УДК 004.9

Чертов О. Р.

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ЛОКАЛІЗАЦІЇ І ПОШУКУ ПОМИЛОК У ПЕРВИННИХ ТА ЗВЕДЕНИХ ДЕМОГРАФІЧНИХ ДАНИХ

В статті описані інформаційні технології локалізації і пошуку помилок, що містяться в первинних та зведених демографічних даних. Викладення базується на успішному досвіді впровадження зазначених технологій в Україні та закордоном.

Ключові слова: *інформаційна технологія; демографічні дані; перепис.*

В статье описаны информационные технологии локализации и поиска ошибок, которые встречаются в первичных и сводных демографических данных. Изложение базируется на успешном опыте внедрения указанных технологий в Украине и за рубежом.

Ключевые слова: *информационная технология; демографические данные; перепись.*

The article describes information technologies of localization and search for errors that occur in primary and summary demographic data. The presentation is based on the successful experience of implementing these technologies in Ukraine and abroad.

Key words: *information technology, demographic data, census.*

Вступ

Сучасні системи обробки демографічної інформації характеризуються дуже великими обсягами даних, що оброблюються, та значною кількістю різнорідних правил контролю первинних та зведених даних [1, с. 76–113; 2]. Тому швидкість отримання результатів перепису населення чи іншого відповідного статистичного спостереження багато в чому залежить від застосовуємих інформаційних технологій локалізації та пошуку помилок в статистичних даних.

Загалом, поняття «інформаційна технологія» визначається в низці законодавчих та нормативних документів. Найбільш точне та лаконічне визначення міститься в ДСТУ 2226-93 [3]: «інформаційна технологія – технологічний процес, предметом перероблення й результатом якого є інформація». Більш розлого це поняття визначено в [4]: «інформаційна технологія – цілеспрямована організована сукупність інформаційних процесів з використанням засобів обчислювальної техніки, що забезпечують високу швидкість обробки даних, швидкий пошук інформації, розосередження даних, доступ до джерел інформації незалежно від місця їх розташування».

По аналогії з технологіями матеріального виробництва кожну інформаційну технологію можна розглядати як процес, що використовує сукупність засобів і методів збору, оброблення та передачі даних для отримання інформаційного продукту – інформації нової якості про стан об'єкту, процесу чи явища (див. рис.1) [5, с. 87]. Як зазначено в [6, с. 6], під інформаційними технологіями «починають розуміти будь-які засоби трансформації даних в корисний для досягнення мети системи управління інформаційний продукт».

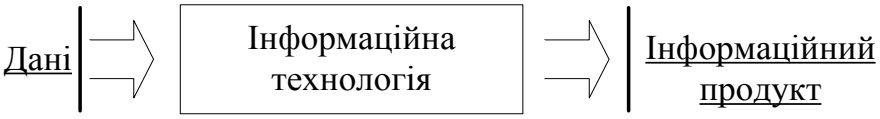

Рис. 1. Інформаційна технологія як процес перетворення даних в інформаційний продукт

Для ефективної організації всіх процесів інформаційної технології вимагається «визначення необхідних робіт (їх відповідного структурування, ув'язування за входом, виходом, термінами реалізації, виконавцями)» [6, с. 90]. Інформаційну технологію прийнято представляти у вигляді ієрархічної структури за різними рівнями [5, с. 91]: від більш загальних до більш детальних, наприклад, за рівнями етапів, операцій та дій.

З точки зору загального менеджменту виділяють три послідовні етапи процесу контролю [7, с. 440 – 449; 8, с. 83–85]:

1) встановлення стандартів;
2) зіставлення досягнутих результатів з встановленими стандартами;
3) прийняття необхідних коригуючих мір.

Враховуючи специфіку обробки демографічних первинних та зведених даних, зазначені етапи можна конкретизувати так, як це показано в табл. 1.

Етапи, виділені курсивом в табл. 1, є ключовими для пришвидшення виявлення та виправлення помилок в даних, оскільки ці етапи не можна повністю автоматизувати.

