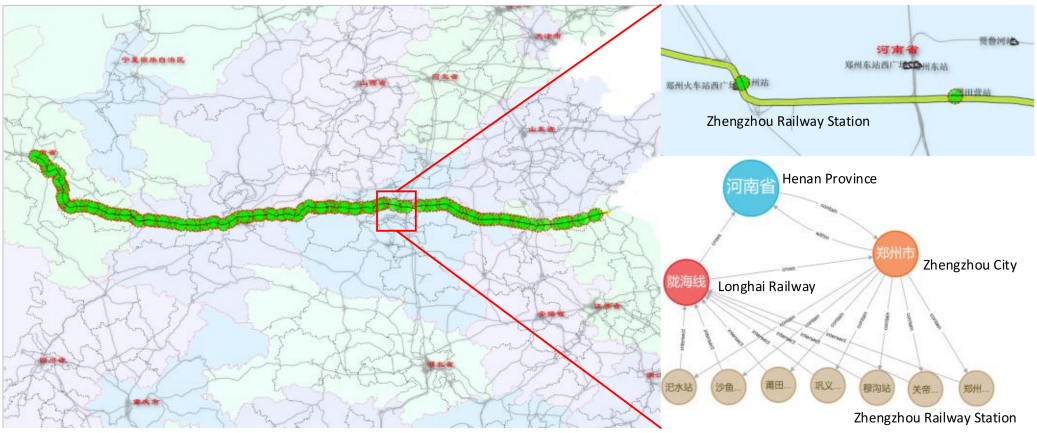

Figure 16. An application case after processing the transportation warehousing layer.

5. Discussion

The advantages and limitations of GeoKG construction are the main focuses of this study. The GeoKG integrates semantic characteristics with spatial characteristics to understand the real world.

The GeoKG is compared with two other KGs: CrowdGeoKG [2] and CKG [4]. Against the OBOR background, CKGs focus on extracting geographical entities from internet encyclopedias about the countries along OBOR. However, CKG does not consider geographic datasets, and lacks precise coordinates. Although CrowdGeoKG integrates knowledge from OSM and Wikidata, it lacks the support of extracting entities from geographic datasets (such as shapefile), and its linkage rate between OSM and Wikidata is only 6.62%. Compared to the above two methods, our GeoKG regards geographic datasets as a main data source and Baidu Baike as an assistant data source, whose linkage rate is increased to 12.17%. It also offers two more advantages. Firstly, the map and KG are integrated to simplify the GIS interactions. Secondly, a spatial database and a graph database are used to better support multi-source heterogeneous data fusion.

There are some design trade-offs of GeoKG. Firstly, geospatial cognition has the characteristics of levels and regions [32]. However, most of the spatial relationships extracted from two-dimensional space are erroneous, because they are in different levels or regions. In light of this, spatial relationships are constructed in the schema layer, and then extracted in real-time. To compensate for extraction time, the spatial database is introduced to improve efficiency. Although our method can acquire abundant geographical knowledge, it cannot extract knowledge from raster and trajectory data. Aiming at completing the GeoKG with more data sources, deep learning and image processing technology will be introduced to extract knowledge from these data.

6. Conclusions

KGs have attracted a lot of attention worldwide, and play an essential role in AI. However, general KGs lack geographical knowledge. In this paper, both geographic datasets and Baidu Baike are taken as data sources to extract geographical knowledge and semantics. In the schema layer, concepts and relationships are modeled to represent geographical knowledge based on GeoSPARQ. In the data layer, geographical knowledge

is extracted, interlinked, and transformed into RDF triples. Then, both graph and spatial databases are used to store geographical knowledge. Furthermore, the GeoKG is verified through the linkage rate, coverage rate, and application cases. The results indicate that the method could automatically extract knowledge from multisource data and combine accurate spatial location with semantics. Additionally, our GeoKG has a higher success rate of linking web pages with geographic datasets, and the accuracy of its coordinates has increased to 100%.

In a word, GeoKGs have become a new research hotspot, which can integrate multi-source geospatial data and promote GIS to realize the combination of accurate spatial location and semantics. Thus, they are of great significance to the extension of geographic data into knowledge.

