Research on Multi-source Place Name Data Integration and Update Methods

Shangwei Lin [1], Xiao Du [1], Xi Guan [2], Meng Xu [3], Xueying Liang [4], Jiage Chen [1], Chenchen Wu [1], Xiaoguang Zhou[5]

[1] National Geomatics Center of China, 100830 Beijing, China - (linshangwei, duxiao, jiagechen, wucc)@ngcc.cn
[2] SinoMaps Press Group Co., Ltd., 100053 Beijing, China - woshi.guanxi@163.com
[3] Heilongjiang Institute of Geomatics Engineering,150081 Haerbin Heilongjiang, China – xumeng92@qq.com
[4] Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, 100048 Beijing, China – liangxueying1992@sina.com
[5] Central South University,410083 Changsha Hunan, China – zxgcsu@foxmail.com

Keywords: Multi source Place names, Matching and fusion, Levenshtein Distance, Similarity, Rule library, Transliteration.

Abstract

Place names are important basic geographic information resources for surveying and mapping. They play a significant role in maintaining national security, national defense construction, economic development, and other aspects. The construction and development of global geographic information resources require the continuous, efficient acquisition, and integration and updating of place names data. This paper takes the boundaries of the lowest administrative units in the task area as the matching scope and employs a string edit distance algorithm based on the substitution rules of common name synonyms for the similarity calculation and integration and updating of place names. It also uses a multilingual translation method based on a transliteration rule library for foreign language translation, with the addition of collaborative processing verification. This ultimately improves the matching efficiency of multi-source place names elements and ensures the accuracy of foreign place names translation. The research findings have been tested in some national regions and can provide a reference for the production and updating of place names data worldwide.

1. Introduction

Geomatics data is not only an important strategic data resource but also a new type of production factor, comparable in importance to traditional production factors such as land, labor, capital, and technology. Place names data, as a type of geomatics data, is an essential basic geographic information resource for surveying and mapping and a public information resource(Zhang & Shi, 2012). It plays a significant role in maintaining national security, national defense construction, economic development, and other areas. The updating of place names reflects the development of urban society and changes in geographic entities. With the increasing globalization of data cooperation and the rapid development of Internet technology, the demand for geographic names data from all sectors of society is growing. The construction and development of global geographic information resources require the continuous and efficient acquisition and integration of multi-source geographic names data.

Place names are unique designation assigned by people to a specific spatial location of a natural or cultural geographic entity (Zhang et al., 2017), it has characteristics such as sociality, temporality, ethnicity, locality, and representativeness. These names include the names of natural geographic entities, administrative divisions, streets and alleys, and geographic entities with significant geographical orientation meanings. They are generally composed of two parts: the specific name and the generic name. Transformation guidelines of geographical names from foreign languages into Chineses clearly stipulates that the specific name is the term used in a palce name to distinguish individual geographic entities, while the generic name is the term used to categorize the type of geographic entity.

Traditional methods of updating place name, such as field surveying, mobile mapping system updates, and periodic overall updates, face challenges including large manual data collection and processing workloads, long data acquisition and update cycles, high costs for updating and translation, and difficulties in conducting field surveys and on-site verification in foreign areas. With the development of crowdsourced mapping technology, the updating of place name data has gradually shifted from comprehensive, centralized, and periodic updates to incremental, crowdsourced, and dynamic updates. This shift also poses higher requirements for the efficiency and quality of data matching, integration, and updating(Zhang et al., 2017). There is a vast amount of open-source place names data available for free on the internet, such as GeoNames, Open Street Map (OSM), Global Administrative Areas Database (GADM), Open Addresses, and Geocode Earth（Dong et al., 2020; Li & Liu, 2024; Wei et al., 2016）. However, these multi-source data often overlap and are redundant, with some data having relatively low precision, weak currency, uneven density of geographic names, missing attribute fields, and errors or non-standard translations of geographic names（Guan et al., 2023）. A single data source cannot meet the dynamic update requirements of projects. Most existing research focuses on comparative analysis or service applications of open-source geographic names data（Song & Liu, 2016; Jiao et al., 2021）, such as downloading and converting a specific type of open-source geographic names data for a particular area to build a geographic names database that meets project application needs. Few studies have focused on the incremental extraction and integration of multi-source place names data for updating and translation to ensure the high currency, high precision, and high richness of geographic names data. Therefore, conducting efficient integration, accurate translation, and rapid updating of multi-source place names data from the internet is an urgent

issue that needs to be addressed in the geographic information industry（Cao et al., 2019）

