УДК 314.1, 004.62

***Чертов О.Р.***

# УЧЕТНЫЕ И АНАЛИТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ В ОБЛАСТИ СТАТИСТИКИ НАСЕЛЕНИЯ

*У зв'язку з початком нового десятилітнього раунду проведення переписів актуальним є проробка питань розробки статистичних інформаційних систем. В статті проаналізовано їх функціональне призначення, відмінність між переписними обліковими і аналітичними інформаційними системами. Виділено новий архітектурний рівень таких систем, відповідальний за інтелектуальний аналіз статистичних даних.*

***Ключові слова:*** *перепис, інформаційно-облікова система, інформаційно-аналітична система, інтелектуальний аналіз даних.*

*В связи с началом нового десятилетнего раунда проведения переписей актуальным является проработка вопросов разработки статистических информационных систем. В статье проанализировано их функциональное назначение, отличия между переписными учетными и аналитическими информационными системами. Выделен новый архитектурный уровень таких систем, ответственный за интеллектуальный анализ статистических данных.*

***Ключевые слова:*** *перепись, информационно-учетная система, информационно-аналитическая система, интеллектуальный анализ данных.*

*The start of new decennial census round turns the issues of statistical information systems development into an urgent problem. In the paper, their functionalities and the differences between census OLTP and OLAP systems are analyzed. Besides, we mark out the novel architectural level which is responsible for data mining of statistical data.*

***Key words:*** *census, informative-registration system; informative-analytic system; intellectual analysis of data.*

## *ВВЕДЕНИЕ*

С 60-х годов XX столетия по рекомендациям ООН абсолютное большинство стран мира проводят переписи населения циклически – раундами каждые 10 лет. Переписи последнего цикла проводились, начиная с 1999 г. Решение о проведении переписи населения в сроки, максимально близкие к 2010 г., было принято в Минске 28 ноября 2006 г. на Совещании глав государств Содружества Независимых Государств. Так, в Российской Федерации ближайшая перепись населения запланирована на 2010 г., в Украине – на 2011 г. Поэтому важной и актуальной является задача проработки вопросов построения информационных систем, предназначенных для обработки данных статистики населения и, прежде всего, – переписных данных.

Изложенные в данной статье идеи базируются на опыте, который получил автор в ходе руководства проектами по разработке автоматизированных систем (АС), предназначенных для обработки данных переписи населения Украины в 2001 г. и населения Молдовы в 2004 г., и последующей работе в статистической отрасли. В частности, были обобщены результаты встреч с зарубежными специалистами, занимавшимися разработкой и/или эксплуатацией:

– АС ВПН-2002 [1] (Москва, Российская Федеральная служба государственной статистики, Росстат, 2004 г.),

- La macro SAS CALMAR, La macro SAS CUBE d'échantillonnage équilibré [2] (Париж, Национальный институт статистики и экономических исследований Франции – Institut national de la statistique et des études économiques, INSEE, 2008 г.),
- The Census 2000 Testing, Experimentation, and Evaluation Program [3] (Соединенные Штаты Америки, Бюро переписи населения США – United States Census Bureau, 2009 г.).

Основы современного подхода к построению информационных систем в статистической отрасли были заложены известным шведским специалистом Б. Сунгреном (Bo Sundgren) в 1999 г. [4]. С того времени ежегодно под эгидой Статистического управления Европейских сообществ (ЕВРОСТАТ) проводятся специализированные семинары, носящие с 2004 г. название «Meeting on management of statistical information systems» [5], на которых обговариваются различные аспекты разработки и внедрения систем обработки переписных данных.

Анализ вышеперечисленных и других существующих информационных систем в области статистики населения показывает, что, обеспечивая основную функциональность по хранению и обработке переписных и/или иных демографических данных, эти системы дополнительно позволяют пользователям строить разнообразные аналитические отчеты. Однако поиск скрытых закономерностей, имеющихся в анализируемых данных, выполняется фактически лишь вручную путем выдвижения определенных гипотез и проверки их справедливости посредством построения аналитических отчетов.

### *ПОСТАНОВКА ЗАДАЧИ*

Объектом исследования являются информационные системы в области статистики населения, а предметом исследования – архитектурные вопросы их построения. Цель статьи – на основании анализа существующих информационно-учетных и информационно-аналитических систем обработки переписных данных определить роль и место систем, обеспечивающих интеллектуальный анализ данных.

