UDC 528.001+681.518

Nadiia LAZORENKO

Department of Geoinformation system and photogrammetry, Kyiv National University of Construction and Architecture, 31 Povitroflotsky Ave, Kyiv, 03037, Ukraine, e-mail: nadiialg@gmail.com, https://orcid.org/0000-0002-1572-49471

https://doi.org/10.23939/istcgcap2022.95.113

INTEGRATION OF GEOSPATIA L DATA BASED ON THE APPLICATION OF THE JOIN OPERATION OF RELATIVE ALGEBRA

The purpose of this work is to study the integration of sets of core reference and thematic geospatial data based on the JOIN operation of relational algebra and its interaction with geocoding of geospatial features, which is implemented in modern geographic information systems (GIS) and database management systems (hereinafter – DBMS) for the development of the national spatial data infrastructure (hereinafter – NSDI). Method. The research is based on the analysis of the possibilities of applying the theory of geospatial databases and knowledge bases, international and national harmonized standards in the field of Geographic Information/ Geomatics to solve the problem of integration of geospatial data using the operation JOIN relational algebra in object-relational database management systems (OR DBMS). Results. The paper examines the models of the Join operation of relational algebra, which underlie the geocoding of features and the creation of electronic gazetteers, and proves its effectiveness: the Join operation integrates of core reference and thematic geospatial datasets. There is a need to define the required geographic identifiers, which must be present among the attributes of the core reference and thematic geospatial datasets to perform the join. The variety of uses of the Join operation covers all possible cases that arise in their practical application. Thus, the use of the Join operation involves identifying these required geographic identifiers at the geospatial database design stage. In particular, it is expedient to determine mandatory geographical identifiers (codes) of features according to the official national systems of features classification (codification) in the relevant sectoral thematic registers, which are responsible for certain holders of thematic data in accordance with Annex 2 of the Decree of Cabinet of Ministers "The order for the functioning of the national spatial data infrastructure" of May 26, 2021, No. 532. Scientific novelty and practical significance. The integration of core reference data and thematic geospatial datasets based on JOIN operation models of relational algebra and their interaction with geocoding of geospatial features is researched, which is implemented in modern GIS and DBMS for the development of national spatial data infrastructure. The research was performed on a set of core reference spatial data, namely: information on administrative-territorial units of the Cherkasy region, including their borders; the data from the statistical bulletin of the socio-economic situation of the Cherkasy region for January 2021 of the Main Department of Statistics in Cherkasy region of the State Statistics Service of Ukraine were selected as thematic data. It has been shown that relational algebra join (JOIN) operations can be used to integrate other thematic geospatial data with core reference data using geographic identifiers that contain these datasets.

Key words: geospatial data integration, interoperability, JOIN operation, national spatial data infrastructure, core reference data, thematic geospatial data.

Introduction

The Law of Ukraine "On the National Spatial Data Infrastructure of Ukraine", No. 554 was adopted on April 13, 2020, which was contributed to the development National Spatial Data Infrastructure (hereinafter – NSDI) in Ukraine as an interconnected set of organizational structure, hardware and software, core reference and thematic geospatial datasets, metadata, services, technical regulations, standards, technical specifications required for production, updating, processing, storage, publication, use of geospatial data and metadata, other activities with such data [Law of Ukraine, 2020]. The Decree of the Cabinet of Ministers of Ukraine "On Approving the Procedure for the Functioning of the National Spatial Data Infrastructure" No. 532 was adopted on May 26, 2021. Core reference data are the core of the NSDI because they "ensure the production and use of thematic geospatial data, constitute a unified single topographic basis for interoperability of geospatial data, their integration and interagency information interaction. The basis for creation sets of core reference data is databases of topographic data and information of the State Land Cadastre, State Register of Geographical Names, State Address Register, Urban Cadastre and Cadastres of Natural Resources, as well as other geographic information resources" [Decree, 2021]. Usually, the attributive characteristics of topographic features that belong

to the core reference data are collected as a result of topographic surveys [Karpinskyi & Lazorenko–Hevel, 2018; Karpinskyi & Lazorenko–Hevel, 2020], so they usually do not contain complete thematic information about these features. Thematic geospatial datasets are formed by executive authorities and local governments and other data holders responsible for these datasets in accordance with Annex 2 of the Decree of the Cabinet of Ministers of Ukraine "On Approving the Procedure for the Functioning of the National Spatial Data Infrastructure" No. 532 [Decree, 2021].

The Ministry of Agrarian Policy and Food of Ukraine approved the Order "On approval of technical requirements for geospatial data, metadata and geoinformation services of the National Spatial Data Infrastructure" No. 347 of 10.11.2021, which states that the unification and integration of core reference and thematic geospatial data provide work NSDI [Order, 2021].

One way to integrate data is to join different geospatial databases with data that do not contain a geospatial component, using a system identifier to compile them from different sources in the NSDI. There are 2 ways to integrate data: direct geocoding, which is carried out as a result of topographic and geodetic works, and indirect (not direct), in which geographical identifiers play a key role. It is possible to connect geospatial coordinate descriptions of features with an array of attribute data that do not have a direct positioning using such geographical identifiers [Geographic information… DSTU ISO 19112: 2017, 2017; Karpinsky & Lazorenko–Hevel, 2020; Karpinskyi et al., 2020; Karpinskyi & Lazorenko–Hevel, 2020; Lazorenko–Hevel, 2021].

Geocoding is one of the most common geospatial tools and services in all popular geographic information systems (hereinafter – GIS) and database management systems (hereinafter – DBMS) which determines geospatial information (for example coordinates x, y) for location escriptions by comparing the elements of the location description with those in the reference data registers: addresses, geographical names, etc.

In other words, geocoding is the matching of the registers to the features on the map, for which there is a need to have a database of the feature's locations, so-called gazetteers – a directory of geographical identifiers describing location instances.

These directories contain additional information regarding the position of each location instance and may include a coordinate reference, but may be purely descriptive. If the gazetteer contains a coordinate reference, it provides a transformation from a spatial reference system using geographical identifiers to another coordinate reference system. If it contains a descriptive reference, it will be a spatial reference using another spatial reference system with geographic identifiers, such as the zip code of the property. There can be more than one gazetteer for any type of location. Descriptive reference can be defined in the spatial reference for the geographic identifier in the gazetteer [The National Standard of Ukraine DSTU 8774:2018, 2018; Lazorenko-Hevel, 2021].

Nowadays there are many desktop and web geocoding tools in GIS. For example, the open-source Quantum GIS (QGIS) geographic information system uses a special MMQGIS Plugin for geocoding, which is a set of Python plugins for manipulating vector map layers in QGIS: input/ output/join in *.CSV format, geocoding with street layer, which allows to create address registers from data sources in CSV format, geocoding in CSV format with a web service that allows to geocode address tables through geocoding services, geometry transformation, and more. There are such services:

1) Nominatim is a worldwide open source geocoding service based on data OpenStreetMap;

2) The Geocoding service OSMNames is derived from OSM data, the entire database is downloadable and there is (at least some) the ability to query through their API;

3) Google world geocoding service or Google gazetteer is Google's worldwide geocoding service, which requires a prior geocoding API key and has a daily limit (currently 2.500 addresses);

4) Companies Mapbox and HERE have their own geocoding services. Geopy is a geocoding client for several popular geocoding web services, including Nominatim and Google;

5) ESRI World Geocoding Service is a world-wide geocoding service from ESRI company. The Python programming language – ArcPy, has access to geocoding tools in ArcGIS [How to Geocode, 2022].

