Construction and Application of Place Name and Address Management System Based on Elasticsearch

Hongping Zhang [1,2,3], Wei Huang [2,3], Lei Ding [2,3], Dejin Tang [2,3], Cong Wang [2,3], Yong Xu [2,3], Zhen Wang [2]

[1] School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, China
[2] National Geomatics Center of China, Beijing, China- (zhanghongping, huangwei, dinglei, tengdejin, wangcong, xuyong, wangzhen) @ngcc.cn
[3] Technology Innovation Center for Geographic Information Public Service, Ministry of Natural Resources, Beijing, China

Keywords: Place name, Address, Elasticsearch, Multi-level address model, Data management, Digital government

Abstract

Place names and addresses serve as essential references for locating information about people's living, production, and various activity sites. They are a crucial component of public geographic framework data and act as tools and bridges for social communication in daily production and life. This paper presents a comprehensive system for managing place names and addresses by integrating a multi-level address model, data synchronization between PostgreSQL and Elasticsearch, and advanced query capabilities to ensure efficient and accurate data management. The system implements full lifecycle management of place names and addresses, fully leveraging the advantages of relational databases and distributed search engines. It provides essential functions such as address standardization, real-time updates, maintenance management, conditional queries, and statistical analysis. It has been successfully implemented in the National platform for common geospatial information service, laying the foundation for the next steps in refining social governance based on place names and addresses, and further advancing the digital transformation of government operations.

1. Introduction

Place names serve as the fundamental basis for spatial orientation in people's daily lives, while addresses provide textual descriptions of place name information access and geographical location (Chen et al., 2016; Sebake and Coetzee, 2013). As an integral component of geographic information public service data resources, place name and address information play a pivotal role in facilitating unified and effective spatial positioning. Their significance extends beyond their close association with people's daily lives, serving as a crucial link in the information infrastructure of national informationization strategies (Tal et al., 2022). They constitute essential information resources for governmental administrative management and economic development, without which these endeavors would be incomplete. With the rapid development of China's social economy and the accelerating process of urbanization, the volume of place name and address data has shown explosive growth and continuous updates. It is known that the place name and address data used online by National platform for common geospatial information service has exceeded tens of millions entries, with millions of entries updated annually. In addition to the vast amount of data, place name and address information involves dozens of major categories, nearly a hundred subcategories, and hundreds of minor categories. Different place name and address data also possess various industry attributes and labels. The massive amount of place name and address data, along with their complex classifications, presents significant challenges in terms of data storage, management, querying, statistics, updates, distribution, and etc.

At present, place name and address data primarily rely on database development and management systems (Liu et al., 2018). The existing technological solutions heavily rely on relational databases for data management and statistical analysis. Querying and statistical analysis primarily utilize database query languages (Cooper et al., 2020; Tredrea et al., 2020). However, when dealing with millions or more data entries, especially in situations involving fuzzy queries or complex conditional combinations, solutions based on database query languages have shortcomings in areas such as full-text search, query efficiency, and support for unstructured data. Consequently, this leads to issues such as complex management and low efficiency (Chen et al., 2021).

Database synchronization technology ensures data consistency across multiple databases by replicating and backing up data changes from the source database to target databases in real-time or at scheduled intervals. This technique has been widely applied in the fields of data migration and big data analysis (Liu et al., 2017; Zhang et al., 2021), establishing itself as a critical component of modern data management. In this paper, we have developed a place name and address management method and system based on ElasticSearch to meet the demands of managing massive geographic information data and conducting complex queries and statistical analyses (Gormley and Tong, 2015). First, the architecture of place name and address is proposed. Second, the key technologies including the multi-level address model and data synchronization are introduced. Third, we provide a detailed explanation from the overall architecture design to the construction of place name and address databases, and the specific functionalities of the management system. Finally, we close with some concluding remarks.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-571-2024

2. Architecture

Aiming to make massive place names and addresses management with high efficiency and standardization, this paper proposed a framework based on cloud-native technologies for data collection, process, storage and service. As Figure 1 shows, the system consists of four components: the data layer, storage layer, service layer and application layer.

