УДК 004.413
doi: 10.18101/2304-5728-2016-4-13-22

© В. В. Парамонов, Р. К. Фёдоров, Г. М. Ружников, Н. В. Ефимова

Технология формирования и анализа данных здоровья населения Азиатского севера[1]

Одним из базовых принципов промышленного освоения Ямала является гармоничное сочетание развития индустрии на полуострове и бережного отношения к традиционному укладу жизни коренных малочисленных народов. Также требуется своевременная коррекция негативного влияния управляемых факторов среды обитания в регионе. В работе рассматривается технология, обеспечивающая возможности сбора гетерогенных данных о здоровье населения, имеющих пространственную привязку и представленных в виде таблиц, их обобщения и анализа. Анализ собранных данных позволит прогнозировать состояние здоровья и связанного с ним качество жизни населения Азиатского Севера на территориях освоения углеводородного сырья. Реализованные в технологии функции представлены в виде сервисов. Это позволяет обеспечить не только отображение результатов анализа данных в виде таблиц и тематических карт, но и предоставить возможность обрабатывать дополнительные данные, а также возможности подключения к сторонним сервисам для обработки накопленного массива информации.

Ключевые слова: геоинформационные системы, информационная технология, здоровье населения, анализ данных

© V. V. Paramonov, R. K. Fedorov, G. M. Ruzhnikov, N. V. Efimova

The technology of formation and analysis of data on Asian northern health population

One of basic principle of commercial development of Yamal Peninsula is a harmonious combination of industrial development and careful attitude for traditional lives of indigenous people. It is also requires a timely correction of negative impact of environment in Arctic. The paper deals with original informational technology. It is directed on gathering and analysis of heterogeneous data. This data should be represented in tabular view and have some spatial

[1] Работа выполнена при финансовой поддержке: программы президиума РАН АЗ РФ-44п, 0348-2015-0007; грантов РФФИ: 14-07-00-166, 16-07-00411, 16-57-4034, 15-47-04348, 15-37-20042; совета по грантам Президента РФ для государственной поддержки ведущих научных школ Российской Федерации (НШ-8081.2016.9). Работа выполнялась в Центре коллективного пользования “Интегрированная информационно-вычислительная сеть ИНЦ ФАНО”.

labels. Data analysis allows predicting the state of health and related quality of life of the North Asian populations in areas of hydrocarbons development. The functions in technology are suggested as services. It helps to provide displaying data analysis results as tables and thematic maps and allows also processing data in third-party services. Also it is allow to use informational system services for third-party data processing.

Keywords: geospatial systems, informational technology, population health, data analysis.

Введение

Азиатский Север России занимает большую по площади территорию страны, географические, климатические, социальные условия которой можно оценить как неблагоприятные. В этом районе наблюдается недостаточность солнечной радиации, редкая рекреационная сеть, крайне неразвитая инфраструктура и невысокая плотность населения. Экономический интерес России к данным районам связан с тем, что их территория рассматривается как источник для социально-экономического роста страны [1].

В районах Азиатского Севера осуществляется интенсивная разведка и освоение залежей углеводородов [2]. Разработана и утверждена президентом России «Стратегия развития Арктической зоны Российской Федерации и обеспечения национальной безопасности на период до 2020 года» [3]. При этом требуется проводить постоянный мониторинг за состоянием окружающей среды, выявлять влияние различных техногенных и антропогенных факторов на здоровье как коренного, так и пришлого населения обозначенной территории.

В настоящее время, отсутствуют какие-либо стандартизованные технологии и подходы, позволяющие достаточно просто агрегировать и анализировать разнородные и разноформатные данные. Это обуславливает важность создания и практического применения технологических решений для обобщения гетерогенных данных их мониторинга и выявления влияния активной хозяйственной деятельности, загрязнения атмосферного воздуха, воды на здоровье населения. Применение подобного рода систем позволит прогнозировать демографический и трудовой потенциал для указанной территории.

В работе предлагается технология, предоставляющая возможности сбора гетерогенных данных из источников, представленных в различных форматах, а также обеспечивающая их надёжное хранение, предоставления к ним избирательного доступа через сеть Интернет и проведение анализа. При этом предлагается использовать сервисный подход для формирования и обработки данных. Реализация аналитических функций в виде сервисов позволяет обеспечить не только отображение результатов анализа данных в виде таблиц и тематических карт, но и предоставить возможность обрабатывать сторонние данные.