Таблиця 1

Етапи та особливості контролю демографічних даних

Загальні етапи процесу контролю	Етапи контролю демографічних даних	Особливості етапів контролю демографічних даних	
		первинних	зведених
встановлення стандартів	визначення правил контролю	визначення внутрішньобланкових та міжбланкових контролів	визначення внутрішньотабличних, міжрозрізних та міжтабличних контролів
зіставлення досягнутих результатів з встановленими стандартами	встановлення наявності помилок чи їх відсутності	в основному, арифметичний та логічний контроль	як правило, арифметичний контроль
	локалізація помилок, якщо вони виявлені	природно обмежується одним чи декількома (на рівні домогосподарства) бланками	як правило, дуже складно локалізувати помилку до рівня відповідного переписного документу
прийняття необхідних коригуючих заходів	*пошук причин виявлених помилок та їх усунення*	протокол контролю є основним джерелом інформації про помилку	помилки заборонено виправляти на зведених (агрегованих) даних

Класична технологія організації пошуку помилок в статистичних даних описана в [9, с. 91] в п. 3.4.3 «Організація автоматизованого розв'язання задач з використанням АРМ економіста-статистика»:

«… виконується перевірка правильності введеної інформації з використанням арифметичного і логічного контролю за формулами, які задаються економістами. Внаслідок контролю утворюється файл протоколу, який містить інформацію про виявлені помилки. Його можна переглянути на екрані або роздрукувати. Після аналізу помилок звіти можуть бути відкориговані.».

Дана технологія до початку XXI століття фактично була єдино можливою, але застосування для введення статистичних даних швидкодіючих промислових сканерів дозволяє організувати процес локалізації та пошуку помилок в первинних даних принципово більш зручним та ефективним для користувачів способом, який описано в наступному розділі.

Локалізація причини помилки в зведених даних, наприклад, коли загальна кількість респондентів певної групи в різних вихідних таблицях не співпадає, є нетривіальною справою, яка вимагає певних навичок та досвіду в побудові відповідних нерегламентних запитів до бази даних. Тому організація обміну досвідом між користувачами, що займаються пошуком помилок в зведених даних, може значно прискорити весь процес обробки наявних даних.

Постановка задачі

Об'єктом дослідження є процеси локалізації та пошуку помилок в демографічних даних – первинних (переписних бланках чи інших статистичних анкетах) та зведених (вихідних таблицях), що породжують проблему необхідності розширення функціональних можливостей інформаційних систем, які реалізують зазначені процеси. Предметом дослідження є інформаційні технології, які забезпечують організацію відповідної обробки демографічної інформації. Мета статті полягає в узагальнені та формалізації опису зазначених інформаційних технологій, впроваджених чи впроваджуємих за безпосередньої участі автора в різноманітних системах обробки даних статистики населення.

Інформаційна технологія локалізації і пошуку помилок в первинних даних. Розглянемо, як в автоматизованих системах «Перепис-2001» та «Перепис-Молдова 2004», які забезпечували обробку даних відповідно першого Всеукраїнського перепису населення 2001 р. та перепису населення Республіка Молдова в 2004 р., здійснювалися локалізація і пошук помилок під час вхідного контролю, тобто під час контролю первинних даних переписних документів [1, с. 95 – 102].

На рис. 2 представлена загальна схема технологічного процесу локалізації та пошуку помилок в первинних даних портфеля з переписними бланками.

Якщо в підсистемі проведення вхідного контролю в процесі контролю завантаженого без помилок портфеля в хоча б одному з переписних документів виявлено хоча б одну помилку, то відповідне поле в цьому документі, цей документ, а також усі документи, до яких цей документ належить, помічаються відповідними кольоровими позначками як помилкові. Ознайомившись з протоколом результатів вхідного контролю, користувач редагує помилкові переписні документи, які зберігаються в базі переписних, розрахункових та агрегованих даних. Під час редагування переписного документа користувач може бачити (за допомогою модуля інтерактивного і контекстного доступу до графічного образу просканованого документу) зображення графічного образу цього документа для порівняння змісту першоджерела з електронною копією, а також за допомогою модулів інтерактивного і контекстного доступу до довідників та класифікаторів і до протоколу помилок має контекстний доступ до відповідної інформації.