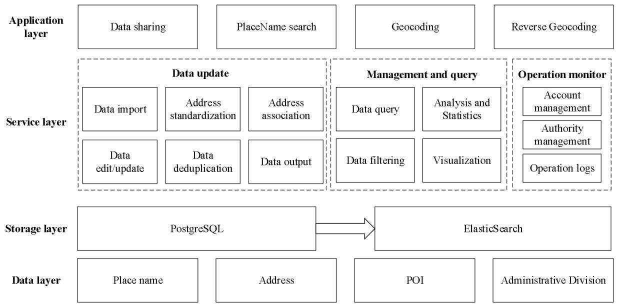


Figure 1. System architecture

Data layer: All the place name and address data are represented by coordinate location points, including information such as names, precise coordinates, addresses, categorical classifications, descriptive tags, and administrative division codes. Notably, these data originate from a diverse array of sources, spanning governmental agencies, enterprises, and even contributions from volunteers, resulting in varying degrees of data quality.

Storage layer: The system used the relational database PostgreSQL to store all the data with the plugin of PostGIS. In modern data architectures, the integration of database and search engine has become increasingly important. To provide faster and more accurate search services, we have implemented real-time data synchronization between PostgreSQL databases and ElasticSearch. This measure ensures the immediate updating of data and the accuracy of search results, providing users with a better experience.

Service layer: According to the logical flow of place name and address data in the system, the service layer implements comprehensive models for data update, management, and operational monitoring. These models encompass detailed functionalities such as data importation, address standardization, data editing, real-time updating, query capabilities, data visualization, account administration, and the management of operational logs, among others.

Application layer: The application layer implements multiple interface-level functional services, including data sharing, place name search, geocoding and reverse geocoding.

3. Method

3.1 Multi-level Address Model

The address model aims to transform human descriptions of spatial locations in both production settings and daily life into granular geographic entities. This refinement process aims to achieve standardized descriptions of locations. As shown in table 1, our multi-level address model incorporates 16 levels, which undergo a rigorous semantic parsing process. The parsed data is then systematically organized and stored in dedicated fields, allowing for efficient retrieval and utilization in various spatial applications. This approach significantly enhances the accuracy and reliability of spatial positioning, thus paving the way for advanced spatial analysis and decision-making.

Level	Field
1	Province/Autonomous Region/Municipality
2	Prefecture-level City/Prefecture
3	District/County
4	Street/Township
5	Community/Village
6	Street/Road/Lane
7	Exclusive District/Development Zone
8	Commercial District
9	Residential Area/Courtyard/Village
10	Point of Interest(POI)
11	Street/Road/Lane Number
12	Building Number/Compound Number
13	Unit
14	Floor
15	Room
16	Intersection/Entrance

Table 1. Multi- level address model

Building upon the multi-level address model, we enrich standard addresses with semantic information, rendering them more vivid and illustrative in describing locations within the real world. This augmentation ensures that the addresses encapsulate not only spatial coordinates but also contextual nuances and interpretable details, thereby fostering a more intuitive comprehension of the respective locations.

Table 2 shows a standard place name and address example. The address is 28 Lianhuachi West Road, Haidian District, Beijing, and the name of POI is China Surveying and Mapping Building. In daily life, people also use alternative names or nicknames to refer to the same location. Table 3 displays the corresponding alternative names for the POI, such as China Surveying and Mapping, Surveying and Mapping building, China Surveying and Mapping Science and Technology Museum, etc. By leveraging the augmentation of semantic information pertaining to place names and addresses, we can substantially enhance the precision of address matching, thereby advancing the accuracy of geolocation services.