In commercial ArcGIS uses several geocoding tools, 3 of them are basic: Add X, Y Coordinate Data as a Map Layer, Create an Address Locator,

Create an Address Locator, and the ArcGIS World Geocoding Service, which allows to find addresses or locations around the world, geocode address tables or perform reverse geocoding, etc., using the ArcGIS Rest API. Companies create their own address registers to work with confidential data.

QGIS also can use the Pelias Geocoding plugin, which allows to perform geocoding using the open-source world geocoder remote service Pelias, which is available on openrouteservice and geocode.earth [Geocoding, 2022].

Analysis of recent research

The issues of integration of geospatial data and its development trends were investigated in such scientific works [Gao et al., 2005; Silberschatz et al., 2011; Franci et al., 2014; Bhattacharya & Painho, 2017; Mardani et al., 2019; Sun et al., 2019].

The peculiarities of integration geospatial data for the development of NSDI, urban cadastre, and other sectors of the economy were considered in such articles [Hansen, 1999; Maksymova, 2016; Pilicheva et al., 2018; Lemenkova, 2020; Lyashchenko & Cherin, 2019, Lyashchenko et al., 2020; Lyashchenko et al., 2021; Stankevich et al., 2021; Shypulin, 2021].

The substantiation and application of the Join operation models of relational algebra are described in scientific works [Codd, 1990; Rayordan, 2001; Connolly & Begg, 2003; Gao et al., 2005; Silberschatz et al., 2011; Glushko, 2013; Bui & Glushko, 2015]. It is well known that relational algebra and the model proposed by E. Codd in the 70s of the XX century and is the basis of many modern DBMS and query languages, in particular SQL (Structured query language), which support the relational model. E. Codd offered 9 relational algebra operations: traditional operations on sets such as union, intersection, difference, and special operations on tables: projection, Cartesian product, θ-join and equijoin, division, selection) [Codd, 1990]. The issue of integration of geospatial data is solved in the international and harmonized national standard DSTU ISO 19112: 2017 (ISO 19112: 2003, IDT) Geographic information. Spatial reference by geographic identifiers [Geographic information… DSTU ISO 19112:2017, 2017].

The article [Lazorenko-Hevel, 2021] examines the subject, idea, role, and meaning of geographic identifiers to ensure the integration of geospatial data in seamless topographic databases and national spatial data infrastructure in accordance with the national standard DSTU ISO 19112:2017 (ISO 19112: 2003, IDT) Geographic information. Spatial referencing by geographic identifiers. The purpose of the national standard DSTU ISO 19112:2017 (ISO 19112:2003, IDT) is the methods to define and describe conceptual-level spatial reference systems using geographic identifiers. The location of the feature is identified by spatial reference. When a geographic identifier is used for spatial reference, it provides a unique identification of the location of the feature. This location is a place or position uses to reference other features [Lyashchenko et al., 2020; Lazorenko-Hevel, 2021].

However, the issue of integration of geospatial data using the JOIN operation in object-relational database management systems (OR DBMS) is not sufficiently considered in the scientific literature.

Aim

The article aims to research the integration of core reference and thematic geospatial datasets based on the JOIN operation of relational algebra and its interaction with geocoding of geospatial features, which is implemented in modern GIS and DBMS to develop national spatial data infrastructure.

Methodology

The research is based on the analysis of the possibilities of applying the theory of geospatial databases and knowledge bases, and international and national harmonized standards in the field of Geographic Information/Geomatics to solve the issue of integration of geospatial data using the JOIN operation of relational algebra in OR DBMS.

Results

The JOIN operation of relational algebra. The JOIN operation in SQL notation is the operation of joining attributes from one or more tables into a new table using Cartesian product and selection. In relational algebra, the join of two relations under a specific condition is called a relation $(A\ TIMES\ B)\ WHERE\ C$, where A and B – any relation; C – logical expression, which may include attributes of relations A and B and/or scalar

expressions (Fig. 1). The join operation is the result of the consistent application of Cartesian product and selection. If there are attributes with the same names in relations A and B, such attributes are renamed before the join is made. [Connolly & Begg, 2003].

The join operation in SQL notation based on relational algebra is divided into the following types: *Inner Join*, *Outer Join*, *Cross Join*, *Self-Join,* and *Spatial Join*. The *Inner Join* is divided into *θ-Join*, *Equi-Join*, *Semi-Join,* and *Natural Join*. The *Outer Join* is divided into *Left Join*, *Right Join,* and *Full Join*. A special case of the *Equi-Join is Natural Join* and a special case of the *Semi-Join* is *Anti-Join*.

The classification of join operation types is in the UML diagram, inheritance connection is established between operation classes (Fig. 2). The DBMS and GIS have additional features and functions for processing geospatial data: spatial data types, spatial operators, spatial application procedures, and spatial indexing.

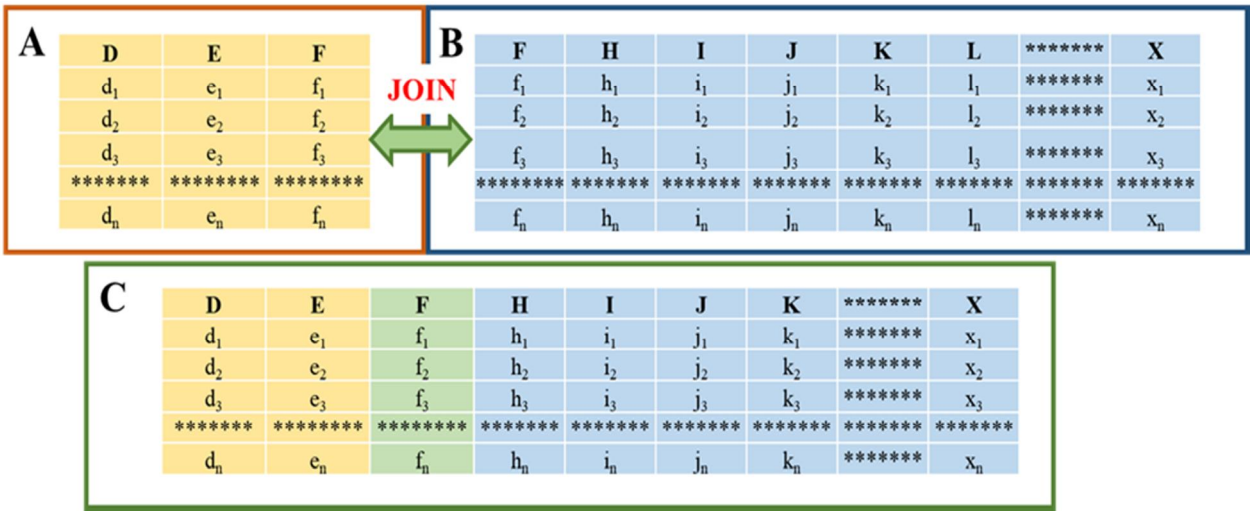

Fig. 1. The integration of core reference (relation A – yellow) and thematic (relation B – blue) geospatial datasets in the geospatial database [Karpinskyi & Lazorenko-Hevel, 2020]