Фатальні помилки (червоний колір підсвічення помилкових переписних документів та показників на бланках вхідних форм в інтерфейсі користувача) необхідно обов'язково виправити. Для цього потрібно проаналізувати помилковий показник і поточну ситуацію (тобто проаналізувати залежність між показниками), при необхідності, звіривши введені переписні документи з їх графічними образами з архіву графічних образів. В деяких випадках фатальна помилка може вказувати на ситуацію, яка існує в реальному житті. В такому випадку необхідно зняти помилку.

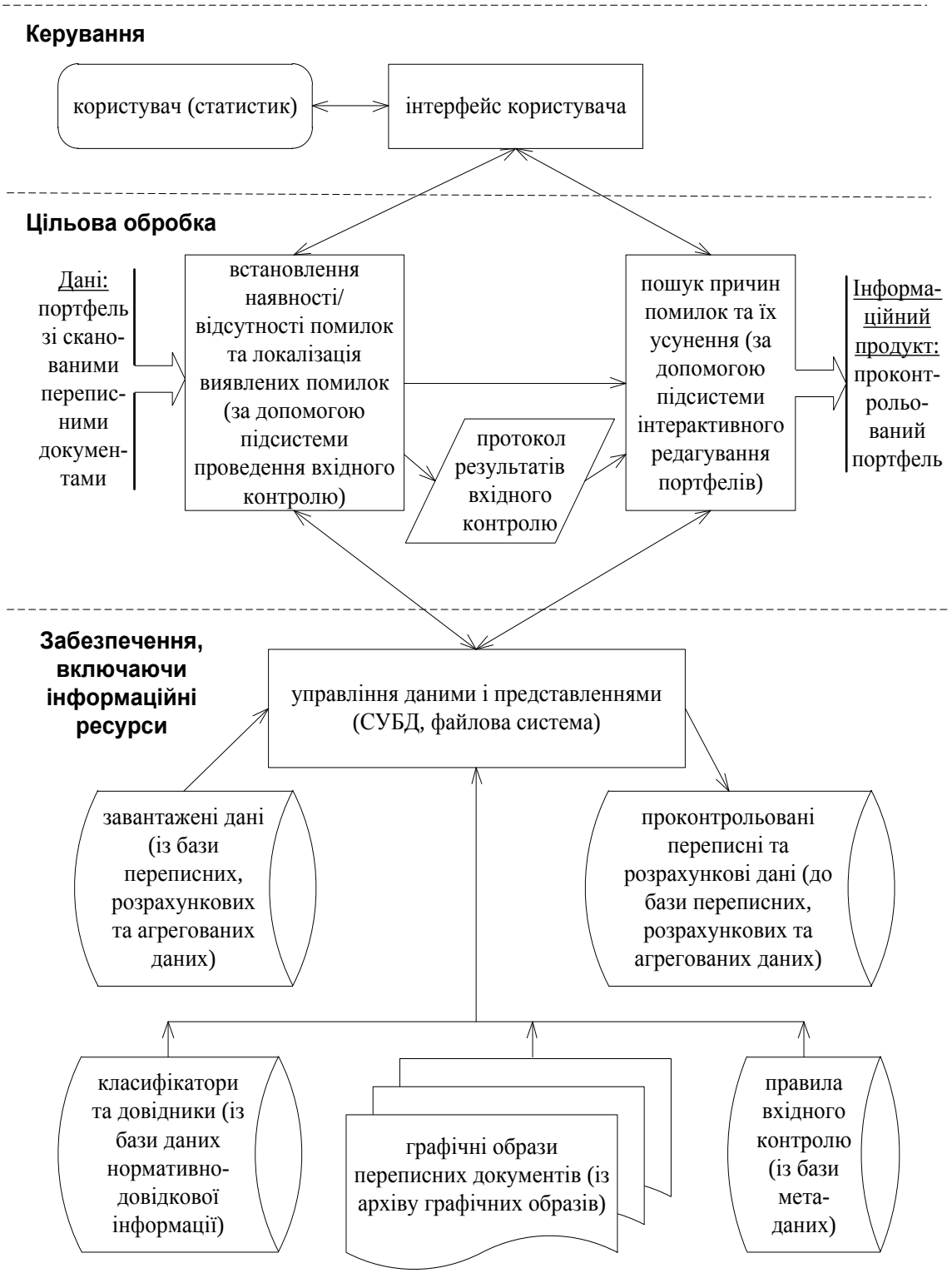

Рис. 2. Схема технологічного процесу локалізації та пошуку помилок в первинних даних портфелю