1. Источники данных

Для проведения исследовательских работ, в Восточно-Сибирском институте медико-экологических исследований (ВСИМЭИ) собраны статистические данные по заболеваемости населения на территориях Ямало-Ненецкого автономного округа (АО) по социально-значимым классам и нозологическим формам болезней.

Данные включают информацию о категории заболевания, нозологическим формам, местности и периода в которых были зафиксированы случаи социально-значимых классов заболеваний. Собранный массив данных является ценным источником информации для анализа здоровья населения на территории Азиатского Севера

Основным источником для наполнения базы данных являлся Территориальный орган государственной статистики по Ямало-Ненецкому АО. Данные прошли предварительную обработку, к ним был применен персентильный [4] метод исследования для изучения заболеваемости.

Предполагается, что в данные могут также поставляться медицинскими и статистическими учреждениями, метеорологическими службами и т.п.

Например, в базу данных (БД) может быть загружена дополнительная информация о состоянии погодных условий в исследуемой местности, загрязнении атмосферного воздуха, воды и т.п. Это позволит поддерживать информацию в наиболее актуальном состоянии и выявлять зависимости между заболеваемостью, её формами и другими, различными факторами, косвенно связанными с качеством жизни.

Для корректной обработки данных, полученных из различных источников, требуется применять методы реконсиляции, трансформации с целью представления их в реляционном виде. Это также обеспечивает возможность применять к нормализованным и очищенным данным разнообразные методы анализа.

Процедура очистки данных позволяет выявить и, в ряде случаев, исправить ошибки и несоответствия в сырых (пользовательских) данных. Такими ошибками являются как дублирование, опечатки, ненужные префиксы, противоречия, пропуски, и т.п. Целью очистки является улучшение качества данных.

Собранная ВСИМЭИ информация [5] организована в виде набора таблиц БД Microsoft Access, модель которой представлена на рис. 1.

Перед загрузкой в интегрирующую БД информация также требует проведения нормализации и очистки.

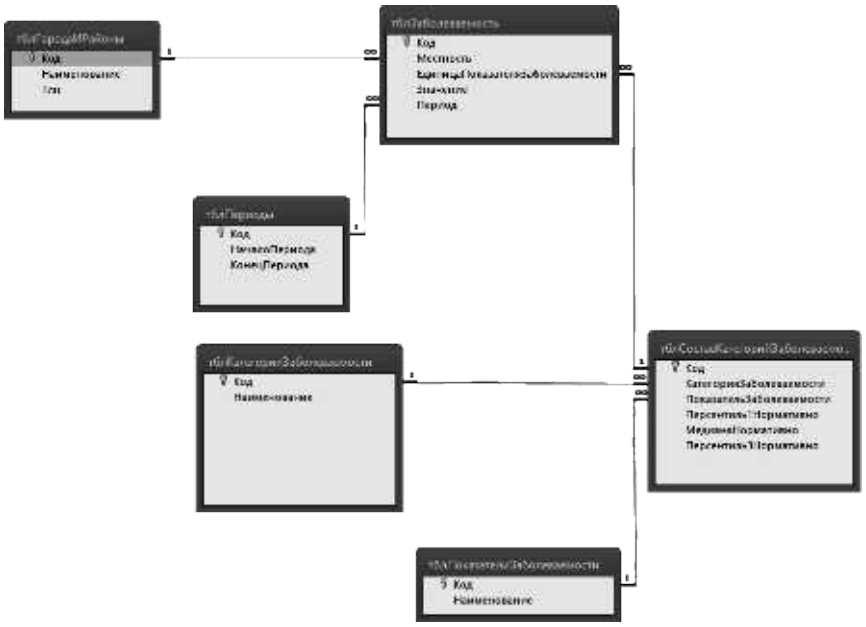

Рис. 1. Модель базы данных

Технология формирования данных

В Институте динамики систем и теории управления имени В.М. Матросова Сибирского отделения РАН (ИДСТУ СО РАН) в сотрудничестве с ВСИМЭИ разработана и апробирована технология обеспечивающая сбор и анализ данных, использующая систему ввода и редактирования реляционных данных «Фарамант» [6]. Работа системы «Фарамант» основывается на модели документа. Данная модель включает три компонента: описание структуры данных; отображение данных; логика работы формы. На основе модели документа создается таблица в СУБД PostgreSQL, формируется форма ввода и отображение данных в виде таблицы. Все этапы технологии выполняются пользователем с помощью браузера. Пользователь должен загрузить документ, представляющий электронную таблицу в систему. Идентификация таблицы будет осуществлена автоматически.