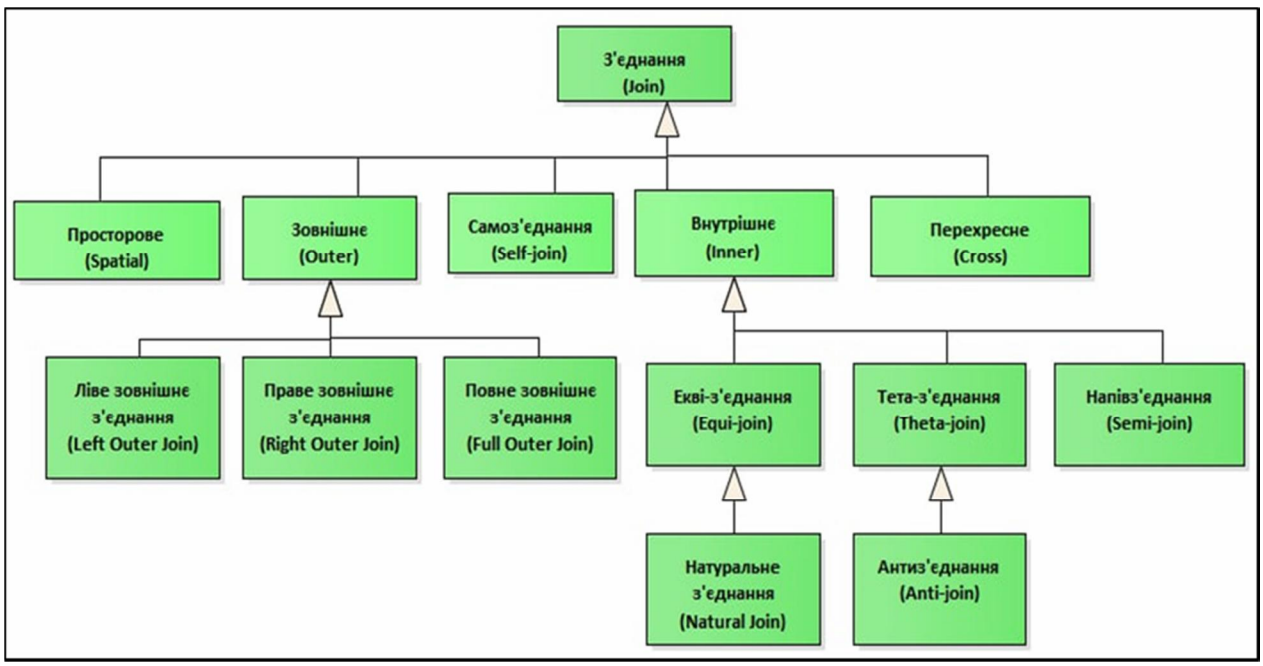

Fig. 2. UML diagram classification of Join operation

These possibilities allow to use of the spatial join operation, taking into account the geometry of the features along with their identifiers.

This paper investigates the possibility of using the JOIN operation of relational algebra to integrate core reference and thematic geospatial datasets in ORDBMS PostgreSQL. The information about administrative-territorial units of the Cherkasy region, including their borders dated 2019 is accepted as the core reference geospatial dataset (relation A) [Geoportal… http://atu.gki.com.ua/], which contains the following attributes: *COATUU, geom, name_ua,* and others. The fragment of A relation is in Table 1, where: *COATUU* is the Classifier of administrative-territorial units of Ukraine; geometry is an attribute that contains the coordinates of features in WKT format (Well-Known Text); name_ua is an attribute which means the Ukrainian name of the feature of administrative-territorial units. It should be noted that there are also newly formed districts in relation A, which are assigned not COATUU, and KATOTTC in accordance with the Resolution of the Verkhovna Rada of Ukraine "On the formation and liquidation of districts" from 17.07.2020 No. 807-IX. So, the COATUU attribute has NULL pseudo-values in the tuples of newly formed districts. The number of tuples in relation A is 30 records.

Table 1

Fragment of A relation. The data on the districts of the Cherkasy region

COATUU	geom	name_ua
7121500000	MULTIPOLYGON(((6392148.42197378 5478804.02538396...)))	Zolotoniskyi
7121200000	MULTIPOLYGON(((6380005.8326648 5402075.15493481...)))	Zvenyhorodskyi
7120600000	MULTIPOLYGON(((63356005.8326648 5412565.15493481...)))	Drabivskyi
7120900000	MULTIPOLYGON(((63569775.8326648 5434895.15493481...)))	Zhashkivskyi
NULL	MULTIPOLYGON(((6386787.50122662 5493188.28142229...)))	Cherkaskyi
NULL	MULTIPOLYGON(((6450505.70254616 5545812.54783928...)))	Umanskyi

The thematic data (relation B) are based on data from the statistical bulletin of the socio-economic situation of the Cherkasy region dated on January 2021 of the Main Department of Statistics in the Cherkasy region of the State Statistics Service of Ukraine, namely indicators of the demographic situation. A fragment of the relation B is in table 2 [Statistical bulletin, 2021]. The State Statistics Service of Ukraine still uses the Classifier of Objects of Administrative and Territorial Units of Ukraine (COATUU) to collect and store statistical data and gradually moves to the Codifier of Administrative-Territorial Units and Territories of Territorial Communities (CATUTTC) in connection with the Verkhovna Rada Resolution from 17.07.2020 No. 807-IX "On formation and liquidation of districts", therefore COATUU lost its relevance. The Ministry of Community and Territorial Development has developed and approved the Codifier of administrative-territorial units and territories of territorial communities (the Order of 26.11.2020 No. 290 as amended by the order of 12.01.2021 No. 3), which corresponds with current legislation and is implemented to replace COATUU [Classifier, 2022].

The number of tuples in relation B is 26 records.

The COATUU and KATOTTC are indirect or geographical identifiers that are available in relations A and B and the article investigated them for the join of the geospatial component of the core reference data with the attribute of thematic geospatial data that do not contain coordinates.

If there is a need to perform a join on the same COATUU tuple and its attributes have the same values, then the operations of natural and external join are used.

The Natural Join. There are the relations $A(A_1, A_2, \dots, A_n, X_1, X_2, \dots, X_p)$ and $B(X_1, X_2, \dots, X_p, B_1, B_2, \dots, B_m)$, which have attributes X_1, X_2, …, X_p with the same names and defined on the same domains. Then the natural join of relations A and B is called the relation with the title $(A_1, A_2, \dots, A_n, X_1, X_2, \dots, X_p, B_1, B_2, \dots, B_m)$ and the body is containing many tuples $(a_1, a_2, \dots, a_n, x_1, x_2, \dots, x_p, b_1, b_2, \dots, b_m)$, such that $(a_1, a_2, \dots, a_n, x_1, x_2, \dots, x_p) \in A$ and $(x, x_2, \dots, x_p, b_1, b_2, \dots, b_m) \in B$.

Table 2

Fragment of B relation. The data on the socio-economic situation of Cherkasy region is dated on January 2021

COATUU	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7121500000	Zolotoniskyi	39031	39339	218	824	441
7121200000	Zvenyhorodskyi	41845	41887	282	804	503
7120600000	Drabivskyi	32602	32806	154	717	269
7120900000	Zhashkivskyi	34831	34726	235	756	430

For a Natural Join use the syntax: *A* JOIN *B*. A Natural Join has the property of associativity: (A JOIN B) JOIN C = AJOIN (B JOIN C), so such joins are written by lowering the brackets: A JOIN B JOIN C [Connolly & Begg, 2003]. The selection was performed for all attributes of the geospatial database table to display the result of the natural join:

SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals

FROM public.cherkasy

NATURAL JOIN public.stat

The natural join returned records for which the join condition was similar (Table 3).