Place names data consist of multiple attribute fields with spatial coordinates, where the attribute values are mostly strings and numerical values. Currently, scholars have proposed various algorithms for calculating string similarity, including literal similarity algorithms, matrix matching similarity algorithms, and edit distance algorithms (Zhao et al., 2019). Among these, the edit distance algorithm, as a commonly used method for solving string similarity, has certain advantages in data cleaning and high precision in detecting spelling errors. It is widely applied, efficient in searching, and has a relatively low time complexity. Based on this, this paper proposes a method for matching, integrating, translating, and updating multi-source place names data. By improving the matching algorithm, standardizing the translation methods for place names, and adding collaborative processing verification, this method aims to address issues such as low matching and integration efficiency, non-standard translations, low accuracy of results, and the lack of multi-source collaborative verification. Additionally, this approach provides a rational and feasible solution for the global integration, updating, and collaborative verification of multi-source place names.

2. Solution

The workflow of the proposed method for matching, integrating, translating, and updating multi-source place name data is shown in Figure 1. First, based on the analysis and comparison of the collected multi-source place names data, a data utilization plan is determined. The GADM administrative division dataset within the task area is obtained, and the lowest administrative surface layer is extracted as the basic unit for multi-source place name matching. Data preprocessing operations are then performed on the multi-source place name elements dataset, including format conversion, unification of spatial reference, structural normalization, data cleaning, classification mapping, and correction of Romanized spellings. Next, within the boundaries of the lowest administrative units, a string edit distance algorithm based on common name synonym substitution rules is used to calculate the similarity of place names and perform integration and updating. A multilingual translation method based on a transliteration rule library is employed for Chinese translation. Finally, the integrated and translated place names data are verified through collaborative processing with high-resolution remote sensing images and vector element data from the same period. This results in the formation of the updated place names dataset.

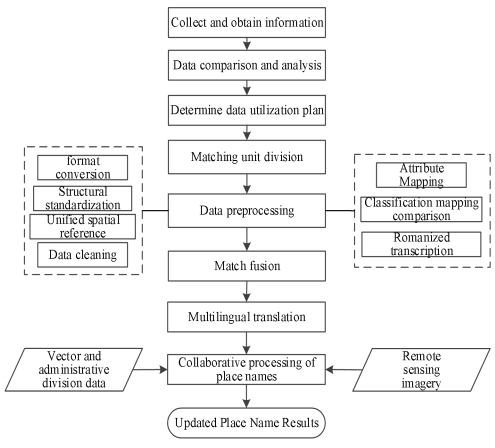

Figure 1. Process of multi-source place name data matching, fusion, translation and update method

3. Method

3.1 Data Utilization Plan

Collect and obtain relevant multi-source place name data within the task area. In addition to the widely used and authoritative place name data sources such as GeoNames and OSM, GADM global administrative division data can also be added. This type of data reflects the latest administrative divisions and place names of relevant countries (regions) and is a relatively reliable graphical material. Other sources include map publications compiled by various national publishing institutions such as the "World Standard Place Name Atlas," "National Boundary Drawing Samples of the World," and "The Times Atlas of the World," as well as Wikipedia, foreign ministry websites, relevant map service websites, real estate websites, cultural and tourism websites, high-resolution remote sensing images, and core vector elements. These sources offer advantages such as high precision, rich information, realistic and intuitive representation, and quick acquisition. The positional relationships between various elements are clear, providing an objective basis for the location and interpretation of different geographical entities.