Технология включает следующие этапы:

1) загрузка данных;
2) определение модели данных;
3) импорт данных в созданные таблицы;
4) разработка форм ввода, определение логики;
5) анализ данных.

Рассмотрим подробнее этапы технологии, с использованием системы ввода и редактирования реляционных данных «Фарамант».

Загрузка данных

Для осуществления удаленной загрузки данных были разработаны специализированные оригинальные сервисы – файловый менеджер, по-

зволяющий производить все основные операции с файловой системой на сервере через браузер пользователя, а также выполнять загрузку и выгрузку документов с компьютера пользователя. Операции проводятся по протоколу HTTP. Также реализована возможность работы по протоколу FTPS (File Transfer Protocol + SSL).

Для каждого пользователя предусмотрен собственный каталог. Доступ к документам пользователя возможен в соответствии с определенной ему ролью. Это позволяет обеспечить защиту доступа к документам пользователей.

На текущем этапе реализована возможность загрузки в БД электронных таблиц, представленных в формате CSV.

Модель данных и права доступа

Предусматривается, что пользователь будет загружать данные в виде отдельных документов. Однако, в дальнейшем, требуется их представление в виде совокупности реляционных таблиц, представленных в третьей нормальной форме. Это является представлением модели данных.

Для описания модели данных пользователю необходимо определить название таблицы и указать список ее полей. Для каждого поля необходимо задать его название, а также специфицировать связанный с ним элемент управления (реализующий ввод/редактирование данных поля). Такого рода элементами могут быть «календарь» для ввода данных типа «дата», «текст» для ввода многострочного текста. В БД, как правило, используются различные классификаторы. Для ввода атрибутов, ссылающихся на таблицы БД без иерархической зависимости, применяется специализированный элемент «classify». При этом классификаторы могут, как содержаться в БД, так и быть загружены пользователем до загрузки основной таблицы. При этом, значения в загружаемом документе должны иметь соответствие со значениями в классификаторе.

Для загруженных в реляционные таблицы данных пользователь определяет права доступа к ним. Рассмотрим механизм предоставления доступа к данным. Первоначально доступ к таблице имеет только её владелец. В ИС имеется три группы пользователей: незарегистрированный пользователь, любой зарегистрированный в системе пользователь и определенный список пользователей. Для каждой группы можно регламентировать следующие операции:

- просмотр своих записей;
- просмотр всех записей;
- редактирование своих записей;
- редактирование всех записей.

Данное разграничение позволяет достаточно полно разграничить права доступа к информации, содержащейся в ИС. Это обеспечит предоставление доступа к материалам ИС с учетом требований пользователей, обеспечивать сохранность авторских прав.

Нормализация данных

При обработке данных, полученных из различных источников, часто требуется проведение их нормализации, т.е. приведение к простым типам данных, представленных в полях таблицы, а также привязка значений к справочникам (если это определено структурой таблицы). Необходимость нормализации продиктована наличием опечаток, измененной последовательности слов в названии, различными форматами записи дат, разделителей и т.п.

Для решения поставленной задачи авторами были разработаны сервисы, обеспечивающие нормализацию данных в полуавтоматическом режиме.

Исходя из указанной пользователем структурной информации, относительно типов загружаемых данных для проведения нормализации применяются различные методы. Если пользователь указал, что поле содержит справочные данные, то происходит сопоставление значений этого поля со стандартными справочниками с применением методов нечеткого сравнения строк, алгоритмов фонетического сопоставления. В случае невозможности установить соответствие между данными пользователя и данными из классификатора строка может быть проигнорирована либо пользователь может вручную сопоставить значение.

В случае, когда поле является единицей измерения возможно применение методов приведения «сырых» данных, основанных на правилах преобразований единиц измерения. Данный подход позволяет автоматически приводить измерения в нужную пользователю шкалу, например, из дюймов в метры.

При загрузке других данных в таком же формате пользователь может повторно использовать составленную им ранее структурную спецификацию.

Значения, содержащие пространственную информацию, подвергаются геокодированию и привязке к базовым пространственным данным.

Создание форм ввода данных

После создания таблицы система «Фарамант» автоматически создает форму заполнения. Это обеспечивает возможность ручного ввода, редактирования и дополнения данных. Для ввода каждого атрибута используются определенные в модели элементы управления. Элементы управления отображаются в соответствии с порядком следования атрибутов в таблице. Пользователю доступно модификация дизайна формы ввода/редактирования. Для этого требуется создать её шаблон. Под шаблоном формы понимается – это HTML код с указанием положения вставки элементов управления данными. Шаблон может быть создан в любом текстовом редакторе.

