НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ МАТЕМАТИЧНИХ МАШИН І СИСТЕМ

ЧЕРТОВ ОЛЕГ РОМАНОВИЧ

УДК 004.9:004.6:314.1

МОДЕЛІ, ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА АРХІТЕКТУРА СИСТЕМ ОБРОБКИ ДЕМОГРАФІЧНОЇ ІНФОРМАЦІЇ

05.13.06 – інформаційні технології

Автореферат дисертації на здобуття наукового ступеня доктора технічних наук

Київ – 2013

Дисертацією є рукопис.

Робота виконана у Національному технічному університеті України «Київський політехнічний інститут» Міністерства освіти і науки України.

Науковий консультант доктор технічних наук, професор
Молчанов Олександр Артемович,
Національний технічний університет України
«Київський політехнічний інститут» МОН України,
завідувач кафедри прикладної математики

Офіційні опоненти: доктор технічних наук, професор
Литвинов Віталій Васильович,
Чернігівський державний технологічний університет
МОН України, завідувач кафедри програмної інженерії

доктор технічних наук, старший науковий співробітник
Ланде Дмитро Володимирович,
Інститут проблем реєстрації інформації НАН України,
завідувач відділу спеціалізованих засобів моделювання

доктор технічних наук, доцент
Танянський Сергій Станіславович,
Харківський національний університет радіоелектроніки
МОН України, професор кафедри електронних
обчислювальних машин

Захист відбудеться «5» березня 2014 р. о 14 годині на засіданні спеціалізованої вченої ради Д 26.204.01 в Інституті проблем математичних машин і систем НАН України за адресою: 03187, м. Київ, проспект Академіка Глушкова, 42.

З дисертацією можна ознайомитись у бібліотеці Інституту проблем математичних машин і систем НАН України за адресою: 03187, м. Київ, проспект Академіка Глушкова, 42.

Автореферат розісланий «_ _» _____ 2014 р.

Учений секретар
спеціалізованої вченої ради В. І. Ходак

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Під демографічною інформацією будемо розуміти відомості про чисельність населення, його структуру, розміщення та динаміку змін. Структура населення досліджується за демографічними ознаками, такими як стать, вік, шлюбний стан, а також за соціальними ознаками – громадянство, етнічне походження, рідна мова, освіта тощо.

За даними Державної служби статистики України (Держстату України), за роки незалежності чисельність населення країни зменшилася на 6,3 млн осіб, кожен десятий українець не доживає до 35 років, а кожен четвертий – до 60. У зв'язку зі старінням і депопуляцією (процесом систематичного скорочення абсолютної чисельності) населення України та необхідністю вироблення адекватних заходів для покращення демографічної ситуації в країні зростає значення наявності точної та актуальної демографічної інформації.

Основним джерелом такої інформації є перепис населення, а в міжпереписний період – вибіркові дослідження та результати оцінювання населення, наприклад розрахунки за даними поточного адміністративного обліку процесів природного та механічного руху населення. Деякі країни, наприклад скандинавські, використовують також загальний демографічний регістр населення.

Зазначені демографічні спостереження характеризуються двома рисами:

– масштабністю, наприклад, перепис населення охоплює за короткий період часу всіх мешканців країни;

– застосуванням складних методів, як правило, розподіленої обробки даних, включаючи правила перевірки їх несуперечності та повноти.

За влучним зауваженням М. Вуда (Monty Wood), прес-секретаря Бюро перепису населення США, «перепис – це найзатратніша операція, яку країна проводить у мирний час», тому створення систем обробки демографічної інформації є важливим завданням національного масштабу.

В Україні ще з радянських часів сформувалися кілька знаних наукових шкіл з розроблення складних територіально-розподілених інформаційно-аналітичних систем, наприклад, в Інституті проблем математичних машин та систем НАН України (А. О. Морозов, В. В. Литвинов), Інституті проблем реєстрації інформації НАН України (О. Г. Додонов). Але що стосується статистичної галузі, то всі значущі програмно-технологічні комплекси оброблення даних створювалися в Центральному статистичному управлінні в Москві. Тому, коли постала задача підготовки автоматизованої системи для обробки даних і матеріалів першого в незалежній Україні перепису населення – перепису 2001 р., фактично заново довелося розроблювати архітектурні рішення та інформаційні технології, необхідні для реалізації зазначеної системи.

Оскільки демографічна інформація отримується, зазвичай, шляхом опитування респондентів, то їй природно притаманна певна недостовірність і суперечність в даних, викликана різноманітними об'єктивними та суб'єктивними

причинами. Наприклад, під час опитування респондент навмисно чи ненавмисно може надати неправдиву чи неточну інформацію, обліковець, записуючи дані, може помилитися, не дочувши правильну відповідь чи не зрозумівши її, кодувальник при обробці переписних форм може механічно помилитися під час присвоєння коду відповідній одиниці інформації, сканувальне програмне забезпечення може неправильно інтерпретувати записані символи (через нерозбірливість почерку обліковця чи неакуратне вписування тексту в комірки для зчитування) тощо. Істотно підвищити достовірність демографічних даних можна лише за рахунок дублювання їх збору та введення, що в силу великих обсягів цих даних на практиці не можливо. Однак для коректного формування вихідної інформації (вихідних таблиць, результатів виконання аналітичних запитів) актуальною залишається проблема забезпечення несуперечності демографічних даних.

З 60-х років XX ст. за рекомендаціями ООН абсолютна більшість країн світу проводить переписи населення циклічно – раундами один раз на 10 років. У межах поточного раунду в грудні 2010 р. у Дергачівському районі Харківської області було проведено пробний перепис населення, надалі планується проведення Всеукраїнських перепису населення та сільськогосподарського перепису.

На кінець 2013 р. у статистичній галузі України заплановано впровадження Інтегрованої системи обробки статистичних даних, створення якої було розпочато в рамках проекту розвитку системи державної статистики України для моніторингу соціально-економічних перетворень за підтримки Міжнародного банку реконструкції та розвитку і спрямовано на заміну більш ніж 150 комплексів електронної обробки інформації, які наразі експлуатуються в Держстаті України.

У зв’язку з цим особливо актуальними стають питання вироблення єдиних методологічних підходів до побудови архітектури зазначених та інших аналогічних систем, в основі яких лежить обробка демографічної інформації.

Підвалини сучасного підходу до розроблення інформаційних систем у статистичній галузі були закладені шведським ученим Б. Сунгреном (Bo Sundgren) у 1999 р. Наразі загальновизнані лідери в цій сфері – науково-інженерні центри в США, Канаді, Франції, Новій Зеландії, Латвії.

Аналіз існуючих інформаційних систем у галузі статистики населення показує, що, забезпечуючи основну функціональність зі збереження та обробки демографічних даних, ці системи додатково дають користувачам можливість будувати аналітичні звіти. Однак пошук прихованих закономірностей, що є в аналізованих даних, виконується фактично лише вручну через висунення певних гіпотез і перевірку їх справедливості за допомогою побудови аналітичних звітів чи, взагалі, здійснюється в межах інших відокремлених систем. Тому робота з вироблення єдиних методологічних та технологічних засад побудови інформаційних систем обробки демографічних даних, розширення можливостей таких систем за рахунок впровадження методів інтелектуального аналізу даних, застосування відповідних моделей метаданих, ліквідації так званої «клаптикової» автоматизації є актуальною та відповідає нагальним потребам ефективного запровадження інформаційних технологій у статистичній галузі.

Останніми роками значного поширення набуває спосіб надання результатів різноманітних статистичних спостережень через мікрофайли – певні вибірки первинних даних про респондентів. Достатньо згадати найбільш масштабний з таких проектів IPUMS-International, у межах якого вже зібрано та відкрито для доступу дослідникам 544 млн персональних записів з 238 переписів 74 країн. Але вилучення в подібних мікрофайлах атрибутів, що однозначно ідентифікують респондента, не гарантує знеособленості інформації в опублікованих даних. Тому також актуальними є дослідження, які спрямовані на підвищення захищеності демографічної інформації шляхом розроблення методів забезпечення індивідуальної та групової анонімності даних про респондентів.

Таким чином, вироблення єдиних методологічних та технологічних засад побудови систем обробки демографічної інформації, забезпечення несуперечності демографічної інформації, підвищення її захищеності шляхом розробки нових моделей, методів, інформаційних технологій та архітектурних рішень для відповідних систем є актуальним напрямком і потребує подальшого дослідження, що й визначило тему даної роботи.

Саме вирішення зазначеної науково-прикладної проблеми і становить сутність цієї дисертаційної роботи.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у межах пріоритетних наукових напрямів «Інформаційні та комунікаційні технології» і «Розвиток сучасних інформаційних, комунікаційних технологій», визначених відповідно законами України «Про пріоритетні напрями розвитку науки і техніки» (№ 2623-III від 11.07.2001 р. у редакції від 05.12.2012 р.) та «Про пріоритетні напрями інноваційної діяльності в Україні» (№ 3715-VI від 08.09.2011 р. у редакції від 05.12.2012 р.), у межах розв’язання проблеми «Розробка теоретичних основ і прикладних методів створення комп’ютерних інформаційно-аналітичних систем, дослідження та розробка методів захисту інформації в комп’ютерних системах і мережах. Методи та системи підтримки прийняття рішень», визначеної спільним Наказом Міністерства освіти і науки України та Національної академії наук України «Про затвердження Основних наукових напрямів та найважливіших проблем фундаментальних досліджень у галузі природничих, технічних і гуманітарних наук на 2009-2013 роки» (№ 1066/609 від 26.11.2009 р.), а також згідно з планами наукових робіт кафедри прикладної математики Національного технічного університету України «Київський політехнічний інститут».

Дисертація вміщує результати досліджень і розробок, здійснених автором під час виконання таких договорів:

1) договір № СН01 від 27.03.2002 між Головним міжрегіональним управлінням статистики (ГМУС) у місті Києві та спільним підприємством «Квазар-мікро ІВМ» на виконання послуг з розробки та впровадження автоматизованої системи (АС) «Перепис-2001» (1-ша черга);

2) договір № СН02 від 4.07.2003 між ГМУС у місті Києві та спільним підприємством «Квазар-мікро ІВМ» на виконання послуг з розробки та впровадження АС «Перепис-2001» (2-га та 3-тя черги);

3) договір № СН03 від 24.03.2004 між ГМУС у місті Києві та АТЗТ «Квазар-Мікро Техно» на виконання робіт з побудови сховища даних, розробки та введення в дію системи оперативного аналізу даних Всеукраїнського перепису населення 2001 року у багатовимірному вигляді (АС «Перепис-2001 Аналітик») і підтримки автоматизованої системи обробки даних Всеукраїнського перепису населення 2001 року (АС «Перепис-2001»);

4) договір № Т512Т2945 від 28.07.2004 між Департаментом статистики і соціології Республіки Молдова та АТЗТ «Квазар-Мікро Техно» на виконання робіт з розробки програмної продукції, сканування і обробки даних перепису населення 2004 р. відповідно до потреб Департаменту статистики і соціології;

5) договір № 116 від 16.10.2008 між Державним комітетом статистики України і АТЗТ «Квазар-Мікро Техно» на надання послуг зі створення єдиної системи оброблення даних із статистики населення (за предметом закупівлі «73.20.1. Послуги з досліджень і розробок у галузі суспільних і гуманітарних наук» по КПКВК 3001020 «Статистичні спостереження та переписи», КЕК 1139);

6) договір № 4729/43 від 1.07.2009 між Державним комітетом статистики України та фізичною особою Чертовим О. Р. на надання консультантських послуг з нагляду за розробленням і впровадженням Інтегрованої системи обробки статистичних даних;

7) договір № 81 від 16.11.2010 між Державним комітетом статистики України і дочірнім підприємством «Квазар Хорте» ТОВ «Юридично-консалтингова фірма «РЕСПОНСА КА» на виконання науково-дослідної роботи з розробки технічного проекту на створення комплексу електронної обробки даних Всеукраїнського перепису населення 2012 року (за предметом закупівлі «73.20. Дослідження та розробки у галузі суспільних і гуманітарних наук» по КПКВК 3001020), державний реєстраційний номер 0110U006826;

8) договір № 120 від 18.11.2011 між Державним комітетом статистики України і дочірнім підприємством «Квазар Хорте» ТОВ «Юридично-консалтингова фірма «РЕСПОНСА КА» на виконання робіт за предметом закупівлі «72.20.2. Послуги з розроблення пакетних програмних засобів (створення програмного забезпечення «Комплекс електронної обробки даних Всеукраїнського перепису населення 2012 року», 1-ша черга)»;

9) договір № SSA 2012/002 від 14.05.2012 між Фондом ООН з питань народонаселення та індивідуальним підрядником Чертовим О. Р. на надання консультантських послуг з підготовки технічної документації для проектів з розроблення і впровадження інформаційних систем для обробки даних перепису населення та сільськогосподарського перепису;

10) договір № 639719 від 24.09.2012 між Університетом Міннесоти, США, та індивідуальним підрядником Чертовим О. Р. на підготовку 10 % мікроданих за результатами Всеукраїнського перепису населення 2001 р.

У роботах, що виконувалися згідно з переліченими вище першими п’яти договорами, автор брав участь як головний аналітик та керівник проекту з боку підприємства – виконавця робіт, у сьомому та восьмому проектах – як головний аналітик, а в усіх інших – як незалежний експерт чи консультант.

Мета і задачі дослідження. Метою дисертаційної роботи є вироблення єдиних методологічних та технологічних засад побудови інформаційних систем з накопичування, обробки та збереження демографічної інформації, забезпечення несуперечності демографічної інформації, підвищення її захищеності.

Для досягнення зазначеної мети розв’язувались такі задачі:

– визначити основні особливості систем обробки демографічної інформації, зокрема, систем обробки даних перепису населення, проаналізувати та систематизувати тенденції їх розвитку;

– розробити функціонально-інформаційну архітектуру статистичної галузі України з урахуванням необхідності реалізації пошуку прихованих закономірностей у статистичних даних і впровадження інтелектуальних інформаційних технологій;

– удосконалити стандартну загальну модель статистичних бізнес-процесів і побудувати відповідну модель метаданих для опису демографічної інформації з урахуванням практики проведення статистичних спостережень в Україні, що склалася за роки незалежності, і явного визначення процесів та інформаційних технологій, які забезпечують анонімність статистичної інформації;

– розробити й апробувати метод забезпечення групової анонімності даних у мікрофайлах на основі застосування вейвлет-перетворення, підготувати загальну методику забезпечення групової анонімності демографічних даних;

– розробити інформаційну технологію обробки демографічних даних для встановлення наявності та локалізації помилок у документах, що містять первинні дані;

– розробити інформаційну технологію обробки демографічних даних для локалізації помилок у зведених даних і пошуку причин їх виникнення.

Об’єктом дослідження є процеси накопичування, обробки та збереження демографічної інформації.

Предметом дослідження є моделі, інформаційні технології та архітектурні рішення, які забезпечують організацію обробки демографічної інформації.

Методи дослідження. Для вирішення завдань з побудови функціонально-інформаційної архітектури статистичної галузі України, вироблення інформаційних технологій для локалізації та пошуку помилок у первинних і зведених даних у системах обробки демографічної інформації як вихідна теоретична база використовувалися основні положення й методи загальної теорії систем та теорії адаптивних систем, теорії корпоративних архітектур, інтелектуального аналізу даних. Крім того, використовувалися відомі методи й способи інженерного проектування автоматизованих баз і сховищ даних та інформаційних систем. При конкретизації загальної моделі статистичних бізнес-процесів, побудові моделі метаданих для опису демографічної інформації додатково до зазначених методів дослідження використовувалися теорія моделювання процесів й апарат теорії

подання метаданих. Нарешті, при розробленні та реалізації методу забезпечення групової анонімності даних у мікрофайлах використовувалися теорія вейвлет-перетворень, теорія матриць, лінійне програмування, теорія рядів і цифрова фільтрація сигналів.

Наукова новизна одержаних результатів складається з таких положень:

– уперше введено і обґрунтовано поняття групової анонімності даних, яке на відміну від існуючих підходів із знеособлення даних про окремих респондентів дає змогу поставити нові класи задач приховування інформації про особливості розподілу групи респондентів і формалізувати їх розв'язок;

– уперше розроблено метод забезпечення групової анонімності даних, що передбачає цілеспрямовану зміну розподілу записів вихідного набору даних і який на відміну від існуючих методів оперує зі складовими вейвлет-розкладення цього розподілу, що дає можливість підвищити захищеність даних та зберегти їх корисність;

– розроблено нову інформаційну технологію встановлення наявності і локалізації помилок у документах з первинними даними, яка за рахунок створення інтегрованого середовища, що містить поточний стан документа, його графічний образ і, на відміну від існуючих технологій, протокол із результатами контролю не тільки певної переписної форми, але й їх пов'язаної сукупності, включаючи результати перевірки логічної та арифметичної узгодженості значень показників різних форм, дає змогу підвищити ефективність роботи користувачів з виявлення суперечностей у первинних даних;

– розроблено нову інформаційну технологію локалізації помилок у зведених даних і пошуку причин їх виникнення за допомогою застосування системи формалізованого опису правил перевірки вихідних таблиць і шляхом побудови нерегламентних запитів, що фільтрують і структурують демографічні дані, реалізація в запропонованій технології адаптивної підтримки співробітництва користувачів дає змогу спростити їм обмін досвідом із побудови відповідних нерегламентних запитів;

– удосконалено модель Захмана функціонально-інформаційної архітектури за рахунок її розширення рівнем системної архітектури, який відповідає, зокрема, за впровадження інтелектуальних інформаційних технологій, забезпечуючи пошук прихованих закономірностей розподілу даних і тим самим доповнюючи існуючі можливості зі структурованого обліку та аналітичної обробки демографічних даних; ця модель була адаптована для статистичної галузі України, що дає можливість її застосовувати при проектуванні національних систем обробки демографічної інформації;

– удосконалено загальну модель статистичних бізнес-процесів шляхом її територіального розподілу, введення нових процесів, що забезпечують знеособлення даних про респондентів, їх індивідуальну і групову анонімність, а також через виокремлення процесу пошуку як в інформаційно-обліковій чи в інформаційно-аналітичній статистичній системі, так і при застосуванні інтелектуальних інформаційних технологій для автоматизованого пошуку в даних

певних закономірностей, що дає змогу забезпечити єдиний методологічний підхід до автоматизації процесів обробки демографічної інформації;

– отримала подальший розвиток модель метаданих Сунгрена для опису статистичної інформації, в якій, на відміну від існуючих, сформована додаткова група метаданих індивідуальної та групової анонімності, що дало змогу враховувати вимоги забезпечення контролю над відкриттям демографічної інформації на всіх етапах її обробки.

Обґрунтованість і достовірність наукових положень і висновків базуються на доведених математичних твердженнях та успішному практичному застосуванні одержаних результатів при проектуванні, розробленні й експлуатації систем обробки демографічної інформації в різних країнах.

Практичне значення одержаних результатів полягає в можливості створення на їх основі територіально-розподілених багатоланкових інформаційних систем з обробки демографічної інформації національного масштабу. Практичні рішення, отримані в процесі дослідницької роботи, дають можливість:

– підвищити ефективність роботи користувачів зазначених систем під час введення, контролювання та подальшої обробки демографічних даних;

– гарантувати анонімність даних про респондентів та їх групи під час поширення результатів статистичних спостережень.

Ефективність розроблених інформаційних технологій та архітектурних рішень підтверджена їх використанням при проектуванні та розробленні інформаційних систем з обробки демографічної інформації, що пройшли успішні випробування і впроваджені в Україні та Молдові.

Теоретичні і практичні результати дисертаційної роботи використані та впроваджені (вих. № 437/0/7-11 від 27.04.2011, вих. № 26-06/12 від 26.06.2012):

– у ГМУС у місті Києві та у відповідних статистичних управліннях 24 областей України, Автономної Республіки Крим та 2 міст центрального підпорядкування (Київ і Севастополь) для обробки даних Всеукраїнського перепису населення 2001 р. у межах АС «Перепис-2001», за допомогою якої було просканована та оброблено загалом близько 67,9 млн переписних документів;

– у ГМУС у місті Києві в системі оперативного аналізу даних Всеукраїнського перепису населення 2001 р. у багатовимірному вигляді (АС «Перепис-2001 Аналітик»), яка з моменту свого створення в 2004 р. і до нинішнього часу постійно використовується для науково-дослідницьких цілей та для підготовки відповідей на запити міжнародних і національних організацій, зокрема, ООН, ЮНЕСКО, Статистичного комітету СНД, Верховної Ради України тощо;

– у Національному бюро статистики Республіки Молдова, м. Кишинів, у АС «Перепис-Молдова 2004» для обробки даних перепису населення республіки Молдова 2004 р., за допомогою цієї системи було просканована й оброблено більше 5,0 млн переписних документів;

– у Державній службі статистики України для обробки даних пробного перепису населення України 2010 р. про 98,5 тисяч респондентів Дергачівського району Харківської області.

Крім того, теоретичні результати дисертаційної роботи:

– були застосовані під час підготовки технічного проекту на створення системи обробки даних з демографічної статистики і технічного завдання на створення прототипу автоматизованої системи оброблення даних Всеукраїнського перепису населення 2012 р. для дослідної експлуатації на матеріалах пробного перепису в рамках виконання робіт згідно з договором № 116 від 16.10.2008 між Державним комітетом статистики України і АТЗТ «Квазар-Мікро Техно» (вих. № 437/0/7-11 від 27.04.2011);

– були застосовані під час виконання НДР (державний реєстраційний номер 0110U006826) з розроблення технічного проекту на створення комплексу електронної обробки даних Всеукраїнського перепису населення 2012 року (вих. № 20-12/10 від 20.12.2010);

– застосовуються фахівцями Державної служби статистики України під час підготовки мікрофайлів по даних Всеукраїнського перепису населення 2001 р. (вих. № 10/2-15/730 від 17.11.2010);

– були використані в Державній службі статистики України при проведенні робіт з проектування Інтегрованої системи обробки статистичних даних (вих. № 12/1-5/168 від 01.07.2011).

Особистий внесок здобувача. Усі наукові результати дисертації отримані автором самостійно і опубліковані, зокрема, в одноосібно підготовлених працях [1, 6, 8–16, 26, 29, 30, 32, 40, 41, 52].

У друкованих працях, підготовлених за участю автора під час роботи в Національному технічному університеті України «Київський політехнічний інститут» і опублікованих у співавторстві, йому належить таке.

У працях [7, 35] – інформаційна технологія локалізації помилок у зведених даних і пошуку причин їх виникнення через забезпечення адаптивної підтримки співробітництва користувачів; у праці [5] – аналіз та опис територіальних архітектурних рішень для інформаційних систем національного масштабу в Україні.

У монографії [2] та працях, що використані у цій монографії чи опубліковані після її виходу за аналогічною тематикою: загальна редакція монографії [2], передмова та післямова в ній; підрозділ 1.1 монографії [2], праці [3, 17–20, 31, 36–39, 42, 46, 50] – огляд та аналіз існуючих методів забезпечення анонімності даних про респондентів, введення поняття групової анонімності даних та напрямів її забезпечення, постановка загальної (кількісної та концентраційної) задачі забезпечення групової анонімності даних і метод її розв’язання на основі застосування вейвлет-перетворень, окремі приклади для демонстрації методу; підрозділ 1.2 монографії [2] та [17–20, 42] – постановка концентраційно-різницевої задачі забезпечення групової анонімності даних і метод її розв’язання, окремі приклади для демонстрації методу; підрозділ 1.3 і [3, 17, 20] – спеціальна методика застосування групової анонімності для отримання оптимального маскування відповідних даних, що захищаються в мікрофайлі; глава 2 монографії [2] і [27, 43] – порівняльний аналіз необхідних і достатніх умов, що накладаються на передатні функції вейвлетів і фільтри для утворення ними базису діадного дискретного

вейвлет-перетворення; підрозділ 3.2 монографії [2] та [21–24, 28, 33, 34, 44, 45] – обґрунтування важливості застосування недіадного вейвлет-перетворення та додатково: в [21, 28, 33, 34, 45] – різні варіанти постановки задачі адаптивного підбору коефіцієнтів масштабування, в [21, 23, 24] – приклади вибору конкретних коефіцієнтів масштабування, обумовлених проблемною областю, в [21, 22] – зіставлення та порівняльний аналіз методів неперервного і дискретного вейвлет-перетворень, в підпункті 3.2.1.2 монографії [2] та в [22, 45] – дослідження недіадного дискретного вейвлет-перетворення з цілочисловим коефіцієнтом масштабування; підрозділ 3.3 монографії [2] та [25] – підхід, що базується на використанні поліномів наближення в просторі з породжуючим елементом для пошуку заздалегідь визначеного шаблону в статистичних даних мікрофайлів.

У працях [4, 47–49, 51] автору належить підхід, який полягає у взаємодоповнювальному застосуванні можливостей інтелектуального аналізу даних та засобів інформаційно-облікових і аналітичних систем, а також додатково в праці [49] – приклад відповідного розв'язання задачі кластеризації для демографічних даних, у працях [4] і [51] – постановка задачі з узагальнення запропонованого в [49] алгоритму пошуку закономірностей із застосуванням нечіткої кластеризації та асоціативних правил відповідно.

Апробація результатів дисертації. Основні ідеї, положення та результати наукових досліджень доповідалися на міжнародних та всеукраїнських наукових симпозіумах і конференціях, зокрема, на: міжнародному симпозіумі «Россияне в зеркале статистики: Всероссийская перепись населения 2002 года» (м. Москва, 2004); X, XI і XIII міжнародних науково-технічних конференціях «Системний аналіз та інформаційні технології» (САІТ) (м. Київ, 2008, 2009, 2011); XII і XIII міжнародних наукових конференціях імені акад. М. Кравчука (м. Київ, 2008, 2010); I, II і III наукових конференціях магістрантів та аспірантів «Прикладна математика та комп'ютинг» (ПМК) (м. Київ, 2009–2011); IX, X, XI і XIII міжнародних наукових конференціях «Інтелектуальний аналіз інформації» (ІАІ) ім. Т. А. Таран (м. Київ, 2009–2011, 2013); VI, VII і IX міжнародних науково-практичних конференціях «Інформаційні технології та безпека в управлінні» (ITSM) (смт Кореїз АР Крим, 2009; смт Затока Одеської області, 2010; м. Севастополь, 2012); International Symposium on Computing, Communication and Control (ISCCC 2009) (Singapore, 2009); III міжнародній науково-практичній конференції «Інформаційна та економічна безпека» (INFECO–2010) (м. Харків, 2010); науково-технічній конференції з міжнародною участю «Комп'ютерне моделювання в наукоємних технологіях» (КМНТ–2010) (м. Харків, 2010); міжнародній науково-практичній конференції «Інформаційні технології та комп'ютерна інженерія» (м. Вінниця, 2010); IV, V і VI всеукраїнських науково-практичних конференціях «Сучасні тенденції розвитку інформаційних технологій в науці, освіті та економіці» (м. Луганськ, 2010–2012); міжнародних науково-практичних конференціях «Ольвійський форум: стратегії України в геополітичному просторі» (м. Ялта, 2010, 2011, 2013; м. Севастополь, 2012); міжнародних науково-технічних конференціях «Штучний інтелект. Інтелектуальні системи» (ШІ–2010 і ШІ–2012) (смт Кацівелі,

АР Крим, 2010, 2012); VI міжнародній науково-технічній конференції «Сучасні інформаційно-комунікаційні технології» (КОМІНФО–2010) (смт Лівадія, АР Крим, 2010); 27th International Conference on Information Systems (BNCOD 2010) (Dundee, Great Britain, 2010); International Conference on Information Processing and Management of Uncertainty (IPMU 2010) (Dusseldorf, Germany, 2010); 2010 International Conference on Signals and Electronic Systems (ICSES 2010) (Gliwice, Poland, 2010); World Conference on Soft Computing (WConSC'11) (San Francisco, USA, 2011); 2nd International Conference on Integrated Information (IC-ININFO 2012) (Budapest, Hungary, 2012).

Матеріали, покладені в основу дисертаційної роботи, також були розглянуті на наукових та науково-виробничих семінарах і нарадах: у Державній службі статистики України (м. Київ, 2004 р., 2008–2012 рр.), у Національному бюро статистики республіки Молдова (м. Кишинів, Молдова, 2006 р.), у Національному інституті статистики та економічних досліджень (м. Париж, Франція, 2008 р.), у компанії ICF Macro (на CSPro Software Workshop, м. Калвертон, штат Меріленд, США, 2009 р.), в Інституті проблем математичних машин і систем НАН України (на семінарі «Обчислювальні машини та інформаційні технології спеціального призначення» Наукової ради з проблеми «Кібернетика» НАН України, м. Київ, 2010 р.), у Навчально-науковому комплексі «Інститут прикладного системного аналізу» Національного технічного університету України «Київський політехнічний інститут» (м. Київ, 2011 р.).

Публікації. Результати дисертації викладені в 52 наукових працях, у тому числі в 4 колективних монографіях, у 2 з яких автор також виступив як єдиний редактор, у 21 статті в наукових журналах та 4 статтях у збірниках наукових праць, що включені до Переліку ДАК України з технічних наук (з них 13 статей одноосібних), у 3 статтях у зарубіжних фахових наукових журналах, у 20 публікаціях у працях і тезах доповідей симпозіумів і конференцій, які відбувалися в Україні, Росії, США, Німеччині, Великій Британії, Польщі, Угорщині та Сінгапурі.

Структура та обсяг дисертації. Дисертаційна робота складається з вступу, п’ятьох розділів, висновків, списку використаних джерел і двох додатків. Повний обсяг дисертації становить 383 сторінок, у тому числі: 274 сторінки основного тексту, 61 рисунок, 44 таблиці, список використаних джерел із 345 найменувань.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** подано загальну характеристику роботи з точки зору критеріїв дисертації.

У **першому розділі** дисертаційної роботи наведено особливості процесів обробки демографічної інформації та проаналізовано сучасний стан систем обробки відповідної інформації. Значну увагу приділено розгляду існуючих архітектурних

моделей і рішень, ключових інформаційних технологій, що застосовуються при побудові сучасних систем обробки демографічної інформації.

На основі проведеного порівняльного аналізу чотирьох основних міжнародних моделей і стандартів, що пов'язані з моделюванням статистичних бізнес-процесів, відібрано загальну модель статистичних бізнес-процесів (ЗМСБП), оскільки вона з точки зору виконання статистичних задач фактично перекриває можливості трьох інших моделей, хоча й має певні недоліки.

Показано, що існуючі наразі нормативні та рекомендаційні системи статистичних метаданих, за винятком моделі Сунгрена, розраховані або на опис архітектури та загальних статистичних бізнес-процесів, або на опис метаданих з точки зору взаємодії з іншими (нестатистичними) інформаційними системами. В той же час для практичного впровадження інтегрованої інформаційної статистичної системи, керованої метаданими, потрібна також детальна проробка моделі метаданих, що буде використовуватися для опису статистичної інформації та конкретних (детальних) процесів її обробки, відповідних інформаційних технологій.

Наведено короткий огляд методів інтелектуального аналізу даних (ІАД), що можуть застосовуватися для обробки демографічної інформації. Показано, що актуальними є роботи, направлені на застосування методів ІАД до демографічних даних із взаємодоповнювальним використанням можливостей інформаційно-облікових систем і систем оперативного аналізу даних.

Первинні дані, зібрані в ході перепису населення чи іншого демографічного спостереження, є інформацією обмеженого доступу, а використання зібраних персональних даних в історичних, статистичних чи наукових цілях може здійснюватися лише в знеособленому (анонімізованому) вигляді. Індивідуальна анонімність даних – це властивість інформації про окремого респондента бути неідентифікованою в наборі даних. У роботі були проаналізовані та класифіковані існуючі методи забезпечення анонімності даних про респондентів – як у вихідних таблицях (зведених даних), так і в мікрофайлах. Під мікрофайлом розуміють таблицю з первинними статистичними даними, як на рис. 1, де кожному респонденту r_k, $k=\overline{1,\mu}$ зіставлені його значення z_{kn} атрибутів u_n, $n=\overline{1,\eta}$.

		А т р и б у т и			
		u_1	u_2	…	u_η
Респонденти	r_1	z_{11}	z_{12}	…	$z_{1\eta}$
	r_2	z_{21}	z_{22}	…	$z_{2\eta}$
	…	…	…	…	…
	r_μ	$z_{\mu 1}$	$z_{\mu 2}$	…	$z_{\mu\eta}$

Рисунок 1 – Дані мікрофайла, оформлені у вигляді таблиці

Якщо є вихідний мікрофайл *X*, то для розв'язання задачі забезпечення анонімності даних необхідно отримати такий новий (захищений) мікрофайл *X'*, який задовольняє вимоги: 1) ризик розголошення інформації, що захищається, має бути мінімальним чи принаймні адекватним її важливості; 2) аналіз даних про респондентів у мікрофайлах *X* і *X'* має давати однакові чи близькі результати; 3) вартість перетворення даних мікрофайла *X* має бути прийнятною.

Серед методів забезпечення анонімності даних можна виділити метод, запропонований L. Liu , J. Wang і J. Zhang, який базується на вейвлет-перетворенні (ВП). У ньому через дискретне ВП для початкових даних розраховуються апроксимаційні та деталізуючі коефіцієнти, а потім за допомогою порогового відсічення високочастотних деталізуючих коефіцієнтів відбувається маскування даних про респондентів. У третьому розділі дисертаційної роботи будується метод, фактично двоїстий до зазначеного: деталізуючі коефіцієнти зберігаються, а апроксимаційні – змінюються.

Найбільш поширеним методом теорії вейвлетів є кратномасштабний аналіз, який базується на представленні даних з різним ступенем деталізації. Це дає можливість вивчати глобальні особливості даних за великомасштабного представлення і виділяти локальні особливості на менших масштабах. Кратність кратномасштабного аналізу є коефіцієнтом масштабування відповідного ВП. При цьому функція масштабування $\varphi(x)$ визначається за допомогою такого рівняння:

$$\varphi(x) = \sqrt{N} \sum_n h_n \varphi(Nx - n), \qquad (1)$$

де N – коефіцієнт масштабування;

n – цілі числа;

h_n – коефіцієнт низькочастотного фільтра ВП, причому $\sum_n |h_n|^2 < \infty$.

Але всі відомі наразі методи забезпечення анонімності направлені, насамперед, на гарантування знеособленості даних про окремих респондентів. У той же час проблема забезпечення анонімності даних про групи респондентів залишається відкритою.

Загалом були виділені такі ключові виклики та тенденції розвитку систем для обробки демографічної інформації:

1. Стрімке збільшення потужностей обчислювальної техніки та подальший розвиток інструментальних програмно-технічних засобів дають змогу економістам-статистикам переходити від аналізу агрегованих (зведених) даних до всебічного дослідження великих вибірок первинних даних, що робить дуже актуальними питання забезпечення анонімності даних не тільки про окремих респондентів, але й про їх певні групи.

2. З кожним новим циклом перепису ускладнюється його програма: збільшується перелік питань і взаємозв'язків між можливими відповідями,

ускладнюються правила перевірки правильності заповнення анкет тощо. Досвід, набутий автором дисертаційного дослідження під час розроблення, впровадження і супроводження автоматизованих систем з обробки даних Всеукраїнського перепису населення 2001 р., перепису населення республіки Молдова 2004 р. і пробного перепису населення в Україні в 2010 р., свідчить, що найбільш трудомісткими, а значить, такими, що вимагають подальшої автоматизації і спрощення, є операції пошуку причини помилки, виявленої на етапі проведення контролів первинних чи зведених даних.

Потрібне розроблення відповідних нових інформаційних технологій. Зокрема, застосування для введення статистичних даних швидкодіючих промислових сканерів дає можливість організувати процес локалізації та пошуку помилок у первинних даних зручним та ефективним для користувачів способом. Своєю чергою організація обміну досвідом між користувачами, що займаються пошуком помилок у зведених даних, також може значно прискорити весь процес обробки наявних даних.

3. За останні кілька раундів проведення переписів населення досягнуто прийнятного рівня якості та точності даних опитування, але все зростаючими є вимоги до відповідних інформаційних систем щодо здійснення детального та оперативного аналізу отриманих даних, а також їх гармонізації з результатами демографічних спостережень інших країн за допомогою метаданих. Після завершення поточного раунду переписів населення 2010 р. статистичні організації країн СНД вперше в своїй історії отримають дані, за якими можна буде провести порівняльне (зіставне) дослідження підсумків двох останніх переписів. Однак проведений аналіз існуючих інформаційних систем у галузі статистики населення показав, що хоча вони й забезпечують основну функціональність зі збереження та обробки переписних даних (на рівні інформаційно-облікових систем) та додатково дають користувачам змогу будувати аналітичні звіти (на рівні інформаційно-аналітичних систем), проте пошук прихованих закономірностей, що є в аналізованих даних, виконується фактично лише вручну через висунення певних гіпотез і перевірку їх справедливості за допомогою побудови аналітичних звітів чи взагалі здійснюється в межах інших відокремлених систем.

Зважаючи на виявлені недоліки існуючих інформаційних систем з обробки демографічних даних і враховуючи необхідність ліквідації «клаптикової» автоматизації статистичних бізнес-процесів, впровадження методів ІАД, застосування відповідних моделей подання метаданих і забезпечення несуперечності та анонімності переписних даних, було виконано постановку завдання дисертаційного дослідження.

У **другому розділі** на основі проведеного аналізу існуючих методик (frameworks) опису корпоративних архітектур для представлення функціонально-інформаційної архітектури статистичної галузі України була відібрана модель Захмана, оскільки:

– вона є універсальною і достатньою для опису корпоративної архітектури будь-якого масштабу;

– більшість інших методик мають відображення на модель Захмана і тому, за необхідності, від моделі Захмана можна перейти до цих моделей;

– модель Захмана є найбільш вичерпною серед найуживаніших методик опису корпоративних архітектур.

Модель Захмана в сучасному варіанті зображується у вигляді таблиці, розміром 6 на 6, колонкам якої відповідають домени, а рядкам – рівні абстракції моделі. Однак спроба її застосувати безпосередньо під час розроблення технічних проектів на систему обробки даних з демографічної статистики Держстату України та на комплекс електронної обробки даних другого Всеукраїнського перепису населення виявили необхідність в удосконаленні моделі Захмана через виділення додаткового рівня абстракції, який названо рівнем системної архітектури, тобто рівнем архітектури систем по роботі з даними. Основна сфера його відповідальності – це пошук інформації, що важливо і характерно саме для статистичної галузі. У табл. 1 цей новий архітектурний рівень розміщений між рівнями методологів та архітекторів. Виділення цього рівня може виявитися корисним і для інших галузей, що також як і статистична – орієнтовані на обробку нематеріальних ресурсів.

На відміну від архітектурних рішень існуючих систем обробки демографічних даних, які базуються на використанні лише інформаційно-облікових та інформаційно-аналітичних систем, запропоновано додати нову складову – систему ІАД.

Загалом, з точки зору рівня системної архітектури, системи обробки демографічної інформації можуть бути трьох видів:

1) інформаційно-облікові (OLTP – OnLine Transaction Processing), що забезпечують базову функціональність, таку як введення даних та матеріалів відповідних статистичних спостережень і опитувань, їх структуроване (як правило, за допомогою СКБД) зберігання та облік, контроль первинних і зведених даних, розповсюдження результатів через різноманітні регламентні вихідні таблиці; прикладами систем такого виду є розроблені та впроваджені під керівництвом автора дисертаційної роботи автоматизовані системи з обробки даних Всеукраїнського перепису населення 2001 р. і даних перепису населення Молдови за 2004 р.;

2) інформаційно-аналітичні (OLAP – OnLine Analytical Processing), за допомогою яких користувачі можуть швидко будувати нерегламентні таблиці і проводити інші аналітичні дослідження статистичних даних, шукаючи передбачувані закономірності їх розподілу. Прикладом системи такого виду є розроблена та впроваджена під керівництвом автора дисертаційної роботи автоматизована система багатовимірного аналізу даних Всеукраїнського перепису населення 2001 р.;

3) системи ІАД, які беруть на себе найбільш громіздку та рутинну аналітичну операцію з пошуку прихованих закономірностей, що, можливо, існують в даних, які аналізуються.

Таблиця 1 – Модель Захмана для функціонально-інформаційної архітектури статистичної галузі України

Рівень абстракції	Домен					
	ЧОМУ (мотивація)	ЩО (дані)	ЯК (функції)	ДЕ (дислокація)	ХТО (люди)	КОЛИ (час)
Контекст (сфера дії), рівень керівництва	Законодавчо визначені цілі	Статистичні обстеження та показники	Перелік послуг і процесів	Географічний розподіл органів статистики, місця проведення спостережень	Ієрархічна (трирівнева) організаційна структура, споживачі інформації	Програма статистичних спостережень
Концептуальна (бізнес-) модель, рівень методологів	Моделі показників звітності	Метамоделі, моделі інформації, класифікатори	Модель статистичних бізнес-процесів	Схема логістики (інформаційних потоків та потоків керування)	Моделі потоків робіт (workflow), перелік ролей користувачів	Календар (і періодичність) спостереження
Системна модель, рівень системних аналітиків	Пошук інформації	Архітектурна модель	Робота з даними, доступ до них	Модель системної архітектури	Моделі запитів	Узгодженість і синхронізація запитів
Логічна модель, рівень архітекторів	Суб'єкти та моделі діяльності	Логічні моделі даних	Архітектура застосувань	Модель розподіленої архітектури	Модель рольових відносин	Діаграма подій, їх синхронізація
Технологічна модель, рівень проектувальників	Посадові інструкції	Фізичні моделі даних	Специфікації використання	Інфраструктура	Інтерфейси користувача	Специфікація подій
Деталі реалізації, рівень виконавців і розробників	Керівництва користувача, інструкції із застосування, регламент робіт	Опис структур даних	Програмний код, програми, що виконуються	Мережева архітектура	Автоматизовані робочі місця	Визначення часових прив'язувань
Практика застосування	Реальні показники точності, своєчасності, надійності інформації, що надається	Первинні та агреговані дані, мікродані	Регламенти, що реалізуються, послуги, що надаються	Фізичне розміщення обладнання, канали надання послуг	Постачальники і споживачі інформації та послуг	Розклад надання (чи доступності) інформації

На рис. 2 наведена загальна схема етапів введення та обробки демографічної інформації з розподілом їх на 3 виділені види систем рівня системної архітектури. Для реалізації найбільш трудомістких етапів з пошуку причин суперечностей у первинних і зведених даних та забезпечення їх захищеності (етапи 2, 4, 5 і 6 на рис. 2) в наступних розділах дисертації запропоновані відповідні методи та інформаційні технології.

Рисунок 2 – Основні етапи обробки демографічної інформації

Серед усіх комірок моделі Захмана для функціонально-інформаційної архітектури статистичної галузі України найважливішою та найспецифічнішою є комірка «модель статистичних бізнес-процесів».

Враховуючи практику проведення статистичних спостережень в Україні, що склалася за роки незалежності, і явно виділивши процеси, які забезпечують анонімність статистичної інформації, модель ЗМСБП можна конкретизувати, утворивши в ній 14 процесів і територіальні рівні статистичної обробки (рис. 3). На місцевому рівні свою діяльність здійснюють 498 районних і міських відділів статистики. Офіційними органами статистики України на регіональному рівні є Головне управління статистики в Автономній Республіці Крим, 25 головних управлінь статистики в областях і місті Києві, а також Управління статистики в місті Севастополі. Центральний рівень забезпечує Державна служба статистики України,

куди також віднесемо ГМУС (реорганізований у 2000 р. колишній Головний обчислювальний центр).

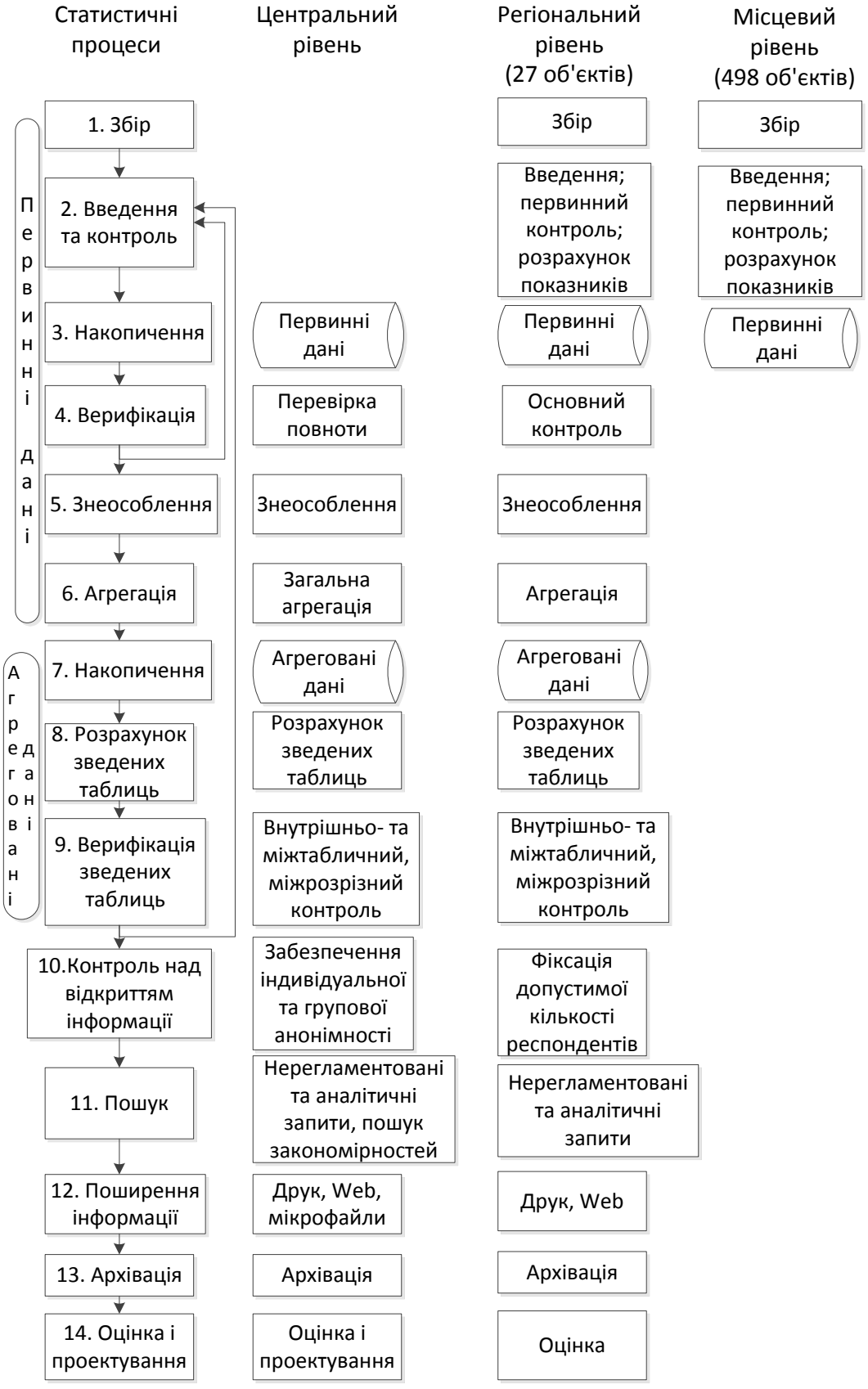

Рисунок 3 – Процеси і територіальні рівні статистичної галузі України

Формально, наведена на рис. 3 модель дає можливість кожному статистичному процесу однозначно зіставити елемент із системи *MBP*, що складається із двох множин:

$$MBP = <Area, Number>, \quad (2)$$

де *Area* – множина територіальних рівнів обробки даних у статистичній галузі України, яка нині складається з трьох елементів: центральний (державний), регіональний (обласний, Автономної Республіки Крим, міст Києва і Севастополя) та місцевий (районний);

Number – множина номерів процесів від 1 до 14 (див. рис. 3).

У зв'язку з тим, що питання уточнення потреб, проектування та формування статистичного процесу мають насамперед методологічний характер і здійснюються, як правило, один раз для конкретного процесу, вони були винесені за межі моделі, що розглядається. З іншого боку, систематичні процеси обробки статистичних даних були ґрунтовно деталізовані порівняно з моделлю ЗМСБП.

З точки зору наближення статистичних даних до їх користувача, дуже важливим є впровадження доступу до пошуку необхідної інформації:

– за допомогою нерегламентних запитів у інформаційно-обліковій статистичній системі;

– через аналітичні запити в інформаційно-аналітичній статистичній системі;

– застосуванням інтелектуальних інформаційних технологій для автоматизованого пошуку певних закономірностей у даних.

Тому в запропонованій на рис. 3 моделі був явно виділений процес «Пошук», який у моделі ЗМСБП розосереджено по процесах «Обробка», «Аналіз» і «Поширення».

У моделі, що пропонується, підпроцес «Контроль над відкриттям інформації» моделі ЗМСБП розділено на два загальні процеси: «Знеособлення» та, власне, «Контроль над відкриттям інформації», оскільки дані процеси застосовуються на різних етапах обробки статистичних даних.

В останньому із зазначених процесів виділено такі підпроцеси:

1) визначення мінімально допустимої кількості респондентів, інформація про яких може бути надана (у вихідній таблиці, у відповіді на аналітичний запит тощо); наприклад, під час обробки даних Всеукраїнського перепису населення 2001 р. ця кількість дорівнювала 3000 осіб, у французькій статистиці застосовується нижня межа в 4000 респондентів тощо;

2) забезпечення індивідуальної та групової анонімності даних.

Таким чином, подана на рис. 3 модель статистичних бізнес-процесів позбавлена недоліків, притаманних ЗМСБП:

– виокремлено процес пошуку інформації, причому він по-різному представлений на регіональному і центральному рівнях статистичної галузі України;

– розширено та деталізовано процеси забезпечення анонімності статистичної інформації;

– модель повністю адаптована до особливостей обробки даних у статистичній галузі України.

З допомогою наскрізного виокремлення процесів забезпечення анонімності даних про респондентів, крім моделі ЗМСБП, була удосконалена ще одна відома модель – модель метаданих Сунгрена для опису статистичної інформації.

Удосконалену модель метаданих Сунгрена для опису статистичної інформації можна представити за допомогою системи *MM*, що складається з чотирьох множин:

$$MM = <Level, Group, Component, Mapping>, \quad (3)$$

де *Level* – множина проекцій (семіотичних рівнів), у якій додатково до синтаксичного, семантичного та прагматичного рівнів моделі Сунгрена додано рівень контролю над відкриттям статистичної інформації;

Group – множина груп і підгруп статистичних метаданих (загальні, декларативні, процесо-зорієнтовані, фактографічні, алгоритмічні, типу зворотного зв'язку), розширена групою метаданих, орієнтованою на контроль над відкриттям інформації;

Component – множина компонент метаопису макроданих, яка складається з 4 компонент (вимірів) опису: 1) α – об’єктний (респондентський) вимір; 2) β – вимір підсумків; 3) γ – класифікаційний вимір; 4) τ – часовий вимір;

Mapping – відображення, що зіставляє одному реквізиту (елементу) опису статистичної інформації один семіотичний рівень та одну чи кілька підгруп статистичних метаданих і компонент метаопису макроданих.

Як ілюстрацію застосування в Інтегрованій системі обробки статистичних даних України запропонованої розширеної моделі метаданих Сунгрена для опису статистичної інформації в дисертації наведено фрагмент опису поняття «статистичне спостереження».

Значну частину другого розділу присвячено розгляду питань інтелектуального аналізу демографічних даних та організації його взаємодії з іншими складовими рівня системної архітектури. Також був наведений фрагмент моделі корпоративної архітектури на мові ArchiMate.

Загалом, модернізація статистичної галузі України, що здійснюється в рамках проекту розвитку системи державної статистики України для моніторингу соціально-економічних перетворень і яка включає в себе впровадження вдосконалених моделей, що описані в другому розділі дисертації, за оцінками експертів згідно спеціалізованої методології PARIS21, дасть змогу підвищити статистичну спроможність системи державної статистики приблизно на 30 %.

Третій розділ фактично складається з двох частин. Перша частина містить формалізацію застосування недіадного та діадного дискретних вейвлет-перетворень до аналізу демографічної інформації. А в другій частині розвинутий апарат вейвлет-перетворення використовується як основа принципово нового методу забезпечення

анонімності даних про респондентів – методу забезпечення групової анонімності даних.

Зрозуміло, що внесення деяких змін, здійснене при маскуванні даних, які захищаються у вихідному мікрофайлі, призведе до втрати певною мірою корисності отриманої інформації.

Однак група російських вчених під керівництвом А. О. Давидова, першого віце-президента російського товариства соціологів, показала, наприклад, за результатами дослідження 44 опитувань суспільної думки громадян Російської Федерації за 1994–2001 рр., що деталізуючі складові вейвлет-розкладення відображають приховані особливості часових рядів, які можна використовувати для коротко- та середньострокових прогнозувань соціальних процесів. Отриманий висновок дає змогу в методі забезпечення групової анонімності даних у мікрофайлах застосовувати ВП для пошуку балансу між зміною початкових даних і запобіганням втрати їх корисності.

На практиці найчастіше використовуються діадні ВП. Проте в деяких випадках двійкове масштабування не завжди є природнім для певної предметної області чи розв’язуваної задачі. Наприклад, визначення особливостей стійкості індикаторів соціально-економічного стану регіону зручно робити, виділивши окремі часові масштаби в 2, 4 і 8 кварталів. Аналіз демографічних даних по сім’ях в Україні краще робити з коефіцієнтом масштабування, рівним 3 (тобто тріадними вейвлетами), – за середньою кількістю людей у сімейному осередку.

Кожен із поширених діадних вейвлетних базисів має певні особливості, завдяки яким він і набув розповсюдження. Так, діадні сплайн-вейвлети дають можливість точно виявляти в розподілі вихідних даних різкі зміни амплітуди сигналу. А оскільки порогове функціонування є загальносистемною властивістю динаміки соціальних і демографічних систем, то такі пошукові задачі характерні й для проблемних областей, що досліджуються, наприклад, задачі виявлення різких змін соціальних процесів у історичній ретроспективі, знаходження системних закономірностей динаміки значень індексу розвитку людського потенціалу Human Development Index для соціуму, пошуку адміністративно-територіальних меж появи певних особливостей розподілу демографічних даних.

Отже, важливою є задача побудови не просто недіадних ВП, а таких недіадних базисів, що зберігають бажані властивості своїх діадних прототипів.

У дисертації застосовується схема Брателлі (O. Bratelli) та Йоргенсена (P. E. T. Jorgensen) для побудови аналітичних коефіцієнтів фільтрів одновимірного тріадного ВП на базі біортогональних сплайн-вейвлетів. У формулі (4) наведено для прикладу розраховану в дисертації передатну функцію $G_0(z)$ відповідного тріадного низькочастотного фільтра відновлення.

Саме зазначені підраховані фільтри з урахуванням справедливості умов Веттерлі точного відновлення сигналу діадним дискретним біортогональним ВП дали змогу довести важливе для практичних застосувань твердження про неможливість збереження в загальному випадку в недіадному дискретному ВП хоча б ще одного фільтра існуючого діадного ВП, крім низькочастотного фільтра

розкладення. Зберігати в недіадному дискретному ВП із існуючого діадного ВП лише фільтр чи фільтри відновлення на практиці немає сенсу, бо в цьому випадку під час реконструкції з відповідним коефіцієнтом масштабування будемо отримувати візуально схожими сигнали, різні за своєю структурою.

$$G_0(z)=\Big((b_{22}b_{33}-b_{23}b_{32})\Big(\big(h^0_{-2}h^0_0-h^0_{-3}h^0_1\big)z^5+\big(h^0_{-3}h^0_2-h^0_{-1}h^0_0\big)z^4+$$
$$+\big(h^0_{-1}h^0_1-h^0_{-2}h^0_2\big)z^3\Big)+\big(b_{21}b_{32}h^0_2-b_{22}h^0_2b_{31}+b_{22}h^0_0b_{33}-b_{23}b_{32}h^0_0- \quad (4)$$
$$-b_{21}h^0_1b_{33}+b_{23}h^0_1b_{31}\big)\big(-h^0_{-2}z^2+h^0_{-1}z\big)+\big(b_{21}b_{33}h^0_{-2}-b_{22}h^0_{-3}b_{33}+b_{22}h^0_{-1}b_{31}-b_{21}b_{32}h^0_{-1}-$$
$$-b_{23}h^0_{-2}b_{31}+b_{23}h^0_{-3}b_{32}\big)\big(-h^0_1z^{-1}+h^0_2z^{-2}\big)\Big)/K,$$

де h^0_i, $i = -3, -2, \ldots, 3$ – коефіцієнти низькочастотного (апроксимативного) фільтру біортогональних сплайн-вейвлетів діадного розкладення;

b_{kj} – елементи другого та третього рядку матриці

$$\begin{pmatrix} h^0_{-3}+h^0_0+h^0_3 & h^0_{-2}+h^0_1 & h^0_{-1}+h^0_2 \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix},$$

які є такими дійсними числами, щоб вона була невиродженою;

$$K\equiv 3\big(h^0_{-2}h^0_2-h^0_{-1}h^0_1\big)\big(K_1\big(h^0_{-3}+h^0_0+h^0_3\big)+K_2\big(h^0_{-2}+h^0_1\big)+K_3\big(h^0_{-1}+h^0_2\big)\big);$$

$$K_1\equiv b_{23}b_{32}-b_{33}b_{22};\ K_2\equiv b_{21}b_{33}-b_{23}b_{31};\ K_3\equiv b_{22}b_{31}-b_{32}b_{21}.$$

У дисертації наводиться запропоноване автором формалізоване зображення недіадного і діадного дискретних ВП, які використовуються для опису прикладів розв'язання задачі забезпечення групової анонімності даних.

Відповідно до нього матричне визначення для апроксимаційної складової $\vec{A}_k$ початкового сигналу на k-му кроці вейвлет-розкладення можна записати такою формулою:

$$\vec{A}_k=\prod_{i=1}^{k}<(\vec{a}_k{}^{\langle L_i|R_i\rangle})_{\uparrow N}*\vec{h}>_{K_i}=\Big\langle \mathbf{M}^c_{\left(N\left([K_1/N]+\tilde{L}_k+\tilde{R}_k\right)+n-1\right)\times\left(N\left([K_1/N]+\tilde{L}_k+\tilde{R}_k\right)\right)}(\vec{h})\times$$
$$\times\mathbf{M}^u_{N\left([K_1/N]+\tilde{L}_k+\tilde{R}_k\right)\times\left([K_1/N]+\tilde{L}_k+\tilde{R}_k\right)}\times\ldots\times\Big\langle \mathbf{M}^c_{\left(N\left([K_k/N]+\tilde{L}_1+\tilde{R}_1\right)+n-1\right)\times\left(N\left([K_k/N]+\tilde{L}_1+\tilde{R}_1\right)\right)}(\vec{h})\times \quad (5)$$
$$\times\mathbf{M}^u_{N\left([K_k/N]+\tilde{L}_1+\tilde{R}_1\right)\times\left([K_k/N]+\tilde{L}_1+\tilde{R}_1\right)}\times\vec{a}_k{}^{\langle\tilde{L}_1|\tilde{R}_1\rangle}\Big\rangle_{[K_{k-1}/N]}\ldots\Big\rangle_m,$$

де $\vec{a}_k$ – вектор апроксимаційних коефіцієнтів;

N – коефіцієнт масштабування;

$\vec{x}^{\langle L|R\rangle} \equiv (y_1,...,y_L,\vec{x},y_{L+1},...,y_{L+R}),\ \forall i\ \exists j\ y_i = x_j$ – розширення вектора $\vec{x}$;

$\vec{h}$ – низькочастотний фільтр вейвлет-перетворення;

$\vec{x}_{\uparrow N}$ – N-кратна інтерполяція вектора $\vec{x}$;

$*$ – згортка векторів;

$<\vec{x}>_K$ – виділення із вектора $\vec{x}$ його центральної частини, довжиною K;

$\lfloor K / N \rfloor$ – результат ділення числа K на N націло;

$\mathbf{M}^c$ – матриця згортки;

$\mathbf{M}^u$ – матриця N-кратної інтерполяції.

Загалом для забезпечення групової анонімності спочатку потрібно по набору даних, що захищається, побудувати зображення, в якому явно виділяються чутливі до розкриття особливості цього набору. Далі в отриманому зображенні зазначені особливості необхідно замаскувати, після чого здійснити відповідні зміни в початковому наборі даних.

Для формалізації постановки задачі забезпечення групової анонімності даних наведемо необхідні означення.

Означення 1. Сутнісною комбінацією значень назвемо вектор $\vec{s}_k^{(v)} \in S_v$, $k = \overline{1, l_v}$, $l_v \le \mu$, де S_v – підмножина декартового добутка $u_{v_1} \times u_{v_2} \times ... \times u_{v_t}$ атрибутів, тобто колонок з рис. 1, з номерами, рівними v_j, $j = \overline{1,t}$. Кожен елемент вектора $\vec{s}_k^{(v)}$ будемо називати сутнісним значенням, а параметри u_{v_j} – сутнісними атрибутами.

Сутнісна комбінація значень окреслює ті атрибути, які буде використано під час визначення записів мікрофайла для переміщення. Наприклад, для зміни регіонального розподілу чоловіків середнього віку потрібно взяти «вік» і «стать» за сутнісні атрибути. У цьому випадку отримаємо сутнісну комбінацію {«чоловік», «45 – 65 років»}.

Означення 2. Параметризуючим значенням назвемо елемент $s_k^{(p)} \in S_p$, $k = \overline{1, l_p}$, $l_p \le \mu$, де S_p – підмножина елементів z_{kp} мікрофайла, що відповідають p-му атрибуту, причому $p \ne v_j \ \ \forall j = \overline{1,t}$. Своєю чергою, відповідний p-й атрибут будемо називати параметризуючим атрибутом.

Параметризуючі атрибути отримали таку назву, оскільки вони використовуються для розділення записів мікрофайла на категорії. Наприклад, взявши «адміністративно-територіальний об’єкт місця проживання респондента» як параметризуючий атрибут, отримаємо категорії респондентів, сформовані за місцем їх проживання.

Означення 3. Під цільовим сигналом розумітимемо числовий вектор $\vec{\theta} \equiv (\theta_1, \theta_2, ..., \theta_m)$, елементами якого виступають (можливо, після деяких

перетворень) кількості повторень записів із певною сутнісною комбінацією значень по параметризуючому атрибуту.

Іншими словами, для кожного параметризуючого значення відберемо ті записи мікрофайла, всі значення сутнісних атрибутів яких є сутнісними. Позначимо кількості таких записів q_i, $i=\overline{1,m}$, де m – кількість параметризуючих значень. Вектор $\vec{q} \equiv (q_1, q_2, ..., q_m)$ будемо називати кількісним сигналом. Цей вектор є окремим випадком цільового сигналу. Для забезпечення групової анонімності даних у мікрофайлі потрібно замінити цей сигнал на інший, який позначимо $\vec{\tilde{q}} \equiv (\tilde{q}_1, \tilde{q}_2, ..., \tilde{q}_m)$.

Кожен графік (розподіл) цільового сигналу можна характеризувати рядом основних особливостей – наявністю екстремумів чи тренду, статистичними характеристиками (середнім значенням, середньоквадратичним відхиленням тощо) і т.п.

Позначимо $G(\vec{V}, P)$ множину атрибутів, яка складається з одного чи кількох сутнісних атрибутів $\vec{V} = \{V_1, V_2, ..., V_l\}$ та параметризуючого атрибута P, $P \neq V_n$, $n=\overline{1,l}$.

Зазначимо, що параметризуючий атрибут не повинен збігатися з сутнісними атрибутами своєї множини атрибутів $G(\vec{V}, P)$, але параметризуючі атрибути різних множин атрибутів $G_i(\vec{V}_i, P_i)$, $i=1,...,k$ можуть збігатися, а множини сутнісних атрибутів – перетинатися.

Означення 4. Задача забезпечення групової анонімності даних полягає в тому, що для кожної множини атрибутів $G_i(\vec{V}_i, P_i)$, $i=1,...,k$, мікрофайла $\boldsymbol{X}$ потрібно здійснити таку модифікацію цільового сигналу, щоб відбулася зміна певної виділеної основної особливості його графіка, побудованого вздовж значень відповідного параметризуючого атрибута P_i, яка привела б до потрібного маскування початкового розподілу сутнісних комбінацій значень атрибутів $\vec{V}_i$. При цьому обов'язково вимагатимемо збереження середнього по цільовому сигналу, тобто

$$\sum_i \theta_i = \sum_i \tilde{\theta}_i , \qquad (6)$$

де $\tilde{\theta}_i$ – елементи цільового сигналу з замаскованими даними.

Таким чином, забезпечення групової анонімності фактично означає перерозподіл записів із сутнісними комбінаціями значень відносно різних значень параметризуючого атрибута.

У загальному випадку, яку саме комбінацію значень сутнісних атрибутів необхідно захищати – випливає із переліку наявних атрибутів мікрофайла та тієї задачі, яка поставлена стосовно захисту розподілу певних даних мікрофайла.

Наступним означенням узагальнимо поняття цільового сигналу.

Означення 5. Цільовим поданням $\Omega(\boldsymbol{X},G)$ мікрофайла $\boldsymbol{X}$ відносно множини атрибутів G будемо називати набір даних, що відображає певні виділені особливості значень цієї множини атрибутів у мікрофайлі $\boldsymbol{X}$ у вигляді, зручному для подальшого забезпечення групової анонімності даних.

Методика забезпечення групової анонімності даних полягає в послідовному застосуванні таких кроків:

1. Побудувати мікрофайл $\boldsymbol{X}$, що містить знеособлені дані про респондентів.

2. Встановити одну чи кілька множин атрибутів $G_i(\vec{V}_i,P_i)$, $i=1,...,k$, які визначають категорії респондентів, що потрібно захистити.

3. Для всіх i від 1 до k:

3.1. Вибрати цільове подання $\Omega_i(\boldsymbol{X},G_i)$ відносно множини атрибутів $G_i(\vec{V}_i,P_i)$.

3.2. Визначити та застосувати функцію перетворення $\Upsilon:\boldsymbol{X}\to\Omega_i(\boldsymbol{X},G_i)$, яка за поточним станом мікрофайла $\boldsymbol{X}$ побудує вибране цільове подання.

3.3. Визначити та застосувати функцію перетворення $\Xi:\Omega_i(\boldsymbol{X},G_i)\to\Omega_i^{'}(\boldsymbol{X},G_i)$, що модифікує цільове подання.

3.4. Здійснити обернене перетворення $\Upsilon^{-1}:\Omega_i^{'}(\boldsymbol{X},G_i)\to\boldsymbol{X}$ й отримати поточний стан модифікованого мікрофайла $\boldsymbol{X}$.

Залежно від виду цільового подання, зокрема, цільового сигналу, можна виділити три різновиди задач забезпечення групової анонімності.

Перший вид – кількісна задача забезпечення групової анонімності – виникає в тих випадках, коли потрібно приховати екстремальні кількості певних сутнісних комбінацій для якогось параметризуючого значення. Для розв'язання цієї задачі спочатку необхідно побудувати кількісний сигнал $\vec{q}=(q_1,q_2,...,q_m)$, елементи якого дорівнюють числу всіх можливих пар виду {«сутнісна комбінація», «параметризуюче значення»} у вихідних даних, а потім замінити його на сигнал $\tilde{\vec{q}}=(\tilde{q}_1,\tilde{q}_2,...,\tilde{q}_m)$.

У багатьох випадках абсолютні кількості не є показовими, і більш прийнятним є аналіз відповідних концентрацій. Розв'язання концентраційної задачі забезпечення групової анонімності повністю зводиться до розв'язання кількісної задачі. Дійсно, спочатку побудуємо кількісний сигнал $\vec{q}=(q_1,q_2,...,q_m)$. Потім, розділивши кожен його елемент на загальну кількість записів по відповідному параметризуючому значенню, отримаємо концентраційний сигнал $\vec{c}=(c_1,\ c_2,\ \ldots,\ c_m)$. Змінюючи останній, прийдемо до нового концентраційного сигналу $\tilde{\vec{c}}=(\tilde{c}_1,\tilde{c}_2,...,\tilde{c}_m)$. Помноживши кожен елемент цього сигналу на кількість записів по відповідному параметризуючому значенню, отримаємо новий кількісний сигнал $\tilde{\vec{q}}=(\tilde{q}_1,\tilde{q}_2,...,\tilde{q}_m)$.

У деяких випадках більш адекватним може виявитися ще один різновид задачі забезпечення групової анонімності, відмінний від двох попередніх. Наприклад,

нехай потрібно приховати якесь військове містечко, населене здебільшого молодими людьми призивного віку. Очевидно, що місце розташування цього містечка легко можна виявити порівнянням концентрацій юнаків і дівчат у кожному регіоні. Різке перевищення перших над другими буде з високою ймовірністю вказувати на таке військове містечко. Це – типовий приклад концентраційно-різницевої задачі. Для її розв'язання потрібно побудувати два кількісні сигнали $\vec{q}^1 = (q_1^1, q_2^1, ..., q_m^1)$ і $\vec{q}^2 = (q_1^2, q_2^2, ..., q_m^2)$, що відповідають кожній групі респондентів (у нашому прикладі – юнакам і дівчатам). Аналогічно до попередньої ситуації можемо отримати два концентраційні сигнали: $\vec{c}^1 = (c_1^1, c_2^1, ..., c_m^1)$ і $\vec{c}^2 = (c_1^2, c_2^2, ..., c_m^2)$. Далі потрібно побудувати та модифікувати концентраційно-різницевий сигнал $\vec{\delta} = (\delta_1, \delta_2, ..., \delta_m) \equiv (c_1^1 - c_1^2, c_2^1 - c_2^2, ..., c_m^1 - c_m^2)$, отримавши в результаті новий сигнал $\vec{\tilde{\delta}} = (\tilde{\delta}_1, \tilde{\delta}_2, ..., \tilde{\delta}_m)$. Наступний етап – побудова нових концентраційних сигналів. Їх можна знайти розв'язанням системи із m лінійних рівнянь з $2m$ невідомими, якою визначається концентраційно-різницевий сигнал. У результаті отримаємо нові концентраційні сигнали $\vec{\tilde{c}}^1 = (\tilde{c}_1^1, \tilde{c}_2^1, ..., \tilde{c}_m^1)$ і $\vec{\tilde{c}}^2 = (\tilde{c}_1^2, \tilde{c}_2^2, ..., \tilde{c}_m^2)$, із яких очевидним чином відновлюються відповідні кількісні сигнали $\vec{\tilde{q}}^1 = (\tilde{q}_1^1, \tilde{q}_2^1, ..., \tilde{q}_m^1)$ і $\vec{\tilde{q}}^2 = (\tilde{q}_1^2, \tilde{q}_2^2, ..., \tilde{q}_m^2)$.

Функція перетворення Ξ, що модифікує цільове подання, має зберігати корисність даних, які будуть отримані в результаті коригувань. Залежно від того, які характеристики цільового сигналу повинні бути збережені, можна по-різному здійснювати забезпечення групової анонімності даних.

Базуючись на виявленому російським соціологом А. О. Давидовим факті, що деталізуючі складові вейвлет-розкладення відображають важливі особливості статистичних, зокрема часових, рядів, запропоновано та розвинуто метод забезпечення групової анонімності даних, який полягає в зміні екстремумів графіка цільового сигналу з одночасним збереженням його високочастотних характеристик.

Суть методу є такою. Цільовий сигнал розкладається за допомогою ВП. Оскільки деталізуючі складові вейвлет-розкладення сигналу залежать тільки від деталізуючих коефіцієнтів і вейвлет-фільтрів, то зміна апроксимаційних коефіцієнтів призведе до зміни лише апроксимаційної складової сигналу. Тому для забезпечення групової анонімності даних можемо (необхідним чином) перерозподілити апроксимаційні значення вейвлет-розкладення. При цьому, будемо запобігати втраті корисності даних, залишаючи незмінними (чи лише пропорційно змінюючи) їх деталізуючі складові. Додатково будемо зберігати і середнє початкових даних.

Образно кажучи, будемо змінювати рельєф (апроксимаційну складову), одночасно намагаючись зберегти всі локальні особливості (деталізуючі складові) початкового розподілу. Отже, запропонований метод забезпечення групової анонімності полягає в перерозподілі записів вихідного набору даних із комбінаціями значень, що захищаються, відносно різних значень певного параметризуючого

атрибута, що відповідальний за розподіл даних у просторі, в часі або по певних інших категоріях.

У **четвертому розділі** наведені детальні приклади та практичні аспекти застосування запропонованого методу до розв'язання задачі забезпечення групової анонімності даних у мікрофайлах у випадку кількісного, концентраційного та концентраційно-різницевого цільових сигналів. Викладення проведено на основі аналізу реальних даних, отриманих під час проведення переписів населення в США, Великій Британії та Італії відповідно.

Після того як знайдене потрібне перетворення початкового сигналу, тобто встановлена функція перетворення $\Xi\colon \Omega_i(\boldsymbol{X}, G_i) \to \Omega_i^{'}(\boldsymbol{X}, G_i)$, яка модифікує цільове подання сигналу, залишається лише створити новий мікрофайл. Іншими словами, потрібно здійснити перетворення $\Upsilon^{-1}\colon \Omega_i^{'}(\boldsymbol{X}, G_i) \to \boldsymbol{X}$ і отримати поточний стан модифікованого мікрофайла $\boldsymbol{X}$.

Це завжди можна зробити через зміну значення параметризуючого атрибута в певної кількості записів із сутнісною комбінацією значень. При цьому змінювати параметризуючі значення потрібно одночасно в двох записах, щоб чисельність записів по кожному параметризуючому атрибуту залишалася сталою. Це можна трактувати як взаємне переселення респондентів, яким відповідають зазначені записи. Такі «переселення» потрібно продовжувати до тих пір, поки не досягнемо шуканого розподілу цільового сигналу.

Визначальними атрибутами назвемо такі атрибути, розподіл значень яких за параметризуючими значеннями становить інтерес чи є корисним при подальшому дослідженні даних мікрофайла. Найкращим будемо визнавати варіант такого перетворення, яке, забезпечивши групову анонімність, мінімально спотворить значення визначальних атрибутів даних мікрофайла через обмін параметризуючими значеннями таких записів, що є «близькими» один до одного.

Як кількісну міру зазначеної близькості двох записів покладемо таку метрику.

Означення 6. Визначальною метрикою InfM назвемо функцію, що двом записам r і r^* зіставляє число, яке обчислюється за формулою

$$\mathrm{InfM}(r, r^*) = \sum_{p=1}^{n_{\mathrm{ord}}} \omega_p \left(\frac{r(I_p) - r^*(I_p)}{r(I_p) + r^*(I_p)} \right)^2 + \sum_{k=1}^{n_{\mathrm{nom}}} \gamma_k \chi^2 \big(r(J_k), r^*(J_k) \big), \qquad (7)$$

де I_p – p-й визначальний атрибут порядкового типу, загальна кількість таких атрибутів дорівнює n_{ord};

ω_p – ваговий коефіцієнт, чим він більший, тим вагомішим є відповідний атрибут I_p порядкового типу;

$r(\cdot)$ – функція, яка дає значення відповідного атрибута для запису r;

$r^*(\cdot)$ – функція, яка дає значення відповідного атрибута для запису r^*;

J_k – k-й визначальний атрибут номінального (категоріального) типу, загальна кількість таких атрибутів дорівнює n_{nom};

γ_k – ваговий коефіцієнт, чим він більший, тим вагомішим є відповідний атрибут J_k номінального типу;

$\chi(\nu_1, \nu_2)$ – функція, яка дорівнює певному числу χ_1, якщо значення ν_1 і ν_2 потрапляють до однієї категорії, та дорівнює іншому числу χ_2 – в протилежному випадку.

На відміну від раніше введених, запропонована метрика InfM(r, r*):

– дає можливість отримати кількісну оцінку відразу як порядкових, так і категоріальних атрибутів;

– бере до уваги не всі, а лише визначальні атрибути, тобто ті атрибути, значення яких важливі для подальшого дослідження даних мікрофайла;

– дає змогу кількісно врахувати (через вагові коефіцієнти ω_p і γ_k) вплив кожного з атрибутів.

Зрозуміло, що на практиці для реальних даних неможливо перебрати всі пари записів (r, r*) так, щоб вибрати серед них ті, обмін параметризуючими значеннями яких дасть для метрики InfM(r, r*) найменше значення. Тому в дисертаційній роботі були розроблені певні евристичні стратегії щодо підбору таких пар записів (r, r*), які давали б результати, прийнятні з точки зору обчислювальної складності та близькості до мінімального значення метрики InfM(r, r*).

Статистичний експеримент, у якому було побудовано 100 різних варіантів розв’язання задачі забезпечення групової анонімності кількісного розподілу військовослужбовців територією американського штату Массачусетс (за даними мікрофайла із записами про 148 815 респондентів), показав, що в середньому ці розв’язки вимагали зміни значень лише 0,12 % всіх записів мікрофайла, стандартне відхилення вибірки склало 0,02 %, а найкраща з розглянутих евристичних стратегій вимагала в середньому зміни значень лише 1,8 визначальних атрибутів із 7 наявних.

Додатково був проаналізований вплив цих змін на зміст результатів можливих запитів до БД, що містить дані мікрофайла. В експерименті було побудовано 7000 запитів, у яких кількість визначальних атрибутів була від 1 до 7, а конкретні визначальні атрибути запиту та обмеження на їх допустимі значення генерувалися випадковим чином. Виявилося, зокрема, що у всіх згенерованих випадках стандартне відхилення кількісного різницевого сигналу, побудованого як арифметична різниця між початковим і сформованим по регіонам сигналами кількостей записів, які задовольняють умовам запиту, не перевищило 2,18 запису, а в найгіршому випадку довірчий інтервал рівня 95 % для середнього значення кількісного різницевого сигналу в регіоні, яке дорівнює 0,15, склав лише [0,13; 0,18].

Отже, результати проведеного статистичного експерименту засвідчують, що на практиці, користуючись запропонованою визначальною метрикою InfM та евристичними стратегіями з підбору пар записів для обміну їх значеннями, можна побудувати новий мікрофайл, у якому забезпечена групова анонімність даних, а кількість спотворень значень визначальних атрибутів є прийнятною.

П’ятий розділ містить детальний опис загальної функціональної схеми інформаційно-облікової системи обробки переписних даних, яка була покладена в основу автоматизованих систем «Перепис-2001», «Перепис-Молдова 2004» та комплексу електронної обробки даних пробного перепису населення України за 2010 р. На прикладі зазначених систем показана можливість ефективної реалізації локалізації і пошуку помилок у первинних і зведених даних, тобто реалізації найбільш трудомістких операцій під час обробки даних перепису.

Зокрема, розроблена нова інформаційна технологія встановлення наявності та локалізації помилок у переписних документах з первинними даними. На відміну від існуючих підходів, запропонований підтримує інтегроване середовище, що містить поточний стан документа, його графічний образ і протокол із результатами контролю не лише конкретної переписної форми, але й їх сукупності, включаючи результати перевірки узгодженості значень показників різних переписних документів.

На рис. 4 наведена загальна схема інформаційної технології, що забезпечує встановлення наявності та локалізації помилок у первинних даних портфеля.

Хоча потенційно помилок у зведених даних набагато менше, ніж у первинних даних, але їх локалізація та пошук причин виникнення – значно складніші. На практиці пошук джерел таких помилок може бути виконаний лише за допомогою підсистеми підтримки нерегламентних запитів (НЗ), яка дозволяє за певним розрізом отримати визначений розподіл даних з БД. З часом, з накопиченням досвіду роботи ряд користувачів придумують певні шаблонні НЗ, які дають можливість швидко знаходити типові причини, що призводять до появи помилок у певних вихідних таблицях. У дисертації запропонована нова інформаційна технологія, яка на відміну від існуючих підходів для локалізації помилок у зведених даних і пошуку причин їх виникнення використовує адаптивну підтримку співробітництва користувачів при побудові НЗ. Суть адаптивної підтримки полягає у тому, що користувач отримує динамічний доступ до НЗ, сформованих іншими користувачами групи, до якої він належить. Причому відповідні запити упорядковані або за рангом, який явно було приписано відповідному запиту іншими користувачами групи, або за рангом, який автоматично розраховується, виходячи з міркування – чим частіше запит копіюється при побудові іншого НЗ, тим краще (тим вагоміший його ранг).

Для перевірки вихідних таблиць застосовуються внутрішньотабличні, міжрозрізні та міжтабличні контролі. В роботі запропонована метамова опису таких правил контролю. Приклад опису міжрозрізного контролю на цій метамові наведено нижче:

Розріз: [Відповідні], [По Україні] = [Відповідні], [По АРК Крим, областям, м. Київ і Севастополь]

Контроль:

для всіх рядків

для всіх граф, окрім 20,

комірка = сума (комірок).

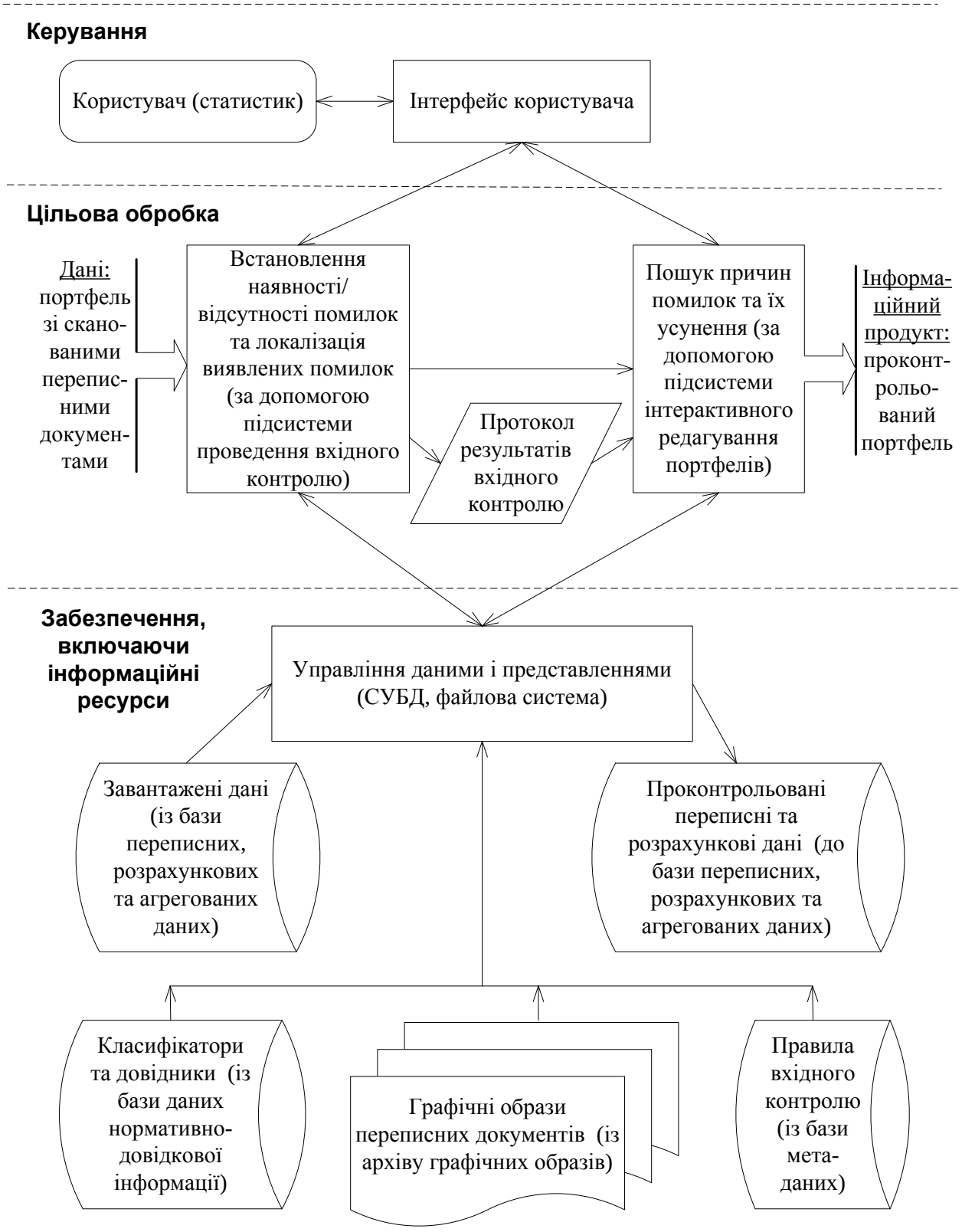

Рисунок 4 – Схема інформаційної технології встановлення наявності і локалізації помилок у первинних даних портфеля

Хоча запропонований формалізований опис правил перевірки вихідних таблиць є досить простим і лаконічним, він дає можливість записати всі контролі зведених даних, які зустрічаються на практиці. Наприклад, у АС «Перепис-2001» за допомогою наведеної метамови для всіх вихідних таблиць були записані відповідні правила перевірки, загальна кількість яких перевищила 200 000.

У табл. 2 для інформаційної технології локалізації помилок у зведених даних і пошуку причин їх виникнення наведена ієрархічна структура, обмежена рівнями етапів і операцій.

Таблиця 2 – Етапи та операції інформаційної технології локалізації помилок у зведених даних і пошуку причин їх виникнення

Етапи	1. Встановлення наявності помилок чи їх відсутності, локалізація виявлених помилок на рівні зведених даних	2. Локалізація виявлених помилок на рівні переписних документів, пошук причин виявлених помилок	3. Забезпечення адаптивності нерегламентних запитів
Операції	1.1. Визначення порядку застосування правил перевірки. 1.2. Застосування правил перевірки. 1.3. Фіксація результатів перевірки	2.1. Перегляд протоколу результатів контролю вихідних таблиць та НЗ. 2.2. Формування/ редагування НЗ. 2.3. Виконання НЗ та перегляд результату	3.1. Призначення групи для користувача. 3.2. Формування даних для забезпечення адаптивності НЗ. 3.3. Адаптивний доступ до НЗ користувачів групи (з можливістю фільтрації за полями запиту)

Як інформаційна технологія встановлення наявності та локалізації помилок у переписних документах з первинними даними, так і інформаційна технологія локалізації помилок і пошуку причин їх виникнення в зведених даних у дисертації подані у вигляді детальної ієрархічної структури за рівнями етапів, операцій та дій.

ВИСНОВКИ

У дисертаційній роботі вирішено актуальну науково-прикладну проблему вироблення єдиних методологічних і технологічних засад створення систем обробки демографічної інформації, забезпечення несуперечності демографічної інформації та підвищення її захищеності через розробку нових моделей, методів, інформаційних технологій та архітектурних рішень для відповідних систем.

При цьому отримані такі нові теоретичні та практичні результати:

1. Введено нове поняття групової анонімності даних, за допомогою якого сформульовані та формалізовані нові класи задач маскування інформації про особливості розподілу групи респондентів: кількісна, концентраційна і концентраційно-різницева. Підготовлено загальну методику забезпечення групової анонімності демографічних даних, що дала можливість розв'язати перелічені задачі на практиці.

2. Уперше розроблено метод забезпечення групової анонімності даних, який базується на застосуванні діадних і недіадних вейвлет-перетворень та полягає в перерозподілі записів вихідного набору даних із комбінаціями значень, що захищаються, відносно різних значень певного параметризуючого атрибута. За допомогою цього метода можна приховати початковий розподіл відповідних комбінацій значень, зберігши при цьому корисність даних шляхом лише пропорційної зміни деталізуючих складових вейвлет-розкладення розподілу.

3. Розроблено нову інформаційну технологію встановлення наявності і локалізації помилок у документах, які містять первинні дані. Головною складовою технології є інтегроване середовище з одночасним контекстним доступом до поточного стану документа, його графічного образу і протоколу із результатами контролю логічної та арифметичної узгодженості значень показників зв'язаних документів. Упровадження даної технології дало змогу пришвидшити в середньому в 3,1 рази виявлення суперечностей у первинних даних пробного перепису населення України 2010 р.

4. Розроблено нову інформаційну технологію локалізації помилок у зведених даних і пошуку причин їх виникнення шляхом створення системи формалізованого опису правил перевірки вихідних таблиць та побудови нерегламентних запитів, що фільтрують і структурують демографічні дані для пошуку джерел виявлених помилок. Реалізація в запропонованій технології адаптивної підтримки співробітництва користувачів дає можливість спростити їм обмін досвідом із побудови відповідних нерегламентних запитів. Упровадження даної технології дало змогу пришвидшити в середньому в 1,5 рази виявлення суперечностей у зведених даних пробного перепису населення України 2010 р.

5. Розроблено функціонально-інформаційну архітектуру статистичної галузі України шляхом розширення моделі Захмана рівнем системної архітектури, який поєднує систему інтелектуального аналізу даних, інформаційно-облікову та інформаційно-аналітичну системи і дає можливість з єдиних взаємодоповнювальних позицій підійти до опису, проектування та реалізації пошуку інформації на всіх ланках таких систем.

6. Удосконалено загальну модель статистичних бізнес-процесів за рахунок її територіального розподілу, розширення процесами знеособлення та забезпечення анонімності даних про респондентів і їх групи, а також виокремлення процесу пошуку інформації, що дало змогу забезпечити єдиний методологічний підхід до автоматизації процесів обробки демографічної інформації, які мають місце в Державній службі статистики України.

7. Розширено практичні можливості моделі метаданих Сунгрена для опису статистичної інформації завдяки додаванню нової групи метаданих, відповідальної за представлення питань забезпечення контролю над відкриттям статистичної інформації, що дало можливість вказані питання враховувати явним чином як при описі демографічних даних, так і під час подальшої реалізації відповідних інформаційних технологій.

8. Практична цінність запропонованого методу забезпечення групової анонімності даних підтверджена розв'язанням різних задач маскування інформації про особливості розподілу окремих груп респондентів на прикладі реальних даних, отриманих під час проведення переписів населення в США, Великій Британії та Італії. У грудні 2012 р. автором був підготовлений захищений мікрофайл за даними першого Всеукраїнського перепису населення. Загалом, результати досліджень були впроваджені в державних службах статистики України і Молдови та лягли в основу автоматизованих систем «Перепис-2001», «Перепис-2001 Аналітик», «Перепис-Молдова 2004» і комплексу електронної обробки даних пробного перепису населення України за 2010 р., за допомогою яких було просканировано, очищено від помилок та оброблено близько 73 млн переписних документів.

Одержані результати можуть бути використані під час проектування та розроблення інформаційних систем з обробки даних другого Всеукраїнського перепису населення та першого Всеукраїнського загального сільськогосподарського перепису, а також аналогічних систем у статистичній галузі України та інших країн.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Чертов О. Р. Архитектура информационных систем по обработке переписных данных / О. Р. Чертов // Информационные технологии: приоритетные направления развития : монография [под общ. ред. О. Р. Чертова]. — Кн. 4. — Новосибирск : СИБПРИНТ, 2010. — С. 76–113.
2. Group methods of data processing : monograph / O. Chertov, D. Tavrov, D. Pavlov, M. Alexandrova, V. Malchikov; ed. O. Chertov. — Raleigh : Lulu.com, 2010. — 155 p.
3. Chertov O. Providing Group Anonymity in Social Networks / Oleg Chertov, Dan Tavrov // Computational social networks. Security and privacy; ed. Ajith Abraham. — London : Springer-Verlag, 2012. — P. 249–268.
4. Chertov O. Fuzzy Clustering with Prototype Extraction for Census Data Analysis / Oleg Chertov, Marharyta Aleksandrova // Soft Computing: State of the Art Theory and Novel Applications. Studies in Fuzziness and Soft Computing; eds. : R. Yager, A. Abbasov, M. Reformat, S. Shahbazova. — Berlin, Heidelberg : Springer-Verlag, 2013. — Vol. 291. — P. 289–313.

5. Лихогруд М. Г. Науково-технічні аспекти створення в Україні єдиної кадастрово-реєстраційної системи / М. Г. Лихогруд, О. Р. Чертов, О. В. Константінов // Землевпорядний вісник. — 2004. — № 1. — С. 4–11.
6. Чертов О. Р. Методи захисту конфіденційної інформації в мікрофайлах / О. Р. Чертов // Вісник Східноукраїнського національного університету ім. В. Даля. — 2009. — № 6 (136), Ч. 1. — С. 16–19.
7. Чертов О. Р. Адаптивная поддержка сотрудничества при поиске информации / О. Р. Чертов, Д. В. Райчук // Штучний інтелект. — 2009. — № 3. — С. 97–104.
8. Чертов О. Р. Учетные и аналитические информационные системы в области статистики населения / О. Р. Чертов // Наукові праці. — (Серія «Комп'ютерні технології»). — Миколаїв : Вид-во ЧДУ ім. Петра Могили, 2010. — Т. 134, Вип. 121. — С. 225–229.
9. Chertov O. R. Models of Statistical Information Description Metadata / O. R. Chertov // Радіоелектронні і комп'ютерні системи. — 2010. — № 3 (44). — С. 81–86.
10. Чертов О. Р. Моделювання метаданих опису статистичної інформації / О. Р. Чертов // Штучний інтелект. — 2010. — № 3. — С. 552–560.
11. Чертов О. Р. Групова анонімність даних / О. Р. Чертов // Збірник наукових праць ВІТІ НТУУ «КПІ». — К. : ВІТІ НТУУ «КПІ», 2010. — Вип. 1. — С. 130–139.
12. Чертов О. Р. Застосування вейвлет-перетворень для забезпечення групової анонімності даних / О. Р. Чертов // Системи обробки інформації : зб. наук. пр. — Х. : ХУ ПС, 2010. — Вип. 3 (84). — С. 90–95.
13. Чертов О. Р. Архітектура систем обробки демографічної інформації / О. Р. Чертов // Реєстрація, зберігання і обробка даних. — 2010. — Т. 12, № 3. — С. 77–84.
14. Чертов О. Р. Інтелектуальний аналіз демографічних даних / О. Р. Чертов // Вісник Тернопільського державного технічного університету. — 2010. — Т. 15, № 3. — С. 132–140.
15. Чертов О. Р. Недіадні одновимірні вейвлет-перетворення / О. Р. Чертов // Наукові вісті НТУУ «КПІ». — 2010. — № 2. — С. 63–73.
16. Чертов О. Р. Застосування недиадних вейвлетів для забезпечення анонімності даних / О. Р. Чертов // Інформаційна безпека. — 2010. — № 2 (4). — С. 96–101.
17. Chertov O. Data group anonymity: general approach / Oleg Chertov, Dan Tavrov // International Journal of Computer Science and Information Security. — 2010. — Vol. 8, № 7. — С. 1–8.
18. Chertov O. R. Providing data group anonymity using concentration differences / Oleg R. Chertov, Danylo Y. Tavrov // Математичні машини і системи. — 2010. — № 3. — С. 34–44.
19. Chertov O. Group anonymity: problems and solutions / Oleg R. Chertov, Danylo Y. Tavrov // Вісник Національного університету «Львівська політехніка». Сер. Інформаційні системи та мережі : зб. наук. пр. — Львів : Вид-во «Львівська політехніка», 2010. — № 673. — С. 3–15.

20. Chertov O. R. Data group anonymity in microfiles / Oleg R. Chertov, Danylo Y. Tavrov // Вісник інженерної академії України. — 2010. — № 2. — С. 159–164.
21. Чертов О. Р. Недиадні вейвлет-перетворення: неперервний випадок / О. Р. Чертов, В. В. Мальчиков // Вісник Східноукраїнського національного університету ім. В. Даля. — 2010. — № 10 (152), Ч. 2. — С. 250–256.
22. Мальчиков В. В. Недиадні вейвлет-перетворення: дискретний випадок / В. В. Мальчиков, О. Р. Чертов // Вісник Харківського національного університету. — (Серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»). — 2010. — № 925, Вип. 14. — С. 140–146.
23. Виявлення аномальної поведінки користувача системи контекстної реклами / О. Р. Чертов, Д. Г. Павлов, В. В. Мальчиков, М. В. Александрова // Штучний інтелект. — 2010. — № 4. — С. 476–483.
24. Чертов О. Р. Виявлення спроб шахрайства в системах контекстної реклами / О. Р. Чертов, Д. Г. Павлов, В. В. Мальчиков // Вісник Східноукраїнського національного університету ім. В. Даля. — 2010. — № 9 (151), Ч. 1. — С. 101–105.
25. Чертов О. Р. Застосування поліномів Кунченка для аналізу статистичних даних / О. Р. Чертов, Д. Ю. Тавров // Східноєвропейський журнал передових технологій. — 2010. — № 4/4 (46). — С. 70–75.
26. Чертов О. Р. Зіставлення з шаблоном на базі поліномів Кунченка в демографічних даних / О. Р. Чертов // Наукові праці. — (Серія «Комп'ютерні технології»). — Миколаїв : Вид-во ЧДУ ім. Петра Могили, 2011. — Т. 160, Вип. 148. — С. 36–43.
27. Чертов О. Р. Необхідні та достатні умови на базис дискретного вейвлет-перетворення / О. Р. Чертов, М. В. Александрова // Вісник Вінницького політехнічного інституту. — 2011. — № 1 (94). — С. 77–83.
28. Чертов О. Р. Апостеріорний метод вибору масштабуючого коефіцієнту вейвлет-перетворення / О. Р. Чертов, В. В. Мальчиков // Вісник Східноукраїнського національного університету ім. В. Даля. — 2011. — № 13 (167). — С. 246–251.
29. Чертов О. Р. Інформаційні технології локалізації і пошуку помилок в первинних та зведених демографічних даних / О. Р. Чертов // Наукові праці. — (Серія «Комп'ютерні технології»). — Миколаїв : Вид-во ЧДУ ім. Петра Могили, 2012. — Т. 191, Вип. 179. — С. 103–110.
30. Чертов О. Р. Мінімізація спотворень при формуванні мікрофайлу з замаскованими даними / О. Р. Чертов // Вісник Східноукраїнського національного університету ім. В. Даля. — 2012. — № 8 (179), Ч. 2. — С. 240–246.
31. Chertov O. Providing online group anonymity / Oleg Chertov, Dan Tavrov // ERCIM News. — 2012. — Vol. 3 (July), N 90. — P. 36–37.
32. Чертов О. Р. Система многомерного анализа данных Всеукраинской переписи населения 2001 года / О. Р. Чертов // Россияне в зеркале

статистики: Всероссийская перепись населения 2002 года : труды междунар. симпозиума, (Москва, 30–31 марта 2004 г.). — М. : Изд-во Федеральной службы государственной статистики, 2004. — С. 234–238.