3.1.1 Acquisition method: The online open source materials such as GeoNames, OSM, and GADM can be obtained, collected, and downloaded through their respective official websites or downloaders. The download scope can be divided into national, provincial, municipal, county, or any polygon; Other Internet place name websites can be crawled using algorithms such as breadth theme first crawler and depth theme first crawler; Map publications from various countries include purchasing from domestic and foreign sources, exchanging with foreign publishing partners, and receiving assistance from embassies of different countries; Map services and remote sensing images can be obtained through various methods such as online map services and offline downloads on the official website.

3.1.2 Data Analysis: The multi-source data materials are analyzed by referring to existing technical methods. The analysis covers several aspects, including whether the number of place names meets the density requirements, whether the attribute information and classification of place names meet the content requirements, and whether the spatial distribution characteristics reflect the actual situation(Song et al., 2016).

GeoNames data is divided into nine feature classes (A - Administrative divisions; H - Hydrography; L - Areas; P - Populated places; R - Roads and railways; S - Independent features; T - Relief features; U - Undersea features; V - Vegetation). These nine classes are further subdivided into over 670 feature codes. GeoNames also integrates various geographic information data, such as place names in different languages, population numbers, and update times.

OSM data includes administrative divisions, hydrography, roads, and place name points. The place name points are divided into six feature classes (transport - transport, traffic - traffic, pois - points of interest, pofw - religious, place - place names, natural - natural features), which are further divided into more than 200 subcategories. However, the geographic information data integrated in OSM place name points is relatively limited. Major place names are mostly duplicated with GeoNames.

Street-level place names such as cafes and retail stores account for a large proportion and can serve as an effective supplementary source for GeoNames place name data.

GADM is an open-access, high-precision global administrative division database that contains administrative boundary data for all levels of administrative units in every country and region worldwide. The dataset is organized by administrative divisions for different countries and regions, including national (regional), first-level, second-level, third-level, and fourth-level administrative divisions. All levels of data are represented as polygon features. The first-level administrative division data includes 16 fields, the second-level includes 18 fields, and the third-level includes 20 fields, such as name, type, HASC code, etc. GADM can serve as an important source for data collection and a reference for matching units in the integration and updating of place names data.

3.1.3 Clarification of Utilization Plan: Existing technical methods indiscriminately utilize multi-source place names data without considering the accuracy and differences among various data sources. In this paper, through comparative analysis and summarization of the coverage, update frequency, level of detail, classification, and positional accuracy of different data sources, a priority scheme for utilizing crowdsourced place names data is established: GeoNames is used as the primary place names dataset, OSM and others as supplementary place names datasets, GADM dataset as the basic unit for matching and integration as well as the boundary for administrative names, and national map publications, internet websites, and remote sensing imagery as auxiliary verification materials(Zhu et al., 2022).

3.2 Division of Matching Units

Traditional place name matching methods typically involve matching all elements or matching within a certain distance threshold from the center point. The former may result in a large computational load and low efficiency, while the latter may lead to the omission of data outside the threshold. Since place names have distinct regional characteristics and administrative attributes, and there is a certain degree of similarity and correlation among place names from different sources within the same administrative boundary, while identical names are rarely found across different administrative divisions, this paper proposes a matching method based on the lowest administrative unit. According to the scope of the task area, the lowest administrative division dataset of the corresponding country is obtained, and the lowest administrative division surface layer is extracted as the basic unit for matching and integrating place names. Within this basic unit, multi-source place name elements are matched and integrated one by one. In addition to recording the geographic spatial location range, the basic unit also records auxiliary verification information such as the administrative division name, administrative unit level, and language to which the place name belongs. By adopting this method, which is similar to an irregular grid indexing approach, the computational load for matching multi-source place name elements can be significantly reduced, thereby improving matching efficiency. At the same time, it minimizes the matching of place name elements across administrative divisions, enhancing the accuracy of matching.

3.3 Data Preprocessing

The GeoNames, OSM, and GADL administrative division datasets are converted to the universal geospatial database Geodatabase (*.gdb) format for storage. The geometric type is uniformly converted to the shapefile vector point feature data format, and the spatial reference is uniformly converted to the “GCS_WGS_1984” geographic coordinate system. Structural normalization, place name classification mapping, data cleaning, attribute mapping, and romanization transcription are then performed successively to ensure the structural consistency of the place name data sources.