Нефатальна помилка (блакитний колір підсвічення помилкових переписних документів та показників на бланках вхідних форм в інтерфейсі користувача) вказує на таку ситуацію, що в загальному випадку може і не бути помилкою, і після звірення введених документів з їх графічними образами та аналізу показників не потребує коригування. Після виправлення помилок користувач знову проводить контроль обраних переписних документів.

В табл. 2 наведено формалізоване представлення інформаційної технології локалізації і пошуку помилок в переписних документах по первинним даним у вигляді ієрархічної структури за рівнями етапів, операцій та дій.

Таблиця 2

Етапи, операції та основні дії інформаційної технології локалізації і пошуку помилок в переписних документах по первинним даним

Етапи					
1. Встановлення наявності помилок чи їх відсутності		2. Локалізація помилок, якщо вони виявлені		3. Пошук причин виявлених помилок та їх усунення	
Операції	**Дії**	**Операції**	**Дії**	**Операції**	**Дії**
Визначення порядку застосування правил перевірки	Отримання із бази метаданих загального порядку застосування правил перевірки та рівня їх критичності	Зіставлення виявлених помилок і показників переписних документів	Отримання із бази метаданих відповідності між помилками і показниками, які можуть їх викликати	Перегляд переписних документів портфелю та результатів вхідного контролю	Перегляд певного переписного документу, значення показників котрого відмарковані відповідно до виявлених помилок
			Фіксація для виявленої помилки відповідного переписного документу		
	Вибір поточного правила перевірки в залежності від результатів попередніх перевірок		Фіксація для виявленої помилки відповідного аркушу переписного документу		Проглядання дерева перегляду переписних документів портфелю, відмаркованих до виявлених помилок
Застосування правил перевірки	Контроль повноти даних (по окремих переписних формах)	Маркування відсутності помилок чи виявлених помилок	Фіксація для виявленої помилки відповідного графічного образу переписного документу	Редагування переписного документу з метою усунення помилок	Контекстний доступ до графічного образу переписного документу
	Контроль форматів даних		Маркування в дереві перегляду позначками синього кольору переписних документів, значення показників котрих викликали нефатальні помилки		Контекстний доступ до протоколу результатів вхідного контролю
	Арифметичний контроль значень даних				Контекстний доступ до нормативно-довідкової інформації
	Логічний контроль значень та наявності даних				Зміна значення конкретного показника переписного документу
	Контроль відповідності даних класифікаторам та довідникам		Маркування в дереві перегляду позначками червоного кольору переписних документів, значення показників котрих викликали фатальні помилки		Маркування в дереві перегляду жовтим кольором переписного документу, в якому було змінено значення показника
Фіксація результатів перевірки з урахуванням критичності помилки	Фіксація результату перевірки — має місце відповідна помилка чи ні	Формування протоколу результатів вхідного контролю	Маркування синім кольором значень показників, що викликали нефатальні помилки	Зняття фатальної помилки контролю	Маркування жовтим кольором зміненого значення показника
	Визначення рівня критичності помилки, якщо він не є фіксованим		Маркування червоним кольором значень показників, що викликали фатальні помилки		Автокоригування значень показників переписного документа
			Запис в протокол контролю повідомлення про виявлену помилку та показники, значення котрих її викликало		Збереження відредагованих (чи всіх поточних) значень показників
			Запис в протокол контролю загальної інформації про виявлені помилки чи їх відсутність		Вибір в протоколі вхідного контролю фатальної помилки, яка по факту наявних даних не є помилкою, та зняття її
					Фіксація інформації про зняту фатальну помилку

Інформаційна технологія локалізації і пошуку помилок в зведених даних (вихідних таблицях)

Пошук причини помилки, виявленої на етапі проведення контролів зведених даних, зокрема, вихідних таблиць (ВТ), на практиці може бути виконаний лише за допомогою підсистеми підтримки нерегламентних запитів, яка дозволяє за певним розрізом отримати визначений розподіл даних з БД.