Field	Value
Province/Autonomous Region/Municipality	Beijing
Prefecture-level City/Prefecture	-
District/County	Haidian
Street/Township	Yangfangdian
Community/Village	-
Street/Road/Lane	Lianhuachi West Road
Exclusive District/Development Zone	-
Commercial District	-
Residential Area/Courtyard/Village	-
Point of Interest(POI)	China Surveying and Mapping Building
Street/Road/Lane Number	28
Building Number/Compound Number	-
Unit	-
Floor	-
Room	-

Intersection/Entrance	-

Table 2. A standard address example

ID	Semantic Address
1	China Surveying and Mapping
2	China Surveying and Mapping mansion
3	Surveying and Mapping building
4	National Bureau of Surveying, Mapping and Geoinformation
5	NASG
6	China Surveying and Mapping Science and Technology Museum

Table 3. Added semantic information for the standard address

3.2 Address Standardization

In the processing of input address data, the initial phase involves thorough cleansing to eliminate unnecessary characters, spaces, punctuation, and rectify conspicuous spelling errors. This preliminary refinement ensures the data's integrity and accuracy for further analysis. Semantic analysis mainly involves address segmentation techniques to accurately extract administrative divisions, streets (lanes), residential areas, natural villages, door (building) numbers, unit-room numbers from address texts. Subsequently, a segmentation process is implemented, leveraging the distinct characteristic words that constitute Chinese address elements. For example, an address such as "28 Lianhuachi West Road, Haidian District, Beijing" is dissected into its constituent parts: "Beijing City," "Haidian District," "Lianhuachi West Road," and "No. 28."

Following segmentation, the process incorporates the identification of the minimum-level administrative division associated with the address through the utilization of coordinate information. This verification step ensures that the administrative division data is accurate and conforms to established standards. A scoring mechanism is then applied to assess the degree of standardization achieved for each address, providing a quantitative metric for evaluating data quality.

Concurrently, a rigorous search is conducted to identify any potential duplicate address entries within the dataset. In the event of duplicate addresses that exhibit consistent or similar locations, the new address is not redundantly entered into the database. Instead, the place name information is intelligently associated with the existing address record, eliminating redundant data and maintaining database integrity.

However, in cases where duplicate addresses are geographically distant, a more nuanced approach is adopted. By referencing additional contextual information such as nearby roads, water systems, and prominent buildings, a determination is made regarding which address elements should be preserved to ensure the most accurate representation of the location.

Finally, the refined and segmented addresses are transformed and mapped into standardized fields, aligning with a predefined address standardization model. This standardization process ensures consistency and compatibility with downstream applications, ultimately culminating in the successful integration of the addresses into the database for further analysis and utilization.

Figure 2 shows the address standardization process

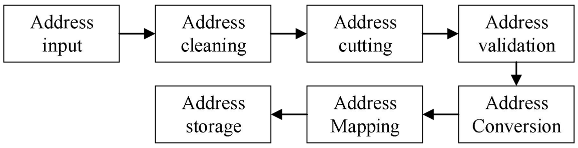


Figure 2. Address standardization process

3.3 Data Synchronization

In the realm of information technology, database synchronization techniques, particularly the synchronization of data between PostgreSQL and Elasticsearch, have become pivotal in leveraging advantages in data querying. PostgreSQL, as a relational database management system, is highly regarded for its transactional capabilities, consistency, and data integrity. However, when it comes to complex search and analytical tasks, Elasticsearch distinguishes itself with its full-text search capabilities, high performance, distributed architecture, flexible data modeling, real-time processing, scalability, multi-tenancy support, rich RESTful APIs, powerful data aggregation and analysis, as well as fault tolerance and self-healing capabilities. By synchronizing data from PostgreSQL to Elasticsearch, significant enhancements in query performance can be achieved, enabling rapid data retrieval and analysis while preserving the robust functionalities of the original database in transaction processing and data consistency. This technological integration not only optimizes data management processes but also provides a solid foundation for building efficient, scalable data-driven applications.

To enhance the query efficiency of massive place name and address data, we employ storage and computation-based separation techniques to manage the data. The original data is stored in PostgreSQL, and ElasticSearch is used to realize complex querying and filtering functionalities. Elasticsearch has the capability of millisecond-level response in retrieving geographical location information, and millisecond-level response is crucial to user experience. As shown in Figure 3, Apache Flink Change Data Capture (CDC) Connectors offer an efficient mechanism for facilitating real-time data synchronization from PostgreSQL databases to Elasticsearch and is used in this paper.