3.3.1 Structural normalization and classification mapping: By establishing classification mapping rules between multi-source place names datasets, a one-to-one correspondence between the same geographic entities is ensured. A partial place name classification mapping table is shown in Table 1, and the basic attribute structure field table for the integrated place names is shown in Table 2. The latter includes fields such as a unique data source identification code for each place name, category code, foreign name, Chinese name, administrative division code, and collection/update date. Additionally, fields can be optionally added based on the incremental data sources. Other attribute data not included in the initial data sources should be retained and aggregated into a single attribute content field named "SRCATTR." The "ENAME1" field is filled with the standard Romanized spelling of the place name. The attribute values of the foreign name fields uniformly adopt the UTF-8 (Unicode Transformation Format-8bit) character encoding, which is efficient, flexible, and highly compatible. This eliminates issues such as software display errors caused by special characters in different foreign languages (e.g., Autovía de Andalucía in Spanish), facilitating subsequent name query matching, collaborative processing, and map annotation display.

GeoNames Category	OSM Category	Category Name
REST	restaurant 、cafe、food_court	Restaurant
GHSE、HTL、RHSE、HLT	guest_house	Hotel
CMPRF	refugee_site	Refugee Camp
MKT	public_market、farmers_market	Farmer's Market
SCH、SCHL	school/college、flight_school、language_school、music_school、dancing_school、art_school、circus_school、riding_school、sailing_school、ballet_school、tailor_school、tuition、place_of_worship、university land、special education、academy	Education
LEPC、CTRM	clinic、doctors、district hospital、diagnostic_center、polyclinic、orthopedic、medical_centre	Medical Institution
RECR、STDM、ATHF	sports_centre、horse_riding、dive_centre、stadium、sports_hall、Coliseum、arena、jump_park、diving_club	Sports Venue
PRK、PRKGT、PRKHQ	park、parklet、skatepark	Park
GOVL、ADMF、USGE	government、politics、politician、administration、agriculture deparment、public prosecution department、political_organizati	Government and Management Agencies

	on、congress_center	
STNB、STNM、STNE、CTRS、ASTR、ITTR、CTRA	archaeological_site、observatory、research_station、research_institute、astronomical_observatory、weather_station、institute	Research Institutions

Table 1. GeoNames and OSM Partial Place Name Classification Mapping Table

Name	Description	Name	Description
ID	Number	ADDRESS	Address
ENTID	Geographic Entity Code	EMAIL	Email
LOCALNAME	Name	TELEPHONE	Phone
ALIASNAME	Alias	PAC	Administrative Division Code (Level 2)
FORMERNAME	Former Name	TAG	Tags, separated by “|”
ABBREVIATION	Abbreviation	SRCCODE	Data Source Number
ENAME1	Foreign Name	SRCID	ID in the Data Source
ENAME2	Foreign Name 2, separated by “|”	SRCATTR	Data Source Attributes
CNAME1	Chinese Name	ACQDATE	Entry Date
CNAME2	Chinese Name 2, separated by “|”	UPDATETIME	Most Recent Update Date
OTHERNAME	Other Names, separated by “|”	DLGBJ	DLG Data Anomaly Mark
CLSID	Category Code	DMBJ	Place Name Data Anomaly Mark

Table 2. Merged Place Name Attribute Field Structure

3.3.2 Data Cleaning: Operations such as deleting abandoned, invalid, and duplicate place names from the normalized multi-source place name data can reduce data redundancy and improve the utilization rate of valid data and matching accuracy. Abandoned place names generally refer to those that are no longer in use or have no value, and place names with keywords such as “abandoned,” “destroyed,” “historical,” or “ruined” in their foreign names should be deleted. Table 3 provides some examples of abandoned place names from GeoNames. Invalid place names with both foreign names and category codes as null values should also be deleted. Duplicate place names generally refer to those that overlap or are close in location due to repeated collection and have the same foreign name. Using 2 kilometers as the threshold, place names with the same foreign name and category are deleted. The one with higher accuracy and stronger currency can be retained with the help of collection time and high-resolution imagery.