Performing a join operation allows to build a variety of thematic maps based on the various attributes attached. For example, Fig. 3 shows the thematic map of Cherkasy region districts by the attribute number of live births.

The **Outer Join** returns all records returned by the inner join and all records involved in the join from one or two datasets. The missing values for which there is no match will be replaced by *Null* values.

The Outer Join can be divided into several groups depending on which additional records are included in them: left, right, and full. A **left outer join** returns all values from the left table and adds attribute values from the right table or NULL if there is no match predicate match, and the **right outer join** returns all values from the right table. A **full outer join** combines the results of the left and right outer joins. The resulting table contains all records from both tables, indicating NULL values with no matches from each table [Connolly & Begg, 2003]. Tables 4 to 6 below show the results of queries using LEFT OUTER JOIN, RIGHT OUTER JOIN, and FULL OUTER JOIN operations.

It is important to note that these results have an identical cartographic image (Fig. 3), but the order of the joined attributes is different from the joining tables. For example, the left and full joins take into account the values of the first table *public.cherkasy*. The results of these queries contain NULL values because the attributes of the COATUU formed for the 4 new districts are not specified, so the join is failed.

SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals

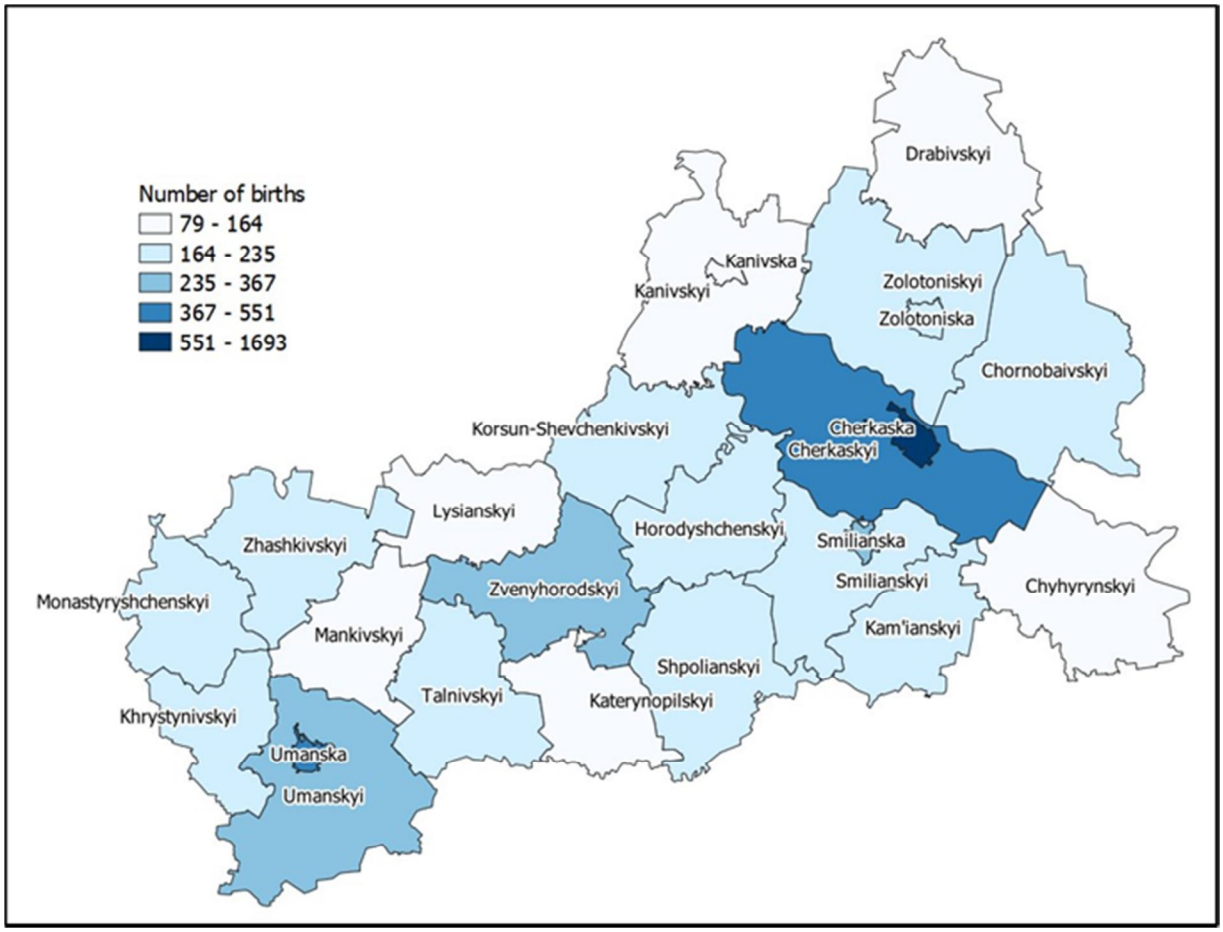

Fig. 3. The resulting thematic map: left, right, and full external join

Table 3

The result of Natural Join

COATUU	geom	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7120600000	MULTIPOLYGON (((6392148.42197378 5478804.02538396...)))	Drabivskyi	32602	32806	154	717	269
7120900000	MULTIPOLYGON (((6380005.8326648 5402075.15493481...)))	Zhashkivskyi	34831	34726	235	756	430
7121200000	MULTIPOLYGON (((63356005.8326648 5412565.15493481...)))	Zvenyhorodskyi	41845	41887	282	804	503
7121500000	MULTIPOLYGON (((63569775.8326648 5434895.15493481...)))	Zolotoniskyi	39031	39339	218	824	441

FROM public.cherkasy
LEFT OUTER JOIN public.stat
ON cherkasy.COATUU = stat.COATUU;

SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals

FROM public.cherkasy
RIGHT OUTER JOIN public.stat
ON cherkasy.COATUU = stat.COATUU
SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals
FROM public.cherkasy
FULL OUTER JOIN public.stat
ON cherkasy.COATUU = stat.COATUU

The query for a θ-Join operation must contain a condition. For example, the value of the attribute "permanent population" of the relation B must be greater than or equal to the value of the attribute "population" of the relation A:

SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals
FROM public.cherkasy
INNER JOIN public.stat
ON cherkasy.COATUU = stat.COATUU
WHERE stat.permanent population >= population

The result of the θ-Join in the form of a thematic map of the districts of the Cherkasy region by the attribute of the available population is in Fig. 4.

Equi-Join. A special case of the θ-Join, when θ is just equality. Equi-join syntax is $A[X = Y]B$ [Connolly & Begg, 2003].