З часом, з накопиченням досвіду роботи ряд користувачів придумують певні шаблонні нерегламентні запити, котрі дозволяють швидко знаходити типові причини, які приводять до появи помилок у вихідних таблицях під час їх перевірки за допомогою внутрішньотабличних, міжрозрізних чи міжтабличних контролів. Тому сильно зростає значення підсистеми підтримки адаптивності, котра як би бере на себе роль по передачі досвіду з побудови нерегламентних запитів.

Ключовою особливістю застосування адаптивних технологій є можливість проведення аналізу діяльності великої кількості користувачів одночасно. Результати цього аналізу можуть використовуватися при налагодженні функціональності для конкретного

користувача. Адаптація на основі колективної взаємодії значно привабливіша за адаптацію на основі тільки особистого досвіду, оскільки в будь-який момент часу можна знайти готове рішення, прийняте на базі вже накопиченого досвіду [10, с. 99].

Розглянемо, як в автоматизованих системах з обробки демографічних даних, розроблених під керівництвом автора, здійснювалися локалізація і пошук помилок в зведених даних.

На рис. 3 представлена загальна схема технологічного процесу локалізації та пошуку помилок в зведених даних.

Якщо в підсистемі проведення контролю матриць вихідних таблиць в процесі контролю зведених даних по якомусь об'єкту адміністративно-територіальному устрою виявлено хоча б одну помилку, то відповідна інформація записується до протоколу результатів контролю матриць вихідних таблиць.

Але на відміну від аналогічної ситуації під час обробки первинних даних протокол результатів контролю зведених даних не дозволяє локалізувати помилку безпосередньо в переписних документах. Тому як для локалізації зазначених помилок, так і для пошуку причин їх виникнення потрібно застосовувати підсистему підтримки нерегламентних запитів.

Дана підсистема повинна реалізовувати наступні функції: формування та зберігання нового запиту, редагування, копіювання, вилучення чи виконання існуючого запиту.

Рис. 3. Схема технологічного процесу локалізації та пошуку помилок в зведених даних

Під час формування нового нерегламентного запиту відбувається автоматичне визначення автора запиту та дати і часу формування запиту. Окрім того, вручну вводиться опис запиту (шифр запиту та анотація дій, що будуть виконуватися).

Кожен нерегламентний запит повинен містити, як мінімум, одне поле із хоча б однієї таблиці БД, що зберігають наступні дані: класифікатори і локальні довідники, первинні дані, тобто дані переписних документів, та розрахункові показники, агреговані дані.

Приклад сформованого в системі «Перепис-2001» нерегламентного запиту для пошуку респондентів, що не вказали свою національність в переписній формі (бланку) 2С, наведено на рис. 4. Фактично підсистема підтримки нерегламентних запитів дозволяє нефахівцеві в області програмування візуально побудувати SQL-запит до БД.