Category	Description
CNLQ	abandoned canal
PPLQ	abandoned populated place
RRQ	abandoned railroad
AIRQ	abandoned airfield
PPLW	destroyed populated place
BDGQ	ruined bridge
DAMQ	ruined dam
PPLCH	historical capital of a political entity
ADM4H	historical fourth-order administrative division
ADMDH	historical administrative division

Table 3. Partial Abandoned Place Names

3.3.3 Attribute Mapping: There are two types of attribute mapping, namely direct mapping and indirect mapping. For example, fields that can be directly mapped from GeoNames include: “name” mapped to “ENAME1”, “geonameid” mapped to “SRCID”, etc. Fields that require indirect mapping include: “alternate names” mapped to “OTHERNAME”, “feature code” mapped to “CLSID”, etc.

3.3.4 Romanization: Place names that are labeled and displayed in local languages need to be romanized according to the national language transliteration rules established by the respective country. Based on the transliteration system adopted in the World Standard Place Names Atlas, the inconsistent Roman spellings of place names should be corrected.

3.4 Place Name Matching and Integration

The Levenshtein distance is a type of edit distance. The principle of the algorithm is to calculate the minimum number of single-character edit operations required to transform one string into another. These edit operations include insertion, substitution, and deletion. It is widely used in spell checking, data matching, and text similarity calculation. However, in text similarity calculation, this algorithm, which is based on character-level editing operations and does not take into account the semantic information of strings, may lead to the judgment that strings with similar semantics but different surface forms are dissimilar. This situation often occurs with place name generics. Therefore, this paper proposes a string edit distance algorithm based on generic synonym replacement rules. By constructing a place name generic synonym replacement rule library and matching similar place name foreign strings one by one, the algorithm first converts shorter generic synonyms in the foreign names into longer generics before calculating similarity. It then compares the calculated similarity with a predefined threshold to determine whether to integrate place names from multiple sources, thereby improving the accuracy of place name matching and integration. The specific algorithm is as follows:

3.4.1 Construction of Synonym Replacement Rule Library: Common generic synonyms with the same semantics and high frequency of occurrence are collected and summarized to establish a place name generic synonym replacement rule library (GSRI). This includes abbreviations, shortenings, acronyms, case variations, and synonyms of character names.

For example, "Road" and "Rd" are generic synonyms, both of which are translated as "road." In string similarity calculations, these generic synonyms are considered to represent the same place name and are converted to the longer string generic "Road" before participating in the calculation.

3.4.2 Place Name Similarity Calculation: Iterate through the place name elements of the same category (i.e., with the same CLSID) in the primary place name dataset and the supplementary place name dataset within the lowest administrative unit. First, refer to the GSRI to compare whether the generics of the two foreign names are generic synonyms. If they are, convert the shorter generic to the longer one and then calculate the string similarity of the foreign name ENAME1 of the place name elements according to formula (1).

$$S_{ij} = \left(1 - \frac{E_{ij}}{L_{ij}}\right) * 100\% \qquad (1)$$

where i = place name category
j = total number of place names in each category
S_{ij} = the name similarity between the two place name elements, with a higher value indicating greater similarity
E_{ij} = the Levenshtein edit distance between two place name elements
L_{ij} = the maximum length of the foreign name strings of the two place name elements

When S_{ij} is greater than 90%, the supplementary place name data is merged, that is, the primary place name data is retained and the supplementary place name data is deleted. When S_{ij} is less than 90%, the supplementary place name data is added, and both the primary and supplementary place name data are retained.