The query using the equi-join operation must contain a condition. For example, the value of an attribute "permanent population" of the relation B must be equal to the value of the attribute "population" of the relation A:

SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals
FROM public.cherkasy
INNER JOIN public.stat
ON cherkasy.COATUU = stat.COATUU
WHERE stat.permanent population = population

Table 4

The result of the left external join

COATUU	geom	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7121500000	MULTIPOLYGON (((6392148.42197378 5478804.02538396...)))	Zolotoniskyi	39031	39339	218	824	441
7121200000	MULTIPOLYGON (((6380005.8326648 5402075.15493481...)))	Zvenyhorodskyi	41845	41887	282	804	503
7120600000	MULTIPOLYGON (((63356005.8326648 5412565.15493481...)))	Drabivskyi	32602	32806	154	717	269
7120900000	MULTIPOLYGON (((63569775.8326648 5434895.15493481...)))	Zhashkivskyi	34831	34726	235	756	430
NULL	MULTIPOLYGON (((6386787.50122662 5493188.28142229...)))	Cherkaskyi	NULL	NULL	NULL	NULL	NULL
NULL	MULTIPOLYGON (((6450505.70254616 5545812.54783928...)))	Umanskyi	NULL	NULL	NULL	NULL	NULL

Table 5

The result of the right external join

COATUU	geom	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7121500000	MULTIPOLYGON (((6392148.42197378 5478804.02538396...)))	Zolotoniskyi	39031	39339	218	824	441
7121200000	MULTIPOLYGON (((6380005.8326648 5402075.15493481...)))	Zvenyhorodskyi	41845	41887	282	804	503
7120600000	MULTIPOLYGON (((63356005.8326648 5412565.15493481...)))	Drabivskyi	32602	32806	154	717	269
7120900000	MULTIPOLYGON (((63569775.8326648 5434895.15493481...)))	Zhashkivskyi	34831	34726	235	756	430

The result of the equi-join in the form of a thematic map of the districts of the Cherkasy region by the attribute of the permanent population is in Fig. 5. The equi-join has a disadvantage: if it occurs on attributes with the same name, then the resulting relation appears in two attributes with the same values. To get rid of this disadvantage by taking a projection on all attributes except one of the duplicates or using the operation of natural join.

Semi-Join. The semi-join operation defines a relation that contains those tuples of relation A that are part of the join of relations A and B. The advantage of a semi-join is that it reduces the number of tuples that need to be processed to obtain a join. [Connolly & Begg, 2003].

The query using a semi-join operation must contain a condition. For example, the value of the "permanent population" attribute of relation B must be equal to the value of the "population" attribute of relation A and the "number of arrivals" attribute of relation B must be greater than 350:

SELECT COATUU, geom, name_ua, existing population, permanent population, number of births, number of deaths, number of arrivals

FROM public.cherkasy

INNER JOIN public.stat

ON cherkasy.COATUU= stat.COATUU

WHERE stat.permanent population = population and number of arrivals > 350

Table 6

The result of the full join

COATUU	COATUU	geom	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
1	2	3	4	5	6	7	8	9
7121500000	7121500000	MULTIPOLYGON (((6392148.42197378 5478804.02538396...)))	Zolotoniskyi	39031	39339	218	824	441

Cont. table 6

1	2	3	4	5	6	7	8	9
7121200000	7121200000	MULTIPOLYGON (((6380005.8326648 5402075.15493481...)))	Zvenyhoro-dskyi	41845	41887	282	804	503
7120600000	7120600000	MULTIPOLYGON (((63356005.8326648 5412565.15493481...)))	Drabivskyi	32602	32806	154	717	269
7120900000	7120900000	MULTIPOLYGON (((63569775.8326648 5434895.15493481...)))	Zhashkivskyi	34831	34726	235	756	430
NULL	NULL	MULTIPOLYGON (((6386787.50122662 5493188.28142229...)))	Cherkaskyi	NULL	NULL	NULL	NULL	NULL
NULL	NULL	MULTIPOLYGON (((6450505.70254616 5545812.54783928...)))	Umanskyi	NULL	NULL	NULL	NULL	NULL

Table 7

The result of the θ-Join

COATUU	geom	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7121500000	MULTIPOLYGON (((6392148.42197378 5478804.02538396...)))	Zolotoniskyi	39031	39339	218	824	441
7121200000	MULTIPOLYGON (((6380005.8326648 5402075.15493481...)))	Zvenyhorodskyi	41845	41887	282	804	503
7120600000	MULTIPOLYGON (((63356005.8326648 5412565.15493481...)))	Drabivskyi	32602	32806	154	717	269
7120900000	MULTIPOLYGON (((63569775.8326648 5434895.15493481...)))	Zhashkivskyi	34831	34726	235	756	430

Table 8

The result of the Equi-Join

COATUU	geometry	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7121500000	MULTIPOLYGON (((6392148.42197378 5478804.02538396...)))	Zolotoniskyi	39031	39339	218	824	441
7121200000	MULTIPOLYGON (((6380005.8326648 5402075.15493481...)))	Zvenyhorodskyi	41845	41887	282	804	503
7120600000	MULTIPOLYGON (((63356005.8326648 5412565.15493481...)))	Drabivskyi	32602	32806	154	717	269
7120900000	MULTIPOLYGON (((63569775.8326648 5434895.15493481...)))	Zhashkivskyi	34831	34726	235	756	430