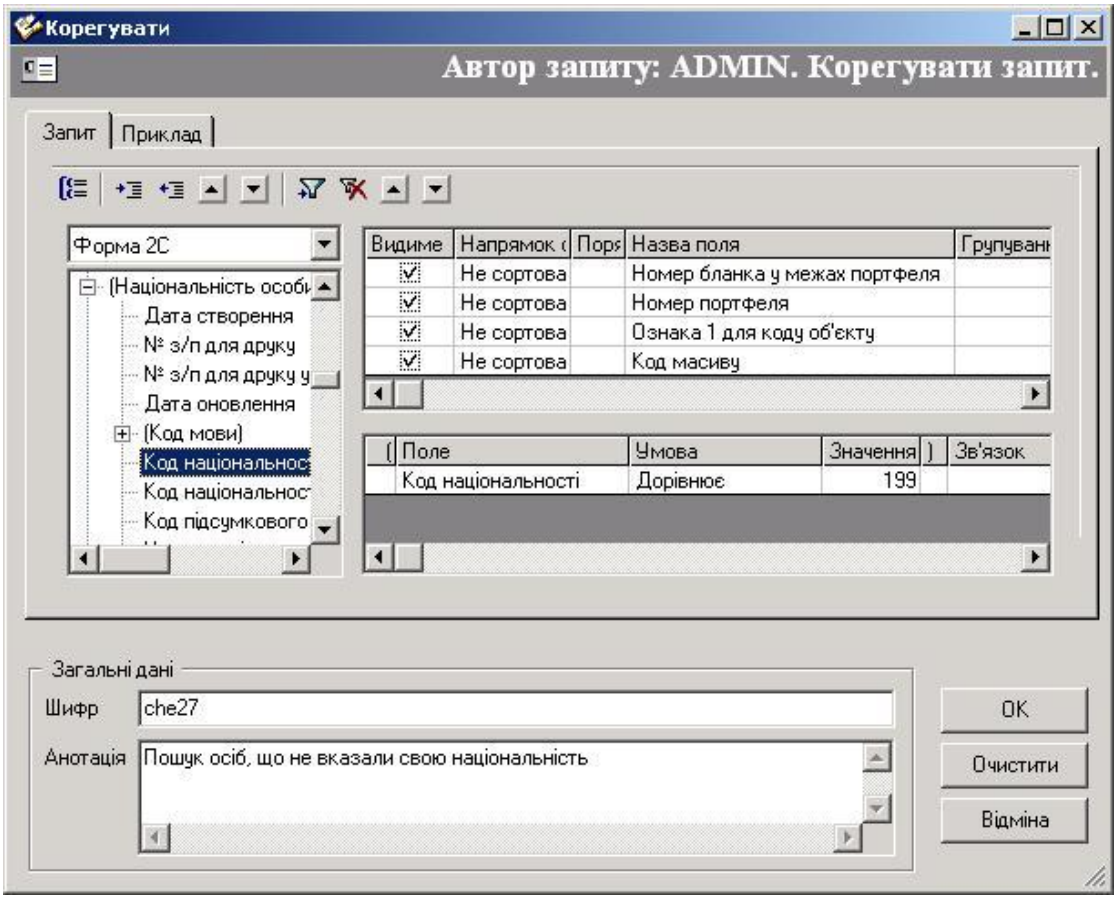

Рис. 4. Приклад нерегламентного запиту в системі «Перепис-2001»

Результат виконання нерегламентного запиту записується до текстового файлу чи видається на екран користувача.

Оскільки заборонено коригувати зведені дані, то за наявності помилок в вихідних таблицях пошук можливих причин цих помилок в первинних даних за допомогою підсистеми підтримки нерегламентних запитів ітеративно чергується з внесенням відповідних змін до первинних документів за допомогою підсистеми інтерактивного редагування портфелів, поки всі помилки в вихідних таблицях не будуть виправлені.

В загальному випадку причиною спрацьовування внутрішньотабличного, міжтабличного чи міжрозрізного контролю, можуть бути наступні помилки:

– помилка, викликана наявністю знятої фатальної помилки;

– помилка, пропущена під час верифікації (на етапі сканування переписних документів) та/чи під час вхідних контролів через похибку в постановці задачі;

– помилка, викликана реальним існуванням певної комбінації значень показників переписних документів, не передбаченої в постановці задачі;

– помилка в програмній реалізації (інформаційної системи обробки переписних даних, СКБД тощо).

Хоча потенційно помилок в зведених даних набагато менше, ніж в первинних даних, але їх локалізація та пошук причин виникнення – значно складніші. Застосування адаптивної підтримки співробітництва користувачів [10] під час побудови нерегламентних запитів за допомогою підсистеми підтримки адаптивності дозволяє як полегшити новачкам початок їх фахової роботи з пошуку причин помилок у вихідних таблицях, так і покращити взаємодію та обмін досвідом інших фахівців, зокрема, за рахунок надання позитивних відгуків на певні нерегламентні запити.

В табл. 3 наведено формалізоване представлення інформаційної технології локалізації і пошуку помилок в зведених даних із забезпеченням адаптивної підтримки співробітництва користувачів у вигляді ієрархічної структури за рівнями етапів, операцій та дій.

Таблиця 3.

Етапи, операції та основні дії інформаційної технології локалізації і пошуку помилок в зведених даних із забезпеченням адаптивної підтримки співробітництва користувачів

<table>
<tr><td colspan="6">Етапи</td></tr>
<tr><td colspan="2">1. Встановлення наявності (відсутності) помилок, локалізація виявлених помилок на рівні зведених даних</td><td colspan="2">2. Локалізація виявлених помилок на рівні переписних документів, пошук причин виявлених помилок</td><td colspan="2">3. Забезпечення адаптивності нерегламентних запитів</td></tr>
<tr><td>Операції</td><td>Дії</td><td>Операції</td><td>Дії</td><td>Операції</td><td>Дії</td></tr>
<tr><td rowspan="2">Визначення порядку застосування правил перевірки</td><td>Отримання із бази метаданих загального порядку застосування правил перевірки</td><td rowspan="3">Перегляд нерегламентних запитів та протоколу результатів контролю ВТ</td><td>Перегляд певного нерегламентного запиту</td><td rowspan="4">Призначення групи для користувача</td><td rowspan="2">Автоматичний перерозподіл користувачів по групам</td></tr>
<tr><td>Вибір поточного правила перевірки в залежності від результатів попередніх перевірок</td><td>Проглядання дерева перегляду нерегламентних запитів</td></tr>
<tr><td rowspan="2">Застосування правил перевірки</td><td>Внутрішньотабличний контроль</td><td>Контекстний доступ до протоколу результатів контролю ВТ</td><td rowspan="2">Ручне призначення групи (наприклад, для новачків)</td></tr>
<tr><td>Міжрозрізний контроль</td><td rowspan="3">Формування/ редагування нерегламентного запиту</td><td>Копіювання існуючого нерегламентного запиту</td></tr>
<tr><td>Міжтабличний контроль</td><td>Визначення службових реквізитів запиту (автор запиту, дата і час формування чи зміни запиту тощо)</td><td rowspan="3">Формування даних для забезпечення адаптивності нерегламентних запитів</td><td>Автоматичний підрахунок рангу запита</td></tr>
<tr><td>Фіксація результатів перевірки</td><td>Фіксація результату перевірки — чи має місце відповідна помилка</td><td>Визначення для запиту поля з таблиці БД, що зберігає класифікатори і локальні довідники, дані переписних документів та розрахункові показники чи агреговані дані</td><td>Ручне визначення користувачем рангу поточного нерегламентного запиту</td></tr>
<tr><td rowspan="3"></td><td>Фіксація комірки, (або рядка чи графи), розрізу і таблиці, в яких була виявлена помилка</td><td rowspan="5"></td><td>Редагування атрибутів поля запиту («Видиме», «Напрямок сортування», «Порядок сортування»)</td><td></td><td>Підрахунок середнього часу на пошук помилки в одній ВТ із певного класу ВТ конкретним користувачем</td></tr>
<tr><td>Маркування в дереві перегляду розрізів ВТ, які пройшли контролі (з помилками чи без них — відповідно)</td><td>Визначення функції агрегування при групуванні по полю запита («Група», «Найбільше», «Найменше», «Середнє», «Кількість»))</td><td rowspan="4">Адаптивний доступ до нерегламентних запитів користувачів групи</td><td rowspan="2">Впорядкування запитів за автоматично сформованими рангами</td></tr>
<tr><td rowspan="2">Формування протоколу результатів контролю ВТ</td><td rowspan="2">Запис в протокол контролю повідомлення про виявлену помилку та розташування в розрізі ВТ помилкового значення</td><td>Редагування списку умов, що накладаються на поле запиту (з визначенням атрибутів «Поле», «Умова», «Значення», «Зв'язок» та дужок)</td></tr>
<tr><td>Зберігання чи вилучення запиту</td><td>Впорядкування запитів за рангами, наданими користувачами групи</td></tr>
<tr><td>Запис в протокол контролю загальної інформації про виявлені помилки чи їх відсутність</td><td rowspan="2">Виконання нерегламентного запиту та перегляд результату</td><td>Запуск нерегламент-ного запиту на виконання</td><td rowspan="2">Впорядкування запитів за особисто визначеними рангами</td></tr>
<tr><td>Перегляд результату виконання запиту</td></tr>
</table>