Указание месторасположения элементов управления задается при помощи специальных дескрипторов вида «#fieldname#», где fieldname это имя атрибута в БД, либо с помощью HTML тегов , у

которых идентификатор совпадает с именем атрибута. Применение контейнеров позволяет управлять отображением заголовков полей.

Для упрощения заполнения формы ввод данных должен подчиняться определенной логике. Так возможные значения атрибутов могут зависеть от других. Задание логики работы формы осуществляется с помощью задания зависимостей между атрибутами и управлением видимостью элементов управления атрибутов. Реализует зависимость элемент управления «classify». Для организации фильтрации текущего элемента управления «classify» необходимо указать, по каким атрибутам таблицы справочника производить фильтрацию, и какие значения используются для фильтрации.

Если в качестве значения указывается «#fieldname», то значение для фильтрации автоматически берется из соответствующего атрибута. При изменении значения ссылаемого атрибута список элемента управления «classify» автоматически обновляется. Если значения фильтрации не определены, т.е. пользователь не указал значения атрибутов, элемент управления не доступен для ввода.

Следующей возможностью управления логикой работы формы является управление видимостью элементов управления атрибутами. В свойстве элемента управления можно задать условие видимости в поле «Visisbility».

Условие выражается в соответствии с синтаксисом языка JavaScript. Может содержать только имена атрибутов, константы и специальные символы (например =). При изменении значения любого атрибута производится проверка условия видимости. Если в шаблоне формы для атрибута задан , то в этом случае можно скрыть атрибут вместе заголовком.

Анализ данных

В технологии реализованы сервисы анализа загруженных данных. Результаты их отображаются как в виде тематической карты, так и в виде таблиц.

Для табличного представления возможно применение различных фильтров, в соответствии с которыми отображаются строки таблицы. Тематическая карта при этом изменяется в соответствии с примененными фильтрами.

Для строковых атрибутов можно определить условие, при котором отображаются записи, значение выбранного атрибута содержит указанную подстроку. Для атрибутов типа Date указывается диапазон дат, при этом будут отображаться все записи со значениями дат из указанного диапазона. Для числовых атрибутов указывается условие с использованием разных знаков сравнения. Для атрибутов, ссылающихся на таблицы справочники, производится фильтрация, используя значения из таблицы справочника. Если заданы условия для нескольких атрибутов, то они все объединяются в запросе по логическому «И». Для атрибута можно определить несколько значений, по которому производится фильтрация, при

этом все указанные значения формируются в запросе по логическому «ИЛИ».

Для обобщения информации реализован механизм группировки данных, который позволяет получить данные о количестве записей, максимальных, минимальных, средних значениях по различным группам записей. В пользовательском интерфейсе для группировки записей необходимо выбрать поля, по оставшимся полям требуется выбрать групповую функцию (max, min, avg, count). Группировка записей производится для всех типов атрибутов по равенству их значений. Для полей типа «Date» можно группировать записи по неделям, месяцам, кварталам, годам и десятилетиям. При обобщении информации используются заданные пользователем фильтры. Данные таблицы выгружаются в формате CSV.

Заключение

Предложенная технология позволяет в сжатые сроки провести сбор информации и представить её в реляционном виде, а затем провести её анализ. Можно отметить недостаток технологии на текущий момент – пользователь ограничен вводом только одной записи в таблицу. Однако, проведенная апробация подтверждает работоспособность и функциональность данного подхода. Достоинствами разработанной технологии являются:

- наличие широкого набора элементов управления, позволяющих упростить ввод различных типов данных с применением динамической загрузки данных (AJAX);
- возможность задания логики с помощью определения зависимостей между атрибутами и управления видимостью;
- гибкий механизм задания дизайна формы;
- наличие методов загрузки и выгрузки данных;
- наличие развитых методов фильтрации и обобщения данных;
- высокая скорость разработки форм без необходимости программирования.

Рассматриваемый подход к формированию данных и обеспечению их анализа позволит интегрировать разрозненные данные по здоровью населения, на территориях Азиатского Севера проводить выявление связей между заболеваемостью и антропогенным влиянием в районах освоения углеводородного сырья, а также позволит прогнозировать демографический и трудовой потенциал для указанной территории. Это, несомненно, будет иметь положительный социально-экономический эффект для промышленного освоения Арктики.