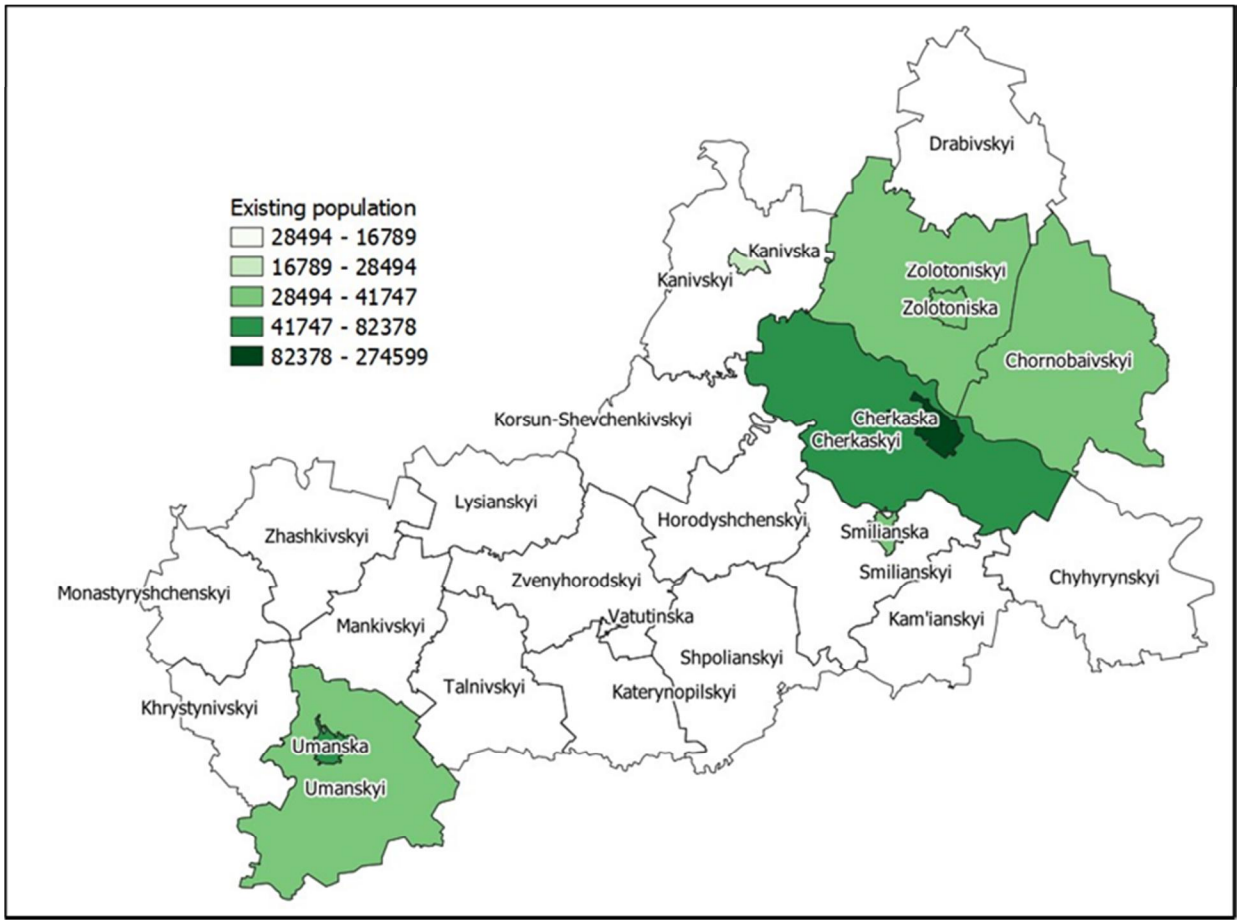

Fig. 4. The resulting thematic map by θ-Join

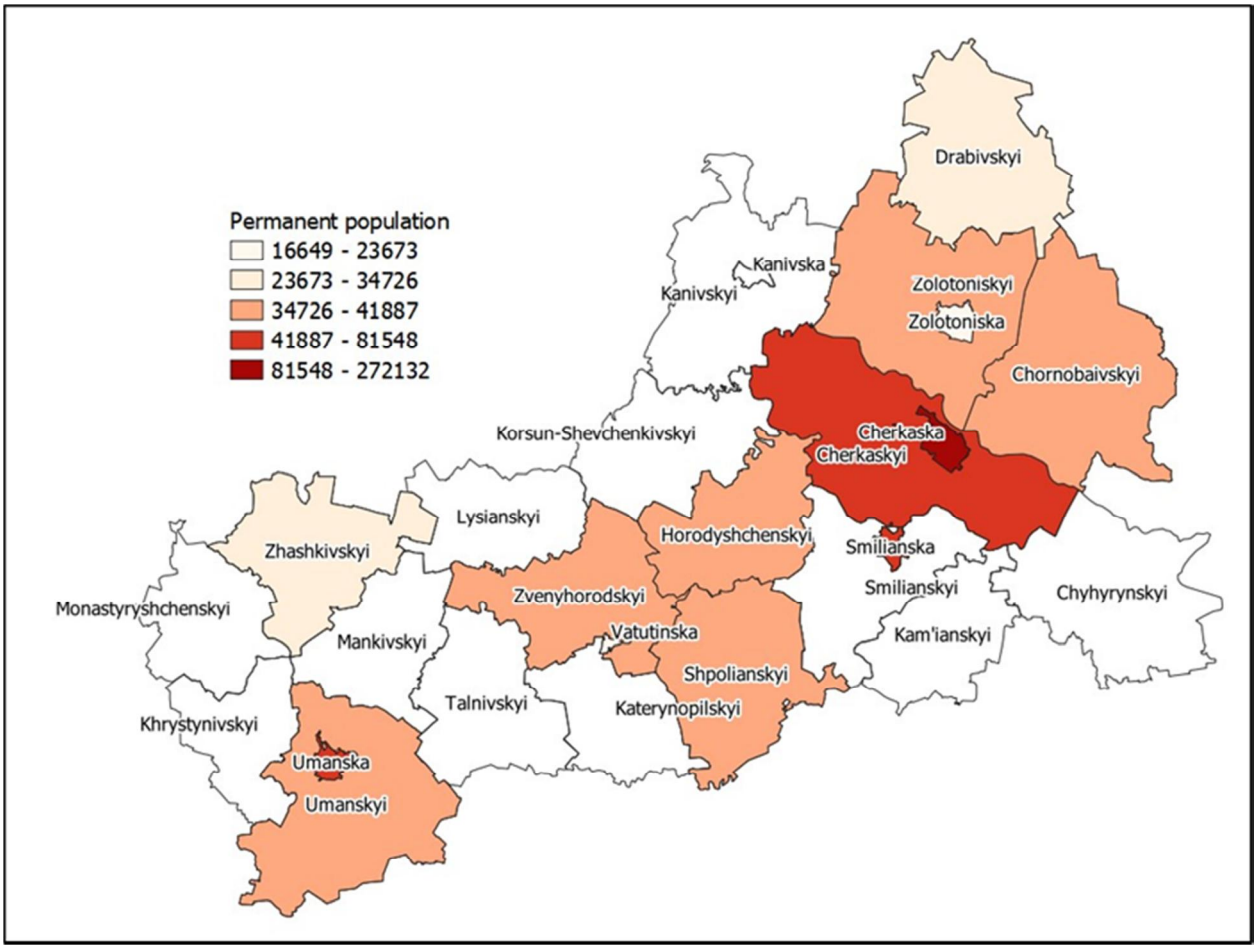

Fig. 5. The resulting thematic map by Equi-Join

Table 9

The result of the Semi-Join

COATUU	COATUU	geometry	name_ua	existing population	permanent population	number of births	number of deaths	number of arrivals
7121500000	7121500000	MULTI POLYGON (((6392148.42197378 5478804.02538396...)))	Zoloto-niskyi	39031	39339	218	824	441
7121200000	7121200000	MULTI POLYGON (((6380005.8326648 5402075.15493481...)))	Zveny-horodskyi	41845	41887	282	804	503
7120900000	7120900000	MULTI POLYGON (((63569775.8326648 5434895.15493481...)))	Zhash-kivskyi	34831	34726	235	756	430

It was found as a result of performing different types of connections of relations A and B that each feature that connects must have a unique identifier for their successful implementation. Besides each type of join has its own peculiarities: the usage of conditions, the order of join, and, accordingly, the results of queries.

Scientific novelty and practical significance

The integration of core reference data and thematic geospatial datasets based on JOIN operation models of relational algebra and their interaction with geocoding of geospatial features is researched, which is implemented in modern GIS and DBMS for the development of national spatial data infrastructure. The research was performed on a set of core reference spatial data, namely: information on administrative-territorial units of the Cherkasy region, including their borders; the data from the statistical bulletin of the socio-economic situation of the Cherkasy region for January 2021 of the Main Department of Statistics in Cherkasy region of the State Statistics Service of Ukraine were selected as thematic data. It has been shown that relational algebra join (JOIN) operations can be used to integrate other thematic geospatial data with core reference data using geographic identifiers that contain these datasets.

Conclusions

The adoption of the Law of Ukraine "On the National Spatial Data Infrastructure of Ukraine», No. 554 on April 13, 2020, the Decree of the Cabinet of Ministers of Ukraine "On Approving the Procedure for the Functioning of the National Spatial Data Infrastructure" No. 532 was adopted on May 26, 2021, the Ministry of Agrarian Policy and Food of Ukraine approved the Order "On approval of technical requirements for geospatial data, metadata and geoinformation services of the National Spatial Data Infrastructure" No. 347 of 10.11.2021 promoted to the development of the National Spatial Data Infrastructure in Ukraine.

The unification and integration of core reference and thematic geospatial data ensure the work of the NSDI. One way to integrate data is to join different geospatial databases with data that do not contain a geospatial component, using a system identifier to compile them from different sources in the NSDI.