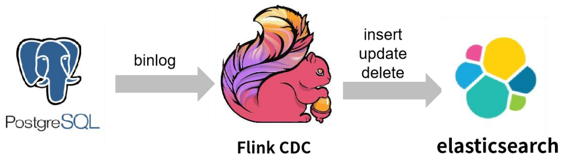


Figure 3. Data synchronization between PostgreSQL and ElasticSearch

3.3.1 Environment Preparation: Prior to the implementation of Flink CDC Connectors, it is essential to ensure that the Apache Flink cluster is operational. Concurrently, PostgreSQL databases and Elasticsearch clusters should be properly installed and configured to ensure version compatibility and support for CDC functionalities.

3.3.2 Flink Job Configuration: In a Flink job, it is initially necessary to define the PostgreSQLCdcSource. This connector serves as the starting point for data capture, establishing a connection to the PostgreSQL database and monitoring for data changes. Subsequently, the ElasticsearchSink must be configured as the endpoint for data synchronization, writing the captured data changes into Elasticsearch.

3.3.3 Data Capture and Transformation: The PostgreSQL CDC Source is responsible for capturing change events within the PostgreSQL database, encompassing insert, update, and delete operations. If adaptation to the Elasticsearch data model or data cleansing is necessary, transformation operations can be introduced within the data stream.

3.3.4 Data Writing into Elasticsearch: The ElasticsearchSink is tasked with receiving the data stream from the PostgreSQL CDC Source and mapping it to the document structure of Elasticsearch. During this process, it is essential to configure the connection properties of Elasticsearch, such as cluster addresses and index names.

3.3.5 Index Construction： Elasticsearch uses an inverted index to store data, with the indexed documents serialized in JSON format. The index data can be stored on a single server or sharded across multiple servers. Each index can have one or more shards, and each shard can have multiple replicas. To improve query retrieval efficiency, this paper constructed nine types of index information, including place names, address suggestions, landmarks, road names, river names, and building indices. The coordinate point uses the field type of "geo_point", while complex geometric objects like lines and polygons use the field type of "geo_shape". Additionally, the IK analyzer was used in "ik_max_word" mode to achieve effective Chinese word segmentation.

In response to the demands for data resource management and services for place names, addresses, administrative divisions, and other geospatial information on the National Geospatial Information Public Service Platform (Tianditu), this paper designs and develops a place name and address management system. This system is based on the study of technologies such as multi-level address models, distributed search engines, full-text search for place names and addresses, real-time database synchronization, and geospatial grids. The system achieves efficient management and application services for place names and addresses. It is primarily divided into four modules: statistical analysis, data updates, query services, and operation monitoring.

4. Implementation

4.1 Statistical Analysis

The system supports statistics based on region (national, provincial, municipal), time (yesterday, today, the last 7 days, custom periods), update type (e.g., added, edited, closed, released), and data category. The statistical information includes total data volume, total released data volume, recent updated volume, and recent released volume. The statistical results can be displayed using bar charts and data tables. Figure 4 shows the number of place names and addresses updates over a period of time.

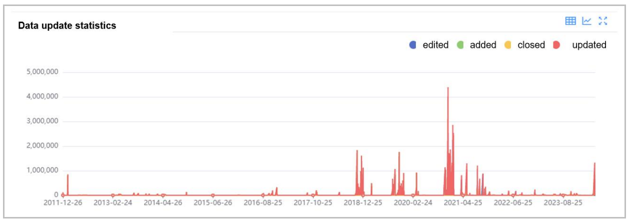
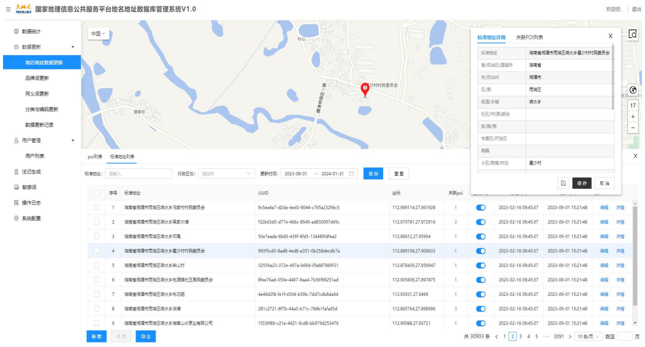


Figure 4. Data update statistics

4.2 Data Updates

Data updates primarily include place names and addresses, synonyms, brand terms, classification codes, and administrative divisions. Corresponding update operations are provided for different data themes.