Литература

1. Фаузер В. В. Демографический потенциал северных регионов России фактор и условие экономического освоения Арктики // Экономика Региона. — №4 (2014). — С. 69 – 81.

2. Татаркин А. И., Логинов В. Г. Оценка природно-ресурсного и про-

изводственного потенциала северных и арктических районов: состояние и перспективы использования. // Проблемы прогнозирования. — № 1. — 2015. — С. 33 – 45.

3. Стратегия развития Арктической зоны Российской Федерации и обеспечения национальной безопасности на период до 2020 года // Геополитика и безопасность. — 2013. — № 4. — С. 136 – 149.

4. Гудинова Ж. В., Овчинникова Е. Л., Нескин Т. А., Жернакова Г. Н., Толькова Е. И., Гегечкори И. В. Новый способ анализа заболеваемости детей в регионах (на примере районов Омской области) // Вопросы современной педиатрии. — 2015. — №1. — Т. 14. — С. 18 – 22.

5. Ефимова Н. В., Мыльникова И. В., Иванов А. Г., Елфимова Т. А. Свидетельство о государственной регистрации базы данных №2015621097 «Заболеваемость населения городов и районов ЯНАО: фоновые показатели и персентиль-профиль для отдельных возрастных групп».

6. Фёдоров Р. К., Шумилов А. С., Ветров А. А., Михайлов А. А., Ружников Г. М. Интернет-система ввода и редактирования пространственных данных «Фарамант». Свидетельство о государственной регистрации программ для ЭВМ № 2014610274. М.: Федеральная служба по интеллектуальной собственности, патентам и товарным знакам, 2014.

References

1. Fauzer V. V. Demograficheskij potencial severnyh regionov Rossii faktor i uslovie jekonomicheskogo osvoenija Arktiki // Jekonomika Regiona. — №4 (2014). — S. 69 – 81.

2. Tatarkin A. I., Loginov V. G. Ocenka prirodno-resursnogo i proizvodstvennogo potenciala severnyh i arkticheskih rajonov: sostojanie i perspektivy ispol'zovanija. // Problemy prognozirovanija. — № 1. — 2015. — S. 33 – 45.

3. Strategija razvitija Arkticheskoj zony Rossijskoj Federacii i obespechenija nacional'noj bezopasnosti na period do 2020 goda // Geopolitika i bezopasnost'. — 2013. — № 4. — S. 136 – 149.

4. Gudinova Zh. V., Ovchinnikova E. L., Neskin T. A., Zhernakova G. N., Tol'kova E. I., Gegechkori I. V. Novyj sposob analiza zabolevaemosti detej v regionah (na primere rajonov Omskoj oblasti) // Voprosy sovremennoj pediatrii. — 2015. — №1. — T. 14. — S. 18 – 22.

5. Efimova N. V., Myl'nikova I. V., Ivanov A. G., Elfimova T. A. Svidetel'stvo o gosudarstvennoj registracii bazy dannyh №2015621097 «Zabolevaemost' naselenija gorodov i rajonov JaNAO: fonovye pokazateli i persentil'-profil' dlja otdel'nyh vozrastnyh grupp».

6. Fjodorov R. K., Shumilov A. S., Vetrov A. A., Mihajlov A. A., Ruzhnikov G. M. Internet-sistema vvoda i redaktirovanija prostranstvennyh dannyh «Faramant». Svidetel'stvo o gosudarstvennoj registracii pro-gramm dlja JeVM № 2014610274. M.: Federal'naja sluzhba po intellektual'noj sobstvennosti, patentam i tovarnym znakam, 2014.

Парамонов Вячеслав Владимирович, кандидат технических наук, научный сотрудник Института динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, e-mail: slv@icc.ru.

Фёдоров Роман Константинович, кандидат технических наук, ведущий научный сотрудник Института динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, e-mail: fedorov@icc.ru.

Ружников Геннадий Михайлович, доктор технических наук, зав. отделения Института динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, e-mail: rugnikov@icc.ru.

Ефимова Наталья Васильевна, доктор медицинских наук, профессор, ведущий научный сотрудник Восточно-Сибирского института медико-экологических исследований, e-mail: medecolab@inbox.ru.

Paramonov Viacheslav Vladimirovich, PhD, researcher of Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences.

Fedorov Roman Konstantinovich, PhD, leading researcher of Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences.

Ruzhnikov Gennadiy Mikhailovich, DSc, chief of department of Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences.

Efimova Natalia Vasil'evna, DSc, leading researcher of East-Siberian Institute of Medical and Ecological Research.