Author Contributions: Conceptualization, X.G.; methodology, X.G. and J.L.; software, J.L. and F.W.; resources, H.Q.; writing—original draft preparation, X.G. and J.L.; writing—review and editing, F.W. and H.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China, grant number 41571442, and the Excellent Youth Foundation of Henan Scientific Committee, grant number 212300410014.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data of the current study are available from the corresponding author based on a reasonable request.

Acknowledgments: We would like to thank the anonymous reviewers for their insightful comments and substantial help on improving this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, W.; Geo, A.I. Where machine learning and big data converge in GIScience. *J. Spat. Inf. Sci.* **2020**, *20*, 71–77. [CrossRef]
2. Chen, J.; Deng, S.; Chen, H. CrowdGeoKG: Crowdsourced Geo-Knowledge Graph. In Proceedings of the China Conference on Knowledge Graph and Semantic Computing, Chengdu, China, 26–29 August 2017; pp. 165–172. [CrossRef]
3. Zhu, Y.; Zhou, W.; Xu, Y.; Liu, J.; Tan, Y. Intelligent Learning for Knowledge Graph towards Geological Data. *Sci. Program.* **2017**, *2017*, 1–13. [CrossRef]
4. Wu, T.; Qi, G.; Li, C.; Wang, M. A survey of techniques for constructing chinese knowledge graphs and their applications. *Sustainability* **2018**, *10*, 3245. [CrossRef]
5. Jiang, B.; Tan, L.; Ren, Y.; Li, F. Intelligent Interaction with Virtual Geographical Environments Based on Geographic Knowledge Graph. *Int. J. Geo Inf.* **2019**, *8*, 428. [CrossRef]
6. Li, W.; Zhu, J.; Zhang, Y.; Cao, Y.-G.; Hu, Y.; Fu, L.; Huang, P.; Xie, Y.; Yin, L.; Xu, B. A Fusion Visualization Method for Disaster Information Based on Self-Explanatory Symbols and Photorealistic Scene Cooperation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 104. [CrossRef]
7. Huang, W.; Mansourian, A.; Abdolmajidi, E.; Xu, H.; Harrie, L. Synchronising geometric representations for map mashups using relative positioning and Linked Data. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1117–1137. [CrossRef]
8. Mai, G.; Janowicz, K.; Yan, B.; Scheider, S. Deeply integrating Linked Data with Geographic Information Systems. *Trans. Gis* **2019**, *23*, 579–600. [CrossRef]
9. Palacio, M.P.; Sol, D.; Gonzalez, J.A. Graph-based knowledge representation for GIS data. In Proceedings of the Mexican International Conference On Computer Science, Tlaxcala, Mexico, 8–12 September 2003; pp. 117–124.
10. Krisnadhi, A.; Hu, Y.; Janowicz, K.; Hitzler, P.; Arko, R.A.; Carbotte, S.M.; Chandler, C.L.; Cheatham, M.; Fils, D.; Finin, T. The GeoLink Framework for Pattern-based Linked Data Integration. In Proceedings of the International Semantic Web Conference, Bethlehem, PA, USA, 11–15 October 2015.
11. Xu, B.; Xu, Y.; Liang, J.; Xie, C.; Liang, B.; Cui, W.; Xiao, Y. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System. In Proceedings of the International Conference Industrial, Engineering & Other Applications Applied Intelligent Systems, Arras, France, 27–30 June 2017; pp. 428–438.
12. Hook, P.A. Domain Maps: Purposes, History, Parallels with Cartography, and Applications. In Proceedings of the IEEE International Conference on Information Visualization, Zurich, Switzerland, 4–6 July 2007; pp. 442–446.
13. Zeyua, L. Review of the 30-year Studies of the Methodology of Science and Technology in China—Based on the Bibliometric Analysis of Journal Articles. *Stud. Philos. Sci. Technol.* **2014**, *31*, 82–89.