4.2.1 Place name and Address Update： The system offers comprehensive management functions for updating place names and addresses. These functions include data import, entry, editing, standardization, association, extraction, query, and browsing. Data import supports various formats, such as TXT, CSV, and SHP. Users can manually add or modify place name and address information directly on a map. Once new addresses are entered, the system automatically performs standardization and validation. New addresses are added to the database, while existing addresses are automatically associated with the new entries. Additionally, the system supports batch editing of specific attributes, allowing for uniform modifications of classification codes, importance levels, and more. As Figure 5 shown, Users can select and edit any individual data entry. Figure 6 shows the details of a standard address and the related POIs .



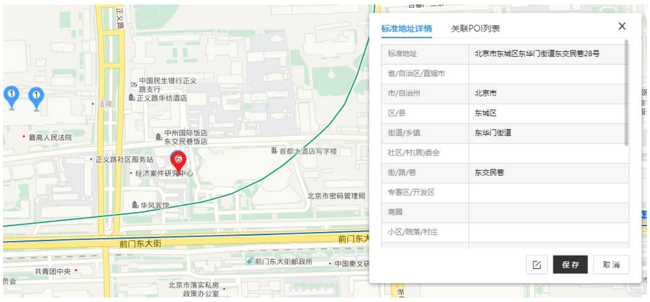
标准地址	湖南省湘潭市雨湖区响水乡星沙村村民委员会
省/自治区/直辖市	湖南省
市/自治州	湘潭市
区/县	雨湖区
街道/乡镇	响水乡
社区/村(居)委会	
街/路/巷	
专署区/开发区	
商圈	
小区/院落/村庄	星沙村

标准地址： 请输入 行政区划： 请选择 更新时间： 2023-08-01 → 2024-01-31

序号	标准地址	UUID	坐标	关联poi				
1	湖南省湘潭市雨湖区响水乡冯家村村民委员会	9c5ea4a7-d2da-4ed3-8044-c765a2329bc5	112.969114,27.961928	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
2	湖南省湘潭市雨湖区响水乡吴家大塘	f32b63d0-d77e-466c-8549-ad850097d45c	112.970781,27.972916	2	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
3	湖南省湘潭市雨湖区响水乡坝尾	50e7eada-6b69-439f-8fd5-134489fdfea2	112.86612,27.95994	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
4	湖南省湘潭市雨湖区响水乡星沙村村民委员会	993f5cd3-8ad8-4ed8-a351-0b258decdb7a	112.899106,27.900633	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
5	湖南省湘潭市雨湖区响水乡缺山村	32556a23-372e-497a-b69d-0fa687869f31	112.876439,27.959947	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
6	湖南省湘潭市雨湖区响水乡桃源路社区居民委员会	8fee76ad-059e-4497-9aa4-7b56f98251ad	112.905606,27.897875	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
7	湖南省湘潭市雨湖区响水乡桃花咀	4e4642f8-fe1f-4504-b59b-7dd7cdb8da64	112.93931,27.9499	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
8	湖南省湘潭市雨湖区响水乡洪塘	281c2721-8f7b-44a3-b71c-784b1fafa65d	112.865764,27.898996	3	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情
9	湖南省湘潭市雨湖区响水乡湖南山水泵业有限公司	1f239f89-c21e-4421-9cd8-bb979d253476	112.90088,27.93721	1	●	2023-02-16 09:45:37	2023-09-01 15:21:48	编辑 详情