Висновки. Системи обробки демографічної інформації, насамперед, – переписних даних, є великими масштабними інформа-ційними системами, в яких отримання остаточних результатів перепису може тривати роками. Ефективна реалізація локалізації та пошуку помилок в первинних і зведених даних, тобто найбільш трудомістких операцій під час обробки даних перепису, може значно покращити кінцеві результати та прискорити терміни їх отримання. Наприклад, описані в статті відповідні інформаційні технології дозволили ліквідувати більш ніж трьохмісячне відставання, що виникло із-за організаційних причин під час обробки даних Всеукраїнського перепису населення 2001 р.

Запропонований в статті формалізований опис інформаційних технологій локалізації та пошуку помилок дозволив з єдиних позицій підійти до їх реалізації як при обробці первинних, так і зведених демографічних даних. Відповідні інформаційні технології наразі застосовуються в Державній службі статистики України для обробки даних пробного перепису населення України 2010 р. про 98,5 тисяч респондентів Дергачівського району Харківської області.

ЛІТЕРАТУРА

1. Информационные технологии: приоритетные направления развития : монография / [Л. Н. Абуталипова, А. Г. Гусейнов, А. С. Дулесов и др.] ; под общ. ред. О. Р. Чертова. – Новосибирск : СИБПРИНТ, 2010. – Кн. 4. – 194 с.
2. Чертов О. Р. Учетные и аналитические информационные системы в области статистики населения / О. Р. Чертов // Наукові праці. – Миколаїв : Вид-во ЧДУ ім. Петра Могили, 2010. – Т. 134. Вип. 121. Комп'ютерні технології. – С. 225—229.
3. Автоматизовані системи. Терміни та визначення : ДСТУ 2226-93. – К. : Держстандарт України, 1994. – 94 с.
4. Про Національну програму інформатизації [Текст]: закон України від 4 лютого 1998 року № 74/98-ВР // Відомості Верховної Ради України. – 1998. – № 27. – С. 181.
5. Информатика : Учебник / [Макарова Н. В., Матвеев Л. А., Бройдо В. Л. и др.] ; под ред. Н. В. Макаровой. – [3-е изд., перераб.]. – М. : Финансы и статистика, 2009. – 768 с.
6. Павлов А. А. Информационные технологии и алгоритмизация в управлении / А. А. Павлов, С. Ф. Теленик. – К. : Техніка, 2002. – 344 с.
7. Мескон М. Х. Основы менеджмента : пер. с англ. / М. Х. Мескон, М. Альберт, Ф. Хедоури. – [2-е изд.]. – М. : Дело, 2001. – 800 с.
8. Кабушкин Н. Л. Основы менеджмента : учеб. пособие / Н. Л. Кабушкин. – [11-е изд., испр.]. – М. : Новое знание, 2009. – 336 с.
9. Годун В. М. Інформаційні системи і технології в статистиці : навч. посібник / В. М. Годун, Н. С. Орленко, М. А. Сендзюк ; за ред. В. Ф. Ситника. – К. : КНЕУ, 2003. – 267 с.
10. Чертов О. Р. Адаптивная поддержка сотрудничества при поиске информации / О. Р. Чертов, Д. В. Райчук // Штучний інтелект. – 2009. – № 3. – С. 97—104.

Дата надходження статті до редколегії 03.05.2012 р.

ЧЕРТОВ О. Р. – к.т.н., доцент, доцент кафедри прикладної математики Національного Технічного Університету України «Київський політехнічний інститут».