3.5 Multilingual Translation

This paper employs a multilingual translation method based on a transliteration rule library to translate the integrated place name data into multiple languages. Chinese is selected as the specific language for multilingual translation to meet the browsing and travel needs of the vast number of users in China. The translation process for other languages can be carried out by referring to this solution. The specific process is as follows: First, analyze the components of the place name to be translated according to the characteristics of character expression, etymology, and spelling rules, based on the constructed transliteration rule library. This involves identifying the language of the place name (such as Russian, German, English, etc.) and extracting the proper noun and generic term. Second, refer to the corresponding language's Chinese transliteration guidelines for translation. Specifically, perform a generic term translation for the generic term and some directional words, adjectives, and characteristic words. For the proper noun, first perform syllable segmentation of the place name, and then conduct a phonetic translation. Next, adjust the word order of the translated proper noun and generic term results. For example, if the foreign generic term is placed at the beginning, it should be moved to the end in the Chinese translation. Finally, output the complete transliterated place name result.

The transliteration rule library is used to store the multilingual transliteration rules and phonetic translation rules required during the place name translation process. If a country or region lacks a phonetic table, new customized phonetic conversion rules need to be added. This paper takes the "Guide to the Transliteration of Foreign Place Names into Chinese Characters," "General Rules for the Transliteration of Foreign Place Names into Chinese Characters," and the standard "Phonetic Table" of each country as the basic standards for place name transliteration. In addition, during translation, it is necessary to follow the general principles of respecting the original name, following established conventions, and retaining customary usages. The translation of personal names and proper nouns in place names refers to the transliteration methods in the "Encyclopedia Britannica" and the "Great Dictionary of World Personal Name Translation." For place names in disputed areas, it is ensured that they conform to China's political stance.

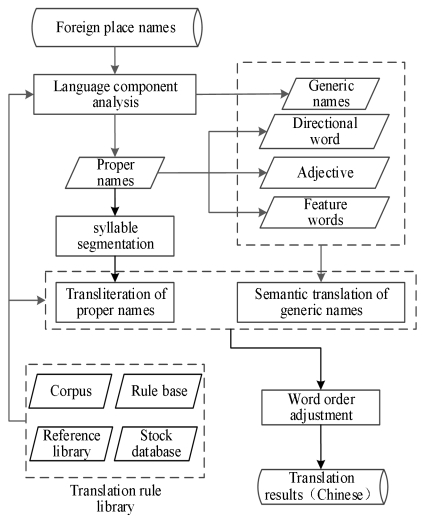

Figure 2. Multilingual translation process based on translation rule library.

Component analysis is used to support the extraction and identification of the components of proper nouns, generic terms, adjectives, directional words, and other word-forming elements in place names. The reference materials include the "General Rules for the Transliteration of Foreign Place Names into Chinese Characters" (China Place Name Committee), "Guide to the Transliteration of Foreign Place Names into Chinese Characters" (GB/T 17693), "Great Dictionary of World Place Name Translation" (Zhou Dingguo), and "Dictionary of World Place Name Translations" (Ministry of Civil Affairs Place Name Research Institute).

Syllable segmentation involves dividing the proper noun of a place name into syllable combinations according to the phonetic rules of different languages. Subsequently, the syllable combinations can be directly converted into Chinese using phonetic tables and other rules. The rule library is a reference for syllable combinations formed on the basis of analyzing a large number of syllable segmentations. Different languages have different segmentation algorithms based on their characteristics. For example, the syllable combinations of Russian, Arabic, and Indonesian have fixed pronunciations, which can be segmented directly based on the order of consonants and vowels. In contrast, the pronunciation of English syllable combinations is not fixed and is difficult to segment directly, requiring the use of the corresponding rule library for syllable segmentation.

3.6 Place Name Collaborative Processing

3.6.1 Data Collaboration: To ensure the consistency and rationality of place name data with basic geographic information data such as images and vectors in terms of spatial location and name, it is necessary to conduct an initial collaborative check between the integrated and translated place name data and Google imagery data. Subsequently, a final collaborative check should be performed with digital orthophoto maps (DOM) and core vector elements (DLG) results. This mainly includes corrections of locations such as bays, straits, harbors, docks, shipping lanes, first and second-level administrative centers, railway stations, waterfalls, airports, dams, and mining areas. The specific technical requirements are as follows.

Place Name	Collaborative Object	Technical Requirements
Bay and Strait	Coastline, Imagery	Located outside the coastline or on areal water system.
Railway Station	Railway, Imagery	No more than 200 meters away from the railway.
Port and Dock	Coastline, Areal River, Linear River, Imagery	The place name point is no more than 200 meters away from the river, or no more than 1000 meters away from the coastline on land.
Waterway	Coastline, Linear River, Areal River, Imagery	No more than 100 meters away from the linear river, or within the areal river, or outside the coastline.
Waterfall Point	Areal River, Imagery	Reasonably retained, such as dams, weirs, lakes, depressions, and rivers. Unreasonable ones should correct the location, such as provincial place names and settlements.
First and Second-level Administrative Centers	Imagery	The place name point is located in the bustling urban area or residential area.
Airport		The place name point is within 100 meters of the airport. Incorrectly collected ones should be deleted.
Dam		The place name point is within 50 meters of the dam. Incorrectly collected ones should be deleted.
Mining Area		The place name point is within 200 meters of the mining area.

Table 3 Technical Requirements for Collaborative Processing of Place Name Data

3.6.2 Edge collaboration: Data edge matching should be conducted between place name datasets submitted by different production units for the same geographic entity, ensuring no duplication or omission of place names at the edges and the uniqueness of national and administrative place names across administrative boundaries.

4. Implementation

4.1 .Overview of the Experimental Area

The Democratic Republic of the Congo has a land area of 2.3449 million square kilometers and a population of 102.3 million (as of 2023). It is divided into 26 provinces and 239 second-level administrative divisions, with Kinshasa as the capital. French is the official language, and the officially recognized ethnic languages are Lingala, Swahili, Kikongo, and Kiluba. It is one of the world's least developed countries as identified by the United Nations. Agriculture and mining are the dominant sectors of the economy, while the processing industry is underdeveloped and the country is not self-sufficient in food.

4.2 Analysis of Data Sources

The initial data was collected in January 2024, including 66,424 Geonames place names and 160,713 OSM place names. After statistical analysis, it was found that within the Congo test area, OSM place names have a smaller spatial scale and a richer variety of place name categories compared to Geonames. In this paper, place name categories such as administrative divisions, roads, railways, hydrography, and those with no productive value (such as small shops, cafes, signposts, etc.) were removed from the Geonames and OSM place name data. After data preprocessing according to the method proposed in this paper, the main Geonames place name dataset consisted of 40,276 entries, and the supplementary OSM place name dataset consisted of 11,149 entries.

4.3 Result Analysis

After the primary place name dataset and the supplementary place name dataset underwent place name matching, integration, translation, and collaborative processing, a total of 51,291 place names across 17 categories within the borders of the Democratic Republic of the Congo were obtained, with the distribution of place name points shown in Figure 3. Among these, 40,114 entries were from GeoNames, 9,991 from OSM, 266 from GADM, and 920 from other sources such as map publications, internet mapping services, and image capture.

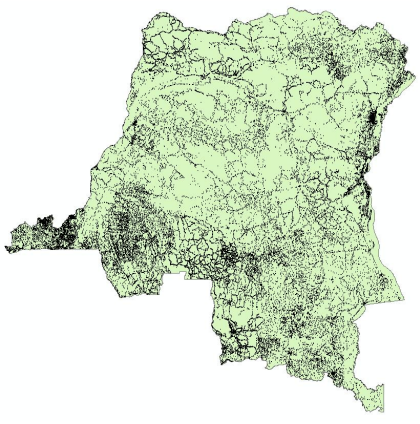

Figure 3 Distribution of Place Name Data Results in the Democratic Republic of the Congo.

Among them, the lowest administrative unit level available in the Democratic Republic of the Congo is the second-level administrative region, which includes 1 national place name, 26 first-level administrative place names, and 239 second-level administrative place names. The first and second-level

administrative regions and their annotations are shown in Figure 4.

a. First-level b. Second-level

Figure 4 Place Names of First and Second-Level Administrative Regions in the Democratic Republic of the Congo

Place names of populated places, administrative regions, and natural features account for a relatively high proportion, among which place names of addresses account for about 76.9% of the total number of place names. Education and culture account for 9.8%, public facilities account for 5.9%, other infrastructure accounts for 2.7%, health and social security account for 1.9%, and other types of place names account for less than 1% each. The detailed situation is shown in Table 4.