Geocoding occupies an important place as a spatial reference of features based on geographical identifiers among all methods of data integration. The main requirements which are in the international and harmonized national standard DSTU ISO 19112:2017 (ISO 19112:2003, IDT) Geographic information – Spatial referencing by geographic identifiers.

The paper examines the models of the Join operation of relational algebra, which underlie the geocoding of features and the creation of electronic gazetteers, and proves its effectiveness: the Join operation integrates of core reference and thematic geospatial datasets. There is a need to define the required geographic identifiers, which must be present among the attributes of the core reference and thematic geospatial datasets to perform the join. The variety of uses of the Join operation covers all possible cases that arise in their practical application. Thus, the use of the Join operation involves identifying these required geographic identifiers at the geospatial database design stage.

In particular, it is expedient to determine mandatory geographical identifiers (codes) of features according to the official national systems of features classification (codification) in the relevant sectoral thematic registers:

1) the COATUU (now – CATUTTC) for the features of the administrative and territorial structure of Ukraine;
2) the identifiers of highways, bridges, crossings, and features of railway according to the registers of the Ministry of Infrastructure of Ukraine, Ukrzaliznytsia, State Agency of Highways of Ukraine (Ukravtodor), and others.
3) the codes of rivers, reservoirs, and drains – according to the State Water Cadastre Classifier;
4) the codes of high voltage power grids – according to the registers of the Ministry of Energy of Ukraine;
5) the codes for forests and vegetation – according to the register of the State Forest Cadastre;
6) the identifiers of buildings and structures – according to the registers of real estate, the registers of the Bureau of Technical Inventory (BTI), and the Urban Cadastre of the Ministry for Communities and Territories Development of Ukraine;

7) the identifiers of pipelines – according to the registers of the Ministry of Energy of Ukraine, National Joint Stock Company "NAFTOGAZ OF UKRAINE" etc.

The usage of the Join operation will reduce the cost of supporting the core reference data attributes of mandatory permanent storage, will ensure the data integration from different information resources, and, thus the creation of different thematic geospatial data based on a set of information available from different state registers and databases. The integration of data based on their Join should ensure the integrity, reliability, informativeness, and thematic diversity of geospatial data due to their interoperability, and compatibility with the primary sources of industry information resources.

REFERENCES

Bhattacharya, D., & Painho, M. (2017). Smart cities intelligence system (smacisys) integrating sensor web with spatial data infrastructures (sensdi). ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 4, 21-28. https://run.unl.pt/handle/10362/28046.

Brandeis Library. Geographic Information Systems (GIS). URL: https://guides.library.brandeis.edu/gis/geocoding (дата звернення: 20.04.2022)..

Bui, D. & Glushko, I. (2015). Expansion of the signature of Codd's relational (table) algebras: current state *NaUKMA Research Papers. Computer Science* (177), 95-107. (in Ukrainian).

Classifier of objects of administrative-territorial organization of Ukraine. (in Ukrainian). http://www.ukrstat.gov.ua/klasf/st_kls/op_koatuu_2016.htm.

Codd, E. F. (2002). A relational model of data for large shared data banks. In Software pioneers (pp. 263-294). Springer, Berlin, Heidelberg.

Codd, E. F. (1990). The relational model for database management: version 2. Addison-Wesley Longman Publishing Co., Inc.

Connolly, Thomas, & Caroline Begg (2003). Database. Design, implementation and maintenance. Theory and Practice. Moscow: Williams, 2003. 1436 p., 145-149.

Decree to the Cabinet of Ministers of Ukraine "On Approval of the Procedure for the Functioning of the National Geospatial Data Infrastructure" No. 532 dated May 26, 2021. https://zakon.rada.gov.ua/laws/show/532-2021-%D0%BF#Text.

ESRI's Geodatabase Website. URL: http://support.esri.com/datamodels (дата звернення: 20.04.2022).

ESRI's Download Website. URL: http://www.esri.com/data/download/census2000_tigerline/index.html (дата звернення: 20.04.2022).

Franci, F., Lambertini, A., & Bitelli, G. (2014, August). Integration of different geospatial data in urban areas: a case of study. In *Second International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2014)* (Vol. 9229, p. 92290P). International Society for Optics and Photonics. https://doi.org/10.1117/12.2066614.

Gao, D., Jensen, C. S., Snodgrass, R. T., & Soo, M. D. (2005). Join operations in temporal databases. The VLDB Journal, 14(1), 2-29. https://link.springer.com/article/10.1007/s00778-003-0111-3.

Geocoding: Longitude and Latitude by Address. URL: https://gisgeography.com/geocoding/ (дата звернення: 20.04.2022).

Geographic information. Spatial referencing by geographic identifiers] (2017). DSTU ISO 19112-2017(ISO 19112:2003, IDT) from 1d October 2019. Kyiv. DP "UkrNDNTs" (in Ukrainian).

Geoportal "Administrative and territorial organization of Ukraine". (in Ukrainian). http://atu.gki.com.ua/

Glushko, I. (2013). *Calculation and extension of signatures of tabular algebras*. (PhD dissertation). Available from Taras Shevchenko National University of Kyiv. URL: (in Ukrainian). http://csc.knu.ua/uk/library/dissertations/hlushko.pdf.

Hansen, H. S. (1999, April). Integrating digital maps and administrative registers-Danish experiences. In 21 st Urban Data Management Symposium (pp. 21-23).

How to Geocode in ArcMap. URL: https://libraries.mit.edu/files/gis/geocoding.pdf (дата звернення: 20.04.2022).

Karpinsky, Y., & Lazorenko-Hevel, N. (2018). The methods of geospatial data collection for topographic mapping. *Modern achievements of geodesic science and industry*. (in Ukrainian). http://gki.com.ua/ua/metodi-zbirannja-geoprostorovih-danih-dlja-topografichnogo-kartografuvannja.

Karpinskyi, Y., Lazorenko-Hevel, N. (2020). The system model of topographic mapping in the national spatial data infrastructure in Ukraine. *Geodesy, Cartography and Aerial Photography*, 92, 24–36. https://doi.org/10.23939/istcgcap2020.92.024.

Karpinskyi, Y., & Lazorenko-Hevel, N. (2020). Topographic mapping in the National Spatial Data Infrastructure in Ukraine. In E3S Web of Conferences (Vol. 171, p. 02004). EDP Sciences. https://doi.org/10.1051/e3sconf/202017102004.

Karpinskyi Y., Lazorenko-Hevel N., Kin D. (2020). INSPIREID implementation in the topographic database of the main state topographic

map of Ukraine. Веб ISTCGCAP, 91, 20–27. https://doi.org/10.23939/istcgcap2020.91.020.

Karpinskyi, Y. & Lyashchenko A. (2006). Strategia formuvannia natsionalnoi infrastruktury geoprostorovych danych v Ukraini, (108 p.). Kyiv: NDIGK. (Ser. "Geodesy, cartography, cadastre") (in Ukrainian).

Law of Ukraine About National Geospatial Data Infrastructure from April 13 2020, No. 554-IX (2020). Vidomosti Verkhovnoi Rady Ukrainy. Bulletin of Verkhovna Rada of Ukraine. (in Ukrainian).

Lazorenko-Hevel N. (2021). Geographic identifiers as a basis for integration of geospatial data. *Mistobuduvannya ta terytorial'ne planuvannya*, (78), 312-326. (in Ukrainian). https://doi.org/10.32347/2076-815x.2021.78.312-326.