14. Zheng, K.; Xie, M.H.; Zhang, J.B.; Xie, J.; Xia, S.H. A knowledge representation model based on the geographic spatiotemporal process. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 1–18. [CrossRef]
15. Kacprzyk, J.; Belyakov, S.; Bozhenyuk, A.; Rozenberg, I. Knowledge Representations for Constructing Chains of Contexts in Geographic Information Systems. *Int. J. Comput. Intell. Syst.* **2021**, *14*, 1388–1395. [CrossRef]
16. Mekni, M. Using GIS Data to Build Informed Virtual Geographic Environments (IVGE). *J. Geogr. Inf. Syst.* **2013**, *5*, 548–558. [CrossRef]
17. Laurini, R.; Favetta, F. About External Geographic Information and Knowledge in Smart Cities. In Proceedings of the 2nd International Conference on Smart Data and Smart Cities, Puebla, Mexico, 4–6 October 2017.
18. Elkin, P.L.; Brown, S.H. Knowledge Representation and the Logical Basis of Ontology. In *Terminology and Terminological Systems*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 11–50.
19. Schade, S.; Smits, P.C. Why linked data should not lead to next generation SDI. In Proceedings of the International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 2894–2897.
20. Huang, W.; Kazemzadeh, K.; Manourian, A.; Harrie, L. Towards Knowledge-Based Geospatial Data Integration and Visualization: A Case of Visualizing Urban Bicycling Suitability. *IEEE Access* **2020**, *8*, 85473–85489. [CrossRef]
21. Varanka, D.E.; Usery, E.L. The map as knowledge base. *Int. J. Cartogr.* **2018**, *4*, 201–223. [CrossRef]
22. Janowicz, K.; Schade, S.; BröRing, A.; KeßLer, C.; Maué, P.; Stasch, C. Semantic Enablement for Spatial Data Infrastructures. *Trans. Gis* **2010**, *14*, 111–129. [CrossRef]
23. Hofer, B.; Mäs, S.; Brauner, J.; Bernard, L. Towards a knowledge base to support geoprocessing workflow development. *Int. J. Geogr. Inf. Syst.* **2017**, *31*, 694–716. [CrossRef]
24. Gould, N.; Mackaness, W. From taxonomies to ontologies: Formalizing generalization knowledge for on-demand mapping. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 208–222. [CrossRef]
25. Almeida, P.D.; Rocha, J.; Ballatore, A.; Zipf, A. Where the Streets Have Known Names. In Proceedings of the International Conference on Computational Science and Its Applications, Beijing, China, 4–7 July 2016; pp. 1–12.
26. Codescu, M.; Horsinka, G.; Kutz, O.; Mossakowski, T.; Rau, R. OSMonto—An Ontology of OpenStreetMap Tags. *State Map Eur. (SOTM-EU)* **2014**, *2011*, 23–24.
27. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowl. Inf. Syst.* **2013**, *37*, 61–81. [CrossRef]
28. Hoffart, J.; Suchanek, F.M.; Berberich, K.; Weikum, G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.* **2013**, *194*, 28–61. [CrossRef]
29. Liu, J.; Liu, H.; Chen, X.; Guo, X.; Zhao, Q.; Li, J.; Kang, L.; Liu, J. A Heterogeneous Geospatial Data Retrieval Method Using Knowledge Graph. *Sustainability* **2021**, *13*, 2005. [CrossRef]
30. Auer, S.; Lehmann, J.; Hellmann, S. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In Proceedings of the International Semantic Web Conference, Chantilly, VA, USA, 25–29 October 2009; pp. 731–746.
31. Hu, Y.; Janowicz, K.; Carral, D.; Simon, S.; Kuhn, W.; Berg-Cross, G.; Hitzler, P.; Dean, M.; Kolas, D. A Geo-Ontology Design Pattern for Semantic Trajectories. In Proceedings of the International Conference on Spatial Information Theory, Scarborough, UK, 2–6 September 2013; Springer: Cham, Switzerland, 2013; pp. 438–456. [CrossRef]
32. Yong, G. Representation and Reasoning of Spatial Relations in Geographical Space. *Geogr. Geo-Inf. Sci.* **2007**, *23*, 1–6.