### *ИНФОРМАЦИОННО-УЧЕТНЫЕ (OLTP) СИСТЕМЫ В ОБЛАСТИ СТАТИСТИКИ НАСЕЛЕНИЯ*

Основой любой информационной системы по обработке данных статистики населения (регистра или переписи населения, включая его естественное движение и миграцию) является так называемая OLTP (online transaction processing)-система [6], т. е. информационно-учетная система, предназначенная для *ввода, структурированного учета и обработки данных*. Специфика статистической отрасли проявляется в выдвижении к информационно-учетным системам следующих дополнительных требований по обеспечению:

- территориально-иерархической распределенности системы (например, в Украине в ходе переписи 2001 г. обработка данных велась на республиканском уровне и в 27 региональных центрах, в Российской Федерации на 2010 г. также планируется два уровня обработки [7]: федеральный центр и 67 региональных центров в соответствующих ТОГСах – Территориальных органах государственной статистики);
- поддержки очень больших объемов данных (например, в Украине при обработке данных переписи 2001 г. было просканировано и проверено более 68 млн переписных документов формата А4, а в Российской Федерации в ходе обработки данных переписи 2002 г. – около 220 млн бланков);
- поддержки большого количества разнородных правил входного контроля и правил внутритабличного, межразрезного и межтабличного контроля (так, Постановка задачи на Всеукраинскую перепись населения 2001 г. содержала около 200 тысяч соответствующих правил).

В отличие от других статистических исследований переписи населения характеризуются: индивидуальным подсчетом респондентов, одновременностью, установленной периодичностью,

тем, что они являются сплошными (охватывается все население страны, включая самые маленькие населенные пункты).

### *ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЕ (OLAP) СИСТЕМЫ В ОБЛАСТИ СТАТИСТИКИ НАСЕЛЕНИЯ*

Понятно, что проанализировать можно лишь те показатели, которые были в опросных листах. Если в украинских или российских переписных листах, в отличие, скажем, от грузинских, не было вопроса к женщинам – сколько они планируют еще родить детей, то проанализировать такие данные, конечно, нельзя.

Однако существуют ограничения и другого рода – для анализа и изучения доступны только те данные, которые предусмотрены функциональными возможностями информационно-учетной системы (фиксированный набор выходных таблиц и разрезов, по которым они строятся). Поэтому для анализа переписных данных дополнительно требуется система другого класса – информационно-аналитическая система, т. е. информационная система многомерного анализа данных, или, в западной терминологии, OLAP-система (online analytical processing) [8].

Термин OLAP был введен Коддом (Edgar Codd) в 1993 г. в работе [9], где он сформулировал 12 определяющих признаков OLAP-данных. В 1995 г. Пендс (Nigel Pendse) переформулировал 12 правил Кодда в более лаконичный критерий FASMI (Fast Analysis of Shared Multidimensional Information) – «быстрый анализ разделяемой многомерной информации».

В отличие от информационно-учетных систем, назначение информационно-аналитических систем состоит в поддержке процессов принятия решений или поиске определенных закономерностей (например, можно выявить зависимость между годом рождения респондента и полученным им образованием) за счет предоставления возможностей *быстрого анализа больших объемов информации*.

Системы многомерного анализа данных характеризуются следующими признаками [10]:

- добавление новых данных в систему происходит относительно не часто и большими блоками (например, переписная информация после корректировки ошибочных данных или по результатам очередной переписи);
- данные, которые вносятся в систему, как правило, никогда не уничтожаются;
- перед загрузкой данных в систему они проходят процесс валидации, который исключает возможность добавления дублирующих или некорректных данных;
- запросы к системе являются нерегламентными (ad hoc) и, в основном, достаточно сложными; очень часто новый запрос формируется аналитиком для уточнения результата, полученного в предыдущем запросе;
- важной является скорость исполнения запросов.

Технологически информационно-учетные и информационно-аналитические системы существенно отличаются друг от друга (табл. 1).

*Таблица 1*

**Сравнительные характеристики информационно-учетных и информационно-аналитических систем**

| Характеристика | Информационно-учетная система | Информационно-аналитическая система |
|---|---|---|
| **характер (уровень) данных** | главным образом, первичные | в основном, консолидированные |
| **изменчивость данных** | высокая (с каждой транзакцией) | низкая |
| **типичная операция** | изменение данных | анализ данных |
| **отчеты** | регламентные | нерегламентные |
| **сроки сохранения данных** | только текущие | исторические и текущие |
| **базовая структура** | таблица / первичный ключ | куб / измерение |
| **приоритет** | производительность | гибкость |

В рамках OLAP-системы пользователь получает естественную, интуитивно понятную модель данных, организовывая их в виде многомерных кубов. Измерениями куба выступают такие характеристики данных, в разрезе которых можно получить, отфильтровать, сгруппировать и отобразить информацию. Что именно за информация предоставляется кубом – определяется так называемой мерой. Например, мерой может выступать количество респондентов, а измерениями: «административно-территориальная единица», «семейное положение», «национальность» и т. п. Каждая ячейка куба хранит количество респондентов, имеющих соответствующие характеристики. Пользователь, анализирующий информацию такого куба, может «разрезать» его по разным направлениям, получить сводные (например, по всей стране) или, наоборот, детальные (по административным районам) сведения и т. д.