Lemenkova, P. (2020). Integration of geospatial data for mapping variation of sediment thickness in the North Sea. *Scientific Annals of the Danube Delta Institute*, *25*, 129-138. https://doi.org/10.7427/DDI.25.14.

Lyashchenko, A. & Cherin, A. (2019). Basic models and methods of geospatial data integration in GIS of urban-planning cadastre. *Mistobuduvannya ta terytorial'ne planuvannya*, (70), 351-365. (in Ukrainian). http://repositary.knuba.edu.ua//handle/987654321/6199.

Lyashchenko, A., Havryliuk, Y., & Smilka, V. (2020). Analysis of methods of unique identification of objects in geospatial data sets. *Mistobuduvannya ta terytorial'ne planuvannya*, (75), 217-232. http://repositary.knuba.edu.ua//handle/987654321/9512.

Lyashchenko, A., Karpinskyi Y., Havryliuk, Y. & Cherin, A. (2021). Methods and means of ensuring the interoperability of the components of the national geospatial data infrastructure. *Mistobuduvannya ta terytorial'ne planuvannya*, (77), 309-319. (in Ukrainian). https://doi.org/10.32347/2076-815x.2021.77.309-319.

Mardani, M., Mardani, H., De Simone, L., Varas, S., Kita, N., & Saito, T. (2019). Integration of machine learning and open access geospatial data for land cover mapping. *Remote Sensing*, *11*(16), 1907. https://doi.org/10.3390/rs11161907.

Maksymova Y. (2016). Creating a database of electronic catalog of object classes for sets of profile geospatial data of urban planning documentation. *Mistobuduvannya ta terytorial'ne planuvannya*, (62 (1)), 367-376. (in Ukrainian). https://repositary.knuba.edu.ua/bitstream/handle/987654321/6932/62a-368-377.pdf?sequence=1.

Order of the Ministry of Agrarian Policy and Food of Ukraine "On approval of technical requirements for geospatial data, metadata and geoinformation services of the national geospatial data infrastructure" from10.11.2021 No. 345. (in Ukrainian).

Pilicheva, M., Kin, D., & Pomortseva, O. (2018). Integration of topographical and cadastral data of the basic dataset of a land parcel. *Mistobuduvannya ta terytorial'ne planuvannya*, (66), 523-531. (in Ukrainian).

Resolution of the Cabinet of Ministers of Ukraine "On approval of the Order for national topographic and thematic mapping" from 04.09.2013 No. 661. (in Ukrainian).

Rayordan, R. (2001) *Relational database fundamentals*. M.: Publishing house "Russian edition".

Shypulin, V. (2021). Integrated real estate information system. Concept for Ukraine: monograph. O. M. Beketov National University of Urban Economy in Kharkiv. (in Ukrainian). http://eprints.kname.edu.ua/57436/.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). Database system concepts (Vol. 5). New York: McGraw-Hill. 1376 p. https://snscourseware.org/snsctnew/files/1581236100.pdf

Stankevich, S., Titarenko, O., & Golubov, S. (2021). Mathematical model of integration of heterogeneous data in assessing the oil and gas prospects of territories. *Kherson –2021*, 86. (in Ukrainian).

Statistical bulletin "Socio-economic situation of Cherkasy region". Main Department of Statistics in Cherkasy Oblast. (2021). (in Ukrainian). http://www.ck.ukrstat.gov.ua/?p=bul_soc_ek.

Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W., & Song, J. (2019). Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data*, *3*(3), 269-296. https://doi.org/10.1080/20964471.2019.1661662.

The National Standard of Ukraine DSTU 8774:2018 "Geographic information. Geospatial data modeling rules". (in Ukrainian). http://gki.com.ua/ua/prinjatonacionalni-standart-ukraiini-dstu-87742018-geografichna-informacija-pravila-modeljuvannjageoprostorovih-danih.

Надія ЛАЗОРЕНКО

Кафедра геоінформатики і фотограмметрії, Київський національний університет будівництва і архітектури, Повітрофлотський проспект, 31, Київ, 03037, Україна, ел. пошта: nadiialg@gmail.com, https://orcid.org/0000-0002-1572-49471

ІНТЕГРАЦІЯ ГЕОПРОСТОРОВИХ ДАНИХ НА ОСНОВІ ЗАСТОСУВАННЯ ОПЕРАЦІЇ З'ЄДНАННЯ (JOIN) РЕЛЯЦІЙНОЇ АЛГЕБРИ

Мета цієї роботи – дослідження інтеграції наборів базових і тематичних геопросторових даних на основі операції з'єднання (JOIN) реляційної алгебри та її взаємодії з геокодуванням геопросторових об'єктів, яку реалізовано в сучасних геоінформаційних системах (далі – ГІС) та системах керування базами даних (далі – СКБД) для розвитку національної інфраструктури геопросторових даних (далі – НІГД). Методика. Основою дослідження є аналіз можливостей застосування теорії баз геопросторових даних і баз знань, міжнародних і національних гармонізованих стандартів у сфері Географічна інформація/Геоматика для вирішення питання інтеграції геопросторових даних за допомогою операції з'єднання JOIN реляційної алгебри в об'єктно-реляційних системах керування базами даних (ОР СКБД). Результати. В статті досліджено моделі операції з'єднання Join реляційної алгебри, які лежать в основі геокодування об'єктів і створення електронних газетирів, і доведено її ефективність: операція з'єднання Join забезпечує інтеграцію наборів базових і тематичних геопросторових даних. Для її виконання необхідно визначити обов'язкові географічні ідентифікатори, які мають бути наявні серед атрибутів наборів базових та тематичних геопросторових даних та за якими виконується з'єднання. Різноманітність видів застосування операції з'єднання Join охоплює всі можливі випадки, які виникають при їх практичному застосуванні. Таким чином, використання операції з'єднання Join передбачає на етапі проектування баз геопросторових даних визначити ці обов'язкові географічні ідентифікатори. Зокрема, доцільним є визначення обов'язкових географічних ідентифікаторів (кодів) об'єктів за офіційними загальнодержавними системами класифікації (кодифікації) об'єктів у відповідних галузевих тематичних реєстрах, за які відповідають визначені держателі тематичних даних відповідно до додатку 2 Постанови Кабінету Міністрів України "Про затвердження Порядку функціонування національної інфраструктури геопросторових даних" від 26 травня 2021 р. № 532. Досліджено інтеграцію наборів базових і тематичних геопросторових даних на основі моделей операції з'єднання (JOIN) реляційної алгебри та їх взаємодії з геокодуванням геопросторових об'єктів, яку реалізовано в сучасних ГІС та СКБД для розвитку національної інфраструктури геопросторових даних. Дослідження виконано на наборі базових геопросторових даних, а саме: відомостей про адміністративно-територіальні одиниці Черкаської області, в тому числі їх меж; за тематичні обрано дані зі статистичного бюлетеня соціально-економічного становища Черкаської області за січень 2021 року Головного управління статистики у Черкаській області Державної служби статистики України. Доведено, що операцію з'єднання (JOIN) реляційної алгебри можна використовувати для інтеграції інших тематичних геопросторових даних з базовими геопросторовими даними за допомогою географічних ідентифікаторів, які містять ці набори даних.

Ключові слова: інтеграція геопросторових даних, інтероперабельність, операція з'єднання (JOIN), національна інфраструктура геопросторових даних, базові геопросторові дані, тематичні геопросторові дані.

Received 05.04.2022