33. Чертов О. Р. Застосування неперервного вейвлет-перетворення для вибору коефіцієнта масштабування вейвлетів / О. Р. Чертов, В. В. Мальчиков // матеріали XII міжнар. наук. конф. ім. акад. М. Кравчука, (Київ, 15–17 травня 2008 р.). Ч. 1. — К. : ТОВ «Задруга», 2008. — С. 854.
34. Чертов О. Р. Выбор коэффициента масштабирования вейвлетов при кратномасштабном анализе сигналов / О. Р. Чертов, В. В. Мальчиков // Системний аналіз та інформаційні технології (САІТ–2008) : матеріали X міжнар. наук.-техн. конф., (Київ, 20–24 травня 2008 р.). — К. : НТУУ «КПІ», 2008. — С. 151.
35. Райчук Д. В. Засоби адаптації інформаційно-пошукових систем / Д. В. Райчук, О. Р. Чертов // Прикладна математика та комп'ютинг (ПМК–2009) : зб. тез допов. I наукової конф. магістрантів та аспірантів, (Київ, 15–17 квітня 2009 р.). — К. : НТУУ «КПІ», 2009. — С. 67–70.
36. Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // Proc. 2009 Int. Symposium on Computing, Communication and Control (ISCCC 2009), (Singapore, October 9–11, 2009). — Singapore : International Association of Computer Science and Information Technology, 2009. — P. 338–342.
37. Чертов О. Р. Методы сокрытия конфиденциальной информации в статистических данных / О. Р. Чертов, А. И. Пилипюк // Интеллектуальный анализ информации (ИАИ–2009) : сб. трудов IX междунар. науч. конф. им. Т. А. Таран, (Киев, 19–22 мая 2009 г.). — К. : Просвіта, 2009. — С. 420–426.
38. Чертов О. Р. Сравнительный анализ методов ограничения раскрытия информации статистических исследований в микрофайлах / О. Р. Чертов, А. И. Пилипюк // Системний аналіз та інформаційні технології (САІТ–2009) : матеріали XI міжнар. наук.-техн. конф., (Київ, 26–30 травня 2009 р.). — К. : ННК «ІПСА» НТУУ «КПІ», 2009. — С. 406.
39. Chertov O. Group Anonymity / Oleg Chertov, Dan Tavrov // Communications in Computer and Information Science. — 2010. — Vol. 81, Part II. — P. 592–601.
40. Чертов О. Р. Методи статистичного контролю за розкриттям інформації / О. Р. Чертов // Прикладна математика та комп'ютинг (ПМК–2010) : зб. тез допов. II наукової конф. магістрантів та аспірантів, (Київ, 14–16 квітня 2010 р.). — К. : Просвіта, 2010. — С. 403.
41. Чертов О. Р. Збереження діадних фільтрів в тріадному вейвлет-перетворенні / О. Р. Чертов // Сучасні інформаційно-комунікаційні технології (КОМІНФО–2010) : зб. тез. доп. VI міжнар. наук.-техн. конф., (Лівадія, 4–8 жовтня 2010 р.). — К. : ДУІКТ, 2010. — С. 108–110.
42. Чертов О. Р. Групповая анонимность данных / О. Р. Чертов, Д. Ю. Тавров // Интеллектуальный анализ информации (ИАИ–2010) : сб. трудов X междунар.

науч. конф. им. Т. А. Таран, (Киев, 18–21 мая 2010 г.). — К. : Просвіта, 2010. — С. 236–242.

43. Чертов О. Р. Необхідні та достатні умови на базис диадного вейвлет-перетворення / О. Р. Чертов, М. В. Александрова // Інформаційні технології та комп'ютерна інженерія : тези доп. міжнар. наук.-практ. конф., (Вінниця, 19–21 травня 2010 р.). — Вінниця : ВНТУ, 2010. — С. 108–109.
44. Чертов О. Р. Використання недиадних вейвлет-перетворень в статистиці та геофізиці / О. Р. Чертов, В. В. Мальчиков // Матеріали XIII міжнар. наук. конф. ім. акад. М. Кравчука, (Київ, 13–15 травня 2010 р.). Т. 2. — К. : НТУУ «КПІ», 2010. — С. 278.
45. Chertov O. Non-dyadic wavelets for detection of some click-fraud attacks / O. Chertov, V. Malchykov, D. Pavlov // Proc. 2010 Int. Conference on Signals and Electronic Systems (ICSES 2010), (Gliwice, September 7–10, 2010). — Gliwice, Poland : The Institute of Electronics of the Silesian University of Technology, 2010. — P. 401–404.
46. Чертов О. Р. Сохранение периодических свойств данных в процессе их обезличивания / О. Р. Чертов, Д. Ю. Тавров // Интеллектуальный анализ информации (ИАИ–2011) : сб. трудов XI междунар. науч. конф. им. Т. А. Таран, (Киев, 17—20 мая 2011 г.). — К. : Просвіта, 2011. — С. 290–296.
47. Чертов О. Р. Використання методів Data Mining для аналізу даних перепису населення / О. Р. Чертов, М. В. Александрова // Системний аналіз та інформаційні технології (САІТ–2011) : матеріали XIII міжнар. наук.-техн. конф., (Киев, 23–28 травня 2011 р.). — К. : ННК «ІПСА» НТУУ «КПІ», 2011. — С. 339.
48. Чертов О. Р. Алгоритм пошуку закономірностей в даних перепису населення / О. Р. Чертов, М. В. Александрова // Прикладна математика та комп'ютинг (ПМК–2011) : зб. тез допов. III наукової конф. магістрантів та аспірантів, (Київ, 13–15 квітня 2011 р.). — К. : Просвіта, 2011. — С. 345–348.
49. Chertov O. Clustering with Prototype Extraction for Census Data Analysis [Electronic resource] / Oleg Chertov, Marharyta Alexandrova // Proc. World Conf. on Soft Computing (WConSC'11), (San Francisco, May 23–26, 2011). — San Francisco, USA : San Francisco State University, 2011. — № 166–26. — 8 p. — 1 electron. medium (USB flash memory) ; 2 Mb. — Mode of access : http://arxiv.org/pdf/1106.5122v1.
50. Chertov O. Providing group anonymity using wavelet transform / Oleg Chertov, Dan Tavrov // Data security and security data : Proc. 27th British National Conf. on databases (BNCOD 27), (Dundee, June 29 – July 1, 2010). — Berlin, Heidelberg : Springer-Verlag, 2012. — P. 25–36. — (Lecture Notes in Computer Science series, Vol. 6121).
51. Chertov O. Using Association Rules for Searching Levers of Influence in Census Data / Oleg Chertov, Marharyta Aleksandrova // Procedia – Social and Behavioral Sciences; eds. : G. Giannakopoulos, D. Sakas, D. Vlachos, D. Kyriaki-Manessi. — Amsterdam : Elsevier, 2013. — Vol. 73. — P. 475–478.

52. Chertov O. Enterprise architecture model that enables to search for patterns of statistical information / O. Chertov // International Journal of Advanced Research in Artificial Intelligence. — 2013. — Vol. 2, N 6. — P. 1–5.

АНОТАЦІЯ

Чертов О. Р. Моделі, інформаційні технології та архітектура систем обробки демографічної інформації. – На правах рукопису.

Дисертація на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Інститут проблем математичних машин і систем НАН України, Київ, 2013.

Дисертація присвячена вирішенню важливої науково-прикладної проблеми вироблення єдиних методологічних і технологічних засад створення систем обробки демографічної інформації, забезпечення несуперечності демографічної інформації та підвищення її захищеності через розробку нових моделей, методів, інформаційних технологій та архітектурних рішень для відповідних систем.

Розроблено нові інформаційні технології, які забезпечують пошук помилок у первинних і зведених статистичних даних.

Основні результати роботи отримали практичне застосування в статистичних галузях України і Молдови при розробленні цілої низки інформаційних систем національного масштабу.

Ключові слова: інформаційна технологія, демографічна інформація, функціонально-інформаційна архітектура, модель Захмана, статистичні бізнес-процеси, метадані, групова анонімність даних, вейвлет-перетворення, мікрофайл.

АННОТАЦИЯ

Чертов О. Р. Модели, информационные технологии и архитектура систем обработки демографической информации. – На правах рукописи.

Диссертация на соискание научной степени доктора технических наук по специальности 05.13.06 – информационные технологии. – Институт проблем математических машин и систем НАН Украины, Киев, 2013.

Диссертация посвящена решению важной научно-прикладной проблемы выработки единых методологических и технологических основ создания систем обработки демографической информации, обеспечения непротиворечивости демографической информации и повышения ее защищенности посредством

разработки новых моделей, методов, информационных технологий и архитектурных решений для соответствующих систем.

Введено и обосновано понятие групповой анонимности данных, которое, в отличие от обезличивания данных об отдельных респондентах, используемого в настоящее время в статистической отрасли, позволяет сформулировать задачу сокрытия информации об особенностях распределения группы респондентов и формализовать ее решение. Поставлены количественная, концентрационная и концентрационно-разностная задачи обеспечения групповой анонимности данных, предложен метод их решения на основе применения диадных и недиадных вейвлет-преобразований, заключающийся в перераспределении записей первоначального набора данных с защищаемыми комбинациями значений относительно различных значений определенного параметризирующего атрибута, ответственного за распределение данных в пространстве, во времени или по каким-то другим категориям. Приведены детальные примеры и рассмотрены практические аспекты применения предложенного метода к решению задачи обеспечения групповой анонимности данных в микрофайлах в случае количественного, концентрационного и концентрационно-разностного целевых сигналов. Изложение выполнено на основе анализа реальных данных, полученных во время проведения переписей населения в США, Великобритании и Италии соответственно.

Усовершенствована модель Захмана функционально-информационной архитектуры за счет ее расширения уровнем системной архитектуры, который отвечает, в частности, и за интеллектуальный анализ данных, обеспечивая поиск скрытых закономерностей распределения данных. Это позволило применить данную модель для описания статистической отрасли Украины при проектировании национальных систем обработки демографической информации. Усовершенствована известная общая модель статистических бизнес-процессов путем ее территориального распределения и выделения процесса поиска. Данный процесс поддерживается не только на уровне информационно-учетной системы (с помощью нерегламентированных запросов) или в информационно-аналитической системе (посредством аналитических запросов), но и в ходе применения интеллектуального анализа данных для автоматизированного поиска в них определенных закономерностей. Дополнительно в модель статистических бизнес-процессов были введены новые процессы, обеспечивающие обезличенность данных о респондентах, их индивидуальную и групповую анонимность. Получила дальнейшее развитие модель метаданных Сунгрена для описания статистической информации, в которой, в отличие от существующих, сформирована дополнительная группа метаданных индивидуальной и групповой анонимности. Благодаря этому данная модель может быть использована как при описании демографических данных, так и в ходе дальнейшей реализации соответствующих статистических процессов. Все это позволяет строить более полные и точные описания процессов обработки демографической информации.

Разработаны новые информационные технологии, обеспечивающие поиск ошибок в первичных и сводных статистических данных. В информационной

технологии локализации ошибок и поиска причин их возникновения в сводных данных (выходных таблицах) организована адаптивная поддержка сотрудничества пользователей, что упрощает обмен опытом относительно построения нерегламентированных запросов. Информационная технология установления наличия и локализации ошибок в документах, которые содержат первичные данные, позволила существенно повысить эффективность работы пользователей за счет создания интегрированной среды, объединяющей текущее состояние документа, его графический образ и протокол с результатами контроля. Внедрение данных технологий позволило ускорить обнаружение и исправление ошибок в среднем в 3,1 раза в первичных данных и в 1,5 раза в сводных данных пробной переписи населения Украины 2010 г.

Основные результаты работы нашли практическое применение в статистических отраслях Украины и Молдовы при проектировании и разработке целого ряда информационных систем национального масштаба, с помощью которых было просканировано и обработано около 73 млн переписных документов.

Ключевые слова: информационная технология, демографическая информация, функционально-информационная архитектура, модель Захмана, статистические бизнес-процессы, метаданные, групповая анонимность данных, вейвлет-преобразования, микрофайл.

ABSTRACT

Chertov O. R. Models, information technologies and architecture of demographic information processing systems. – Manuscript.

Thesis for a doctor's technical science degree majoring in 05.13.06 – information technologies. – Institute of mathematical machines and systems problems of the National Academy of Sciences of Ukraine, Kyiv, 2013.

The thesis is devoted to solving an important scientific and applied problem of making of uniform methodological and technological fundamentals of demographic information processing systems, demographic information coherency control and its higher protection by developing new models, methods, information technologies, and architecture solutions for such systems.

New information technologies that enable searching for errors in primary and aggregated statistical data alike are developed.

Main results of the work have been successfully implemented in statistical branches of Ukraine and Moldova while developing a series of national information systems.

Keywords: information technology, demographic information, functional and information architecture, Zachman Framework, statistical business processes, metadata, data group anonymity, wavelet transforms, microfile.