共 30903 条 < 1 2 3 4 5 ··· 3091 > 10 条/页 跳至 页

Figure 5. Place name address update



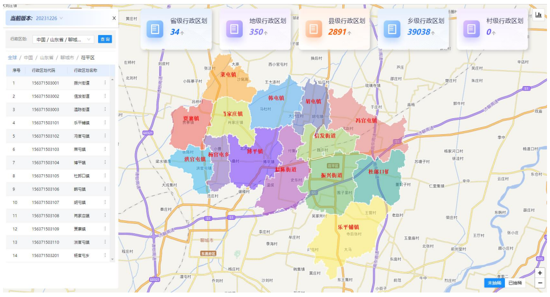
标准地址	北京市东城区东华门街道东交民巷28号
省/自治区/直辖市	
市/自治州	北京市
区/县	东城区
街道/乡镇	东华门街道
社区/村(居)委会	
街/路/巷	东交民巷
专署区/开发区	
商圈	
小区/院落/村庄	

Figure 6. A standard address details

4.2.2 Synonyms and Brand terms Update: Synonyms and brand terms are essential for improving the precision of fuzzy search matches. Synonyms refer to short names, aliases, or former names of places; for example, "Lize Business District" may simply be referred to as "Lize." Brand terms are specific keywords or phrases associated with POIs such as tourist attractions, restaurants, or shops. For instance, the POI information for a renowned shopping center might include brand names like "Nike," "Adidas," and "Apple." This feature supports operations for creating, editing, deleting, and associating synonyms and brand terms with POIs.

4.2.3 Classification Code Update: POI classification codes consist of three levels: major category, subcategory, and minor category. Major and subcategory codes are immutable, allowing only the creation, modification, or deletion of minor category codes. When classification codes are updated, any associated place names and addresses will be synchronized accordingly.

4.2.4 Administrative Division Update: The functions for importing, modifying attributes, version management, maintaining hierarchical relationships, browsing, querying, and visualizing spatial and attribute data of administrative divisions are provided. Administrative division data is utilized to assign administrative division codes to place name and address data. Administrative division data includes national, provincial, municipal, county, and township levels. When administrative divisions undergo updates, the associated place names and addresses within those regions are synchronized accordingly. Figure 7 shows the visualisation of the administrative division data.



序号	行政区划代码	行政区划名称
1	156371503001	振兴街道
2	156371503002	信发街道
3	156371503003	温陈街道
4	156371503101	乐平铺镇
5	156371503102	冯官屯镇
6	156371503103	菜屯镇
7	156371503104	博平镇
8	156371503105	杜郎口镇
9	156371503106	韩屯镇
10	156371503107	胡屯镇
11	156371503108	肖家庄镇
12	156371503109	贾寨镇
13	156371503110	洪官屯镇
14	156371503201	杨官屯乡

Figure 7. Administrative division visualisation

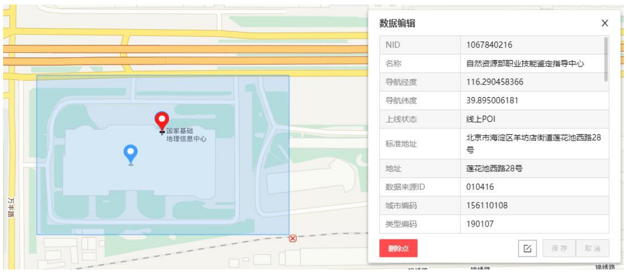
4.3 Place name and Address Query

The system offers simple query, conditional query, and polygon query for place names and addresses.

4.3.1 Simple Query: It supports filtering and querying data based on unique ID, name, classification code, administrative division code, and update time. Filtering methods include fuzzy search, exact matching, left matching, and right matching. Leveraging search engine capabilities, the system implements commonly used database query patterns, enhancing the user-friendly interaction of data querying.

4.3.2 Conditional Query: Queries can be constructed using keywords, classification codes, administrative divisions, update times, and other information to achieve precise searching. Custom sorting based on specific fields is also available.

4.3.3 Spatial Query: Spatial queries support polygon and buffer query. Polygon query refers to drawing an area on the map to retrieve place name and address data within that region. Buffer query refers to retrieving information within a certain distance from a central location. Figure 8 shows the result of polygon query.