### *СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ (DATA MINING) КАК ТРЕТИЙ АРХИТЕКТУРНЫЙ УРОВЕНЬ ИНФОРМАЦИОННЫХ СИСТЕМ В ОБЛАСТИ СТАТИСТИКИ НАСЕЛЕНИЯ*

В настоящее время еще не сложилось четкого понимания места, которое в общей архитектуре должны занимать системы интеллектуального анализа данных – то ли их выделять отдельно, то ли объединять вместе с OLAP и хранилищами данных в системы business intelligence (бизнес-аналитики). Однако в любом случае, задача автоматического или автоматизированного *поиска шаблонов распределения данных и/или их скрытой (неявной) взаимосвязи* является потенциально очень интересной. Особенно ее важность возрастает в связи с тем, что в ходе переписей раунда 2010 г. статистические организации стран СНГ впервые в своей истории получат возможность автоматизированного многомерного анализа данных переписей в их исторической взаимосвязи (за два последних переписных цикла).

Поэтому представляется принципиально важным выделение отдельного уровня для систем подобного класса.

Таким образом, с архитектурной точки зрения современная система по обработке данных статистики населения, в частности, переписных данных, должна содержать три уровня:

- *информационно-учетный* уровень, обеспечивающий базовую функциональность, без которой подобные системы не имеют смысла; речь идет о вводе данных и материалов соответствующих статистических наблюдений и опросов, их структурированному (как правило, с помощью СУБД – системы управления базами данных) хранению и учету, контролю в первичном и сводном виде, распространению результатов в виде различных регламентных выходных таблиц;
- *информационно-аналитический* уровень, с помощью которого пользователи могут быстро строить нерегламентные таблицы и проводить другие аналитические исследования статистических данных в поиске предполагаемых ими закономерностей распределения этих данных;
- уровень *интеллектуального анализа данных*, который берет на себя наиболее громоздкую и рутинную аналитическую операцию по поиску скрытых закономерностей, существующих, например, в переписных данных.

### *ЗАКЛЮЧЕНИЕ*

В статье впервые для информационных систем в области статистики населения предложено выделить архитектурный уровень, связанный с интеллектуальным анализом данных, в частности, с поиском скрытых закономерностей распределения переписных данных. Перспективным представляется апробация различных методов такого анализа для изучения данных переписей населения.

### ЛИТЕРАТУРА


1. Всероссийская перепись населения 2002 г. Математико-статистическое обеспечение [Электронный ресурс]. – Режим доступа: http://www.perepis2002.ru/index.html?id=92.
2. La macro SAS CALMAR, La macro SAS CUBE d'échantillonnage équilibré [Электронный ресурс]. – Режим доступа: http://www.insee.fr/en/methodes/default.asp?page=outils/liste-outils.htm
3. The Census 2000 Testing, Experimentation, and Evaluation Program [Электронный ресурс]. – Режим доступа: http://www.census.gov/pred/www/.
4. Sundgren B. Information systems architecture for national and international statistical offices. Guidelines and recommendations // Conference of European statisticians. Statistical standards and studies, № 51. – Geneva: United Nations Statistical Commission, 1999. – 56 p.
5. UNECE Statistics. Management of statistical information systems [Электронный ресурс]. – Режим доступа: http://www.unece.org/stats/archive/04.01a.e.htm.
6. Weikum G., Vossen G. Transactional information systems: theory, algorithms, and the practice of concurrency control and recovery. – San Diego: Morgan Kaufmann, 2001. – 852 p.
7. Федеральная служба государственной статистики. Основные методологические и организационные положения Всероссийской переписи населения 2010 года. Проект. – М.: Росстат, 2009. – 46 с.
8. Thomsen E. OLAP solutions: building multidimensional information systems, 2nd ed. – N.Y.: John Wiley & Sons, 2002. – 661 p.
9. Codd E.F., Codd S.B., Salley C.T. Providing OLAP (On-Line Analytical Processing) to user-analysts: An IT mandate. Technical report. – Codd & Date, Inc., 1993.
10. Чертов О.Р. Система многомерного анализа данных Всеукраинской переписи населения 2001 года // Россияне в зеркале статистики: Всероссийская перепись населения 2002 года: Международный симпозиум, 30-31 марта 2004 г.: труды симп. – М.: Изд-во Федеральной службы государственной статистики, 2004. – С. 234-238.


© Чертов О.Р., 2010

*Стаття надійшла до редакції 22.04.10 р.*