Place name categories	Number of place names
Accommodation	101
Wholesale, Retail	423
Finance, Insurance	141
Education, Culture	5031
Health, Social Security	980
Sports, Recreation	182
Public Facilities	3030
Commercial Facilities, Services	0
Residential Services	28
Enterprises	92
Transportation, Storage	356
Scientific Research and Technical Services	11
Agriculture, Forestry, Animal Husbandry, Fishery	31
Place Names, Addresses	39444
Other Infrastructure	1397
Military Facilities	14
Ecology	30

Table 4 Statistics of Place Names by Category (Unit: Entries)

5. Conclusion

This paper categorizes multi-source place name data based on priority, utilizes GADM to reduce misalignments of place names across administrative boundaries, uniformly designs multiple data preprocessing steps, and constructs a database of rules for replacing common synonyms to calculate place name similarity, thereby enhancing the efficiency and precision of matching and integration for multi-source place name data. The proposed multilingual translation method based on a translation and writing rule database not only efficiently translates Chinese results but also offers reference experience for inter-translation among other languages. The collaborative processing of place name data not only validates the accuracy between different multi-source place name data but also aids in the interpretation of place name location, improving the consistency and rationality of the spatial alignment relationship between place name data and other fundamental geographic information data products. The technical approach and methods presented in this paper can provide reference ideas and experience for the integration and translation of multi-source data from different countries and languages.

References

Cao, C.X., Huo, L., & Zhu, X.L., 2019. Research on global place name translation and matching method based on multiple data sources. Science of Surveying and Mapping, 44(7), 171-176. DOI: 10.16251/j.cnki.1009-2307.2019.07.027.

Dong, H.F., Zhou, X.G., & Zhao, B.L., 2020. Automatic Metho d of Ouality Control and Evaluation for Incremental Updating o f SimpleObiectives in OSM. Geomatics World, 27(1), 13-19.

Guan, Q., Long, Y.T., Si, L.F., Wang, M.H., Zhang, D., He, F., & Hou, X.Y., 2023. A method for constructing a global Chinese -foreign language place name data resource based on crowd-sou rced place name data. Bulletin of Surveying and Mapping, (02), 167-171+181. DOI: 10.13474/j.cnki.11-2246.2023.0059.

Jiao, C.Y., Cheng, Y., Ge, W., & Xu L., 2021. Study on Quality and Usability of Open Source Geographic Database in Africa. Geomatics World, 28(2), 95-99.

Li, Y.P., & Liu, G.D., 2024. Evaluation of street spatial quality based on street view and POI data. Geospatial Information, 22(2) , 64-67.

Song, H.B., & Liu, X.G., 2016, Comparative Analysis of Multi-source Geographic Data. Geomatics & Spatial Information Technology, 39(10), 114-117+121.

Wei, Y., Hu, D.L., Li, X., & Wang, F., 2016. Geographic name full text query based GeoNames and Solr. Engineering of Surveying and Mapping. 25(2), 28-32. DOI: 10.19349/j.cnki.issn1006-7949.2016.02.006.

Zhang, B.G., & Shi, S.Z., 2012. Application of Place Names to Information-enabled Geomatics. Bulletin of Surveying and Map ping, (08): 60-61+86.

Zhang, X.Y., Lv, G.N., Du, M., & Ye, P., 2017. Acquisition and Application on Geographical Names Information Based on Large Data Driving. *Modern Surveying and Mapping*, 40(2), 1-5.

Zhao, W.Q., 2019. Research on Multi-source Global Toponymi c Data Fusion and Updating Methods. (Doctoral dissertation), N anjing Normal University, DOI: 10.27245/d.cnki.gnjsu.2019.00 2390.

Zhu, Y., Cao, Y.X., & Dong, M.L., 2022. Road Vector Data Up dating Method Based on Remote Sensing Image. Geomatics & Spatial Information Technology, 45(4), 108-111+114. DOI:10.1 9349/j.cnki.issn1006-7949.2016.02.006.