NID	1067840216
名称	自然资源部职业技能鉴定指导中心
导航经度	116.290458366
导航纬度	39.895006181
上线状态	线上POI
标准地址	北京市海淀区羊坊店街道莲花池西路28号
地址	莲花池西路28号
数据来源ID	010416
城市编码	156110108
类型编码	190107

Figure 8. Polygon query

4.4 Operation Monitoring

Operations monitoring encompasses modules such as system parameter management, user and authority management, and log management.

4.4.1 System Parameter Management: System parameter management primarily oversees built-in parameters, interface control parameters, and data dictionaries. Built-in parameters include database connection information and data synchronization configuration parameters. Interface control parameters manage query return limits, data extraction template parameters, and buffer query radii.

4.4.2 User and Authority Management: The system comprises administrator, auditor, and updater users, establishing a user authentication mechanism. When users undergo authentication, the system allocates corresponding access permissions based on their level or type. Users with different permissions have access to different information resources, and the data accessible and operable within application systems also vary. Administrator users are primarily responsible for system parameter configuration, as well as adding or removing updater personnel. Auditor users are tasked with reviewing data update information but do not have permissions to update or maintain place name and address data. Updater users are responsible for the specific operations related to updating and maintaining place name and address data.

4.4.3 Log Management: The log management module primarily offers users the functionality of recording operation logs, which include user login/logout activities, data queries, updates, exports, and other operations. Administrators and auditors have the permission to query these logs.

5. Conclusion

The place name and address data management system is designed and developed in this paper fully integrates the respective advantages of spatial databases and the distributed full-text search engine Elasticsearch. It enables unified and efficient management of billions of place name and address data entries, enhancing the customization and flexibility of querying and analysis. It has been used in the National platform for common geospatial information service in China, which is government-led with nearly one minion developers and about 850 thousand application systems. This system addresses the low efficiency issues present in existing technological solutions, effectively simplifying the challenges associated with place name and address management and distribution. It meets the demands of managing and utilizing massive place name and address data in the era of big data, allowing for rapid updates and shared applications of continuously updated place name and address data.

References

Chen, D., Cheng, C., Tong, X., Yuan, J., 2016. Research on the multi-scale spatial location coding model for address. *Geoinformation Science* 18, 727–733.

Chen, Jian, Chen, Jianpeng, She, X., Mao, J., Chen, G., 2021. Deep contrast learning approach for address semantic matching. *Applied Sciences* 11, 7608.

Cooper, A.K., Katumba, S., Coetzee, S. 2020. South Africa needs a national database of addresses: How it could be done. *The Conversation*

Gormley, C., Tong, Z., 2015. Elasticsearch: the definitive guide: a distributed real-time search and analytics engine. O’Reilly Media, Inc.

Liu, Y., Liu S., 2017. Research and Implementation of Database Synchronization Technology. *SOFTWARE ENGINEERING* 20(01):1-4.

Liu, Y., Sun, W., Li, C., Liu, X., Fang, C., 2018. Design and Implementation of Dynamic Update System for Geographical Names and Addresses. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, 379–382.

Sebake, M.D., Coetzee, S.M., 2013. Address data sharing: Organizational motivators and barriers and their implications for the South African spatial data infrastructure. *Int. J. Spatial Data Infrastructures Res.* 8: 1-20.

Tal, Y., Keinan, E., Haj Yehia, B., Roi, Y., Berkowich, T., Ronen, H., 2022. Development and Dissemination of AN Israeli First Responders Addresses Geospatial Dataset for Management of Emergency Events. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43, 567–572.

Tredrea, G., Coetzee, S.M., Rautenbach, V., 2020. Cloud-based integration and standardization of address data for disaster management–a South African case study. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIV-3/W1-2020*, 145–150.

Zhang, J., Wang, R., Jiang, X., 2021. Research and Implementation of Synchronization Technology of Heterogeneous Database. *SOFTWARE ENGINEERING* 24(01):6-9.