УДК 004.031.4

© А.В. Качанов

# АКТУАЛИЗАЦИЯ БАЗ ДАННЫХ ГЕОИНФОРМАЦИОННЫХ АТЛАСОВ В АВТОМАТИЗИРОВАННОМ РЕЖИМЕ

Статья посвящена вопросам автоматизации процесса сбора и подготовки данных для электронных атласов. Для решения этой задачи предложен подход к проектированию агентно-ориентированных интерфейсов для ГИС. Описана структура системы автоматического сбора данных и роли агентов в ней. Описаны основные способы анализа веб-страниц.

Стаття присвячена питанням автоматизації процесу збору й підготовки даних для електронних атласів. Для вирішення цієї задачі запропонований підхід до проектування агентно-орієнтованих інтерфейсів для ГІС. Описана структура системи автоматичного збору даних і ролі агентів у ній. Описані основні способи аналізу веб-сторінок.

The paper covers problems of automatic fetching and preparing data for electronic atlases. The approach to designing agent-oriented interfaces for GIS that is able to solve this task is proposed. The structure of the system for automatic data collection and agent's roles in it are described. Techniques of automated parsing of webpages are discussed.

**Введение.** Один из самых трудоёмких этапов создания электронных атласов – это процесс сбора, согласования и обработки атрибутивных (статистических) данных. Это обусловлено, в основном, их большим объемом и разнородностью. Особенно важен этот этап при создании пилотной версии атласа, когда отбор исходных данных выполняется с привлечением экспертов, принимающих участие и в других этапах разработки. Под атласом в данной работе понимается система электронных карт, объединённая единой целевой установкой в целостное картографическое произведение [1]. В случае электронных атласов, используемых при поддержке принятия решений в процессе регулирования устойчивого развития, основным их свойством должна быть актуальность информации (преимущественный формат которой – статистические таблицы, отражающие эколого-социально-экономические особенности изучаемых территорий).

Поэтому после создания готового продукта основной задачей должно стать поддержание атласа в актуальном состоянии. Этот процесс включает в себя сбор и добавление в базу данных атласа новых табличных, атрибутивных и статистических данных. В самом простом и наименее затратном случае обновление выполняется без создания новых тематических разделов, исключительно путем обновления и расширения временных диапазонов, в которых представлены разделы атласа.

На практике же после разработки электронных картографических продуктов интерес разработчиков, равно как и финансирование, угасает, за редкими исключениями в виде либо прибыльных (Google Earth), либо открытых для свободного редактирования «общественных» проектов (OpenStreetMap). Кроме того, многие электронные атласы выпускаются в виде CD-дисков, в этом случае оперативное обновление информационного наполнения технически невозможно, как и в случае бумажных изданий.

Поэтому актуальным является вопрос автоматизации процесса сбора исходных данных, который позволит поддерживать атлас в актуальном состоянии без значительных трудозатрат.

**Постановка задачи и существующие способы ее решения.** Одной из проблем в организации процесса поиска и сбора данных для атласа является разнородность источников данных. В настоящее время наиболее удобным и быстрым каналом поступления данных являются глобальные компьютерные сети, в частности, Интернет. Специфика использования сети Интернет состоит в том, что она, с точки зрения процесса создания атласа, представляет собой распределенную систему разнородных источников данных. Поиск и сбор данных в Интернет для атласа не может выполняться по одинаковому алгоритму. Каждый источник данных требует особых методов доступа: обращение и выборка из различных баз данных, обработка табличных и PDF-документов, получение данных с web-страниц, причем в каждом случае технические детали получения данных могут значительно отличаться.

Предложено различные способы решения этой проблемы:

- создание единого Банка геопространственных тематических данных [2];
- создание национальных и региональных инфраструктур пространственных данных (ИПД) [3, 4, 5];
- использование семантической паутины, включающей семантические сети и онтологии [6], и пригодной для машинной обработки данных.

Реализация любого из этих методов дала бы несомненно положительный результат.

Очевидно, что после создания единых ИПД и банков геоданных и практической реализации концепций семантических сетей можно будет в значительной степени упростить, формализовать и автоматизировать поиск и извлечение данных для целей создания электронных атласов. Но в настоящее время практически нет реализованных систем, собирающих и дающих доступ к единым базам гео- и статистических данных на национальных уровнях.

На практике же в большинстве случаев перечисленные подходы ограничиваются теоретическими исследованиями и разработкой рекомендаций. Основная помеха для их практической реализации – огромные трудозатраты на переработку данных, причем на этом этапе их отбором и обработкой должен заниматься человек, а не компьютер.

В качестве исключения можно привести, пожалуй, только систему Eurostat статистической службы Европейского союза (www.ec.europa.eu/eurostat), занимающуюся сбором статистической информации по странам-членам ЕС и гармонизацией статистических методов, используемых данными странами.

В данной статье предлагается метод решения задач сбора и предварительной обработки данных из разнородных источников для использования в БД электронного атласа с применением легко реализуемой на практике технологии программных агентов и мультиагентных систем.

**Описание технологии.** Технология базируется на использовании агентов для выполнения рутинных задач поиска и подготовки исходных данных и обновления и актуализации атрибутивной информации в разделах атласа.

Агент в данном случае - это программный объект (сущность), обладающий определенным искусственным интеллектом и автономно функционирующий для сбора данных из внешних источников. Для каждого источника данных создаются отдельные независимые агенты, которые «знают» о технических подробностях получения данных, таких как:

- язык или протокол обращения к базе данных,
- способ анализа веб-страницы и извлечение данных из нее с отсеиванием HTML-тегов,
- распознавание текста и таблиц в PDF-файле,
- использование подходящего архиватора,
- и т.п.

Агент может быть не только программной сущностью, но и программно-аппаратной – например, для получения и подготовки данных из бумажных источников (комплекс – сканер и соответствующее ПО для распознавания текста и таблиц, последующей предобработки).

Поскольку агенты-сборщики никаким анализом или обработкой полученных данных не занимаются и не обладают сведениями об общей структуре системы, их основная задача – получить актуальные данные из источника и передать их агенту соответствующей специализации для дальнейшей обработки. Для работы на следующих этапах создаются отдельные агенты (рис. 1):

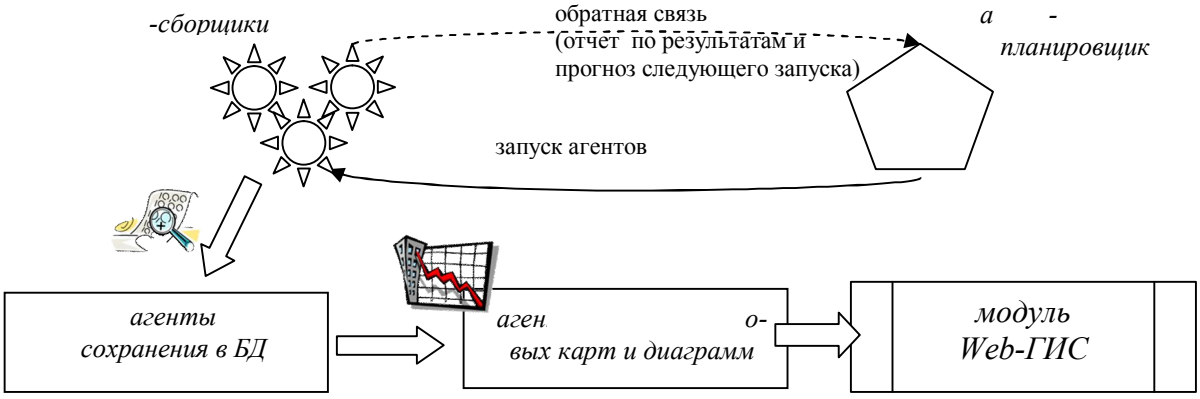


Рис. 1. Структура агентной системы сбора данных для электронного атласа

- агент для предварительной обработки данных,
- агент для сохранения данных в БД,
- агент для построения новых карт и диаграмм (интегрируется в геоинформационную систему),
- агент-планировщик для запуска сборщиков по варьируемому в зависимости от активности (частоты обновления) источника графику.

Агент-планировщик не определяет время следующего запуска агента самостоятельно. Вместо этого каждый агент сам рассчитывает время и передает

его планировщику. Алгоритм расчета периодичности позволяет оптимизировать нагрузку на источники данных и работает следующим образом: изначально для агентов задается периодичность по умолчанию (например, одна неделя); если после очередного запуска получены новые данные из источника, то время следующего запуска этого агента-сборщика будет уменьшено на несколько дней. Если же, наоборот, новые данные в источнике не найдены, период опроса увеличивается. Таким образом, после нескольких итераций, агент сам определяет частоту обновления данных для источника и выбирает оптимальный период повтора запросов, вплоть до режима непрерывного обновления или обновления раз в год/квартал и т.п. Планировщик все же должен учитывать нагрузку в системе и может откладывать запуск в зависимости от приоритета, установленного для различных агентов.

Рассмотрим подробнее разработку агентов-сборщиков. Как уже указано выше, они могут обращаться к хранилищам данных по разным протоколам. В случае работы с базами данных способы получения данных очевидны: использование соответствующего языка запросов или протокола работы с конкретной СУБД, на выходе - уже готовые к дальнейшей обработке табличные или иерархические данные.

В более общем случае обращение происходит к неструктурированным источникам, не имеющим специальных протоколов для выборки данных. К ним можно отнести веб-страницы. И именно они в настоящее время являются основным поставщиком данных для электронных атласов. Наибольший эффект от использования агентов для сбора данных проявляется именно при работе с данными в Веб.

Агенты-сборщики необходимо настраивать и обучать для каждого источника отдельно, благодаря чему в дальнейшем они позволяют получать данные в автоматическом режиме. Для автоматического анализа и разбора (*парсинга*) веб-страницы на практике можно выделить такие основные подходы:

- простой разбор документа с помощью анализа строк и регулярных выражений;
- использование объектной модели HTML-документа;
- преобразование HTML-документа в «правильно построенный» (well-formed) XML-документ и использование языка запросов XPath.

Наиболее удобным и универсальным, а в то же время наименее трудоемким, способом является парсинг веб-страниц с помощью запросов XPath. Ограничением этого способа является необходимость подачи на вход анализатора (парсера) XML-документа, соответствующего строгим стандартам структуры и оформления. Но стандарты языка HTML (особенно до появления версии xHTML 1.0 Strict) допускали достаточно свободную трактовку, поэтому не все HTML-документы удается правильно и без потерь преобразовать в XML, что приводит к появлению различных ухищрений при составлении запросов XPath в процессе обучения агента.

**Заключение.** Предложенный в статье подход к проектированию агентно-ориентированных интерфейсов для ГИС, основанный на мультиагентных сис-

темах, позволяет автоматизировать процесс поиска атрибутивных данных, наполнение и обновление баз данных электронных атласов.

Такая технология не сможет сделать процесс обновления атласа полностью автоматическим, ведь первоначальным обучением агентов занимается эксперт-программист. Обучение состоит в указании правил (создании процедур и программ), по которым получают информацию из каждого источника. Но после обучения данные собираются автоматически с адаптивной подстройкой интервала следующей проверки, и на основе собранных данных проводится актуализация соответствующих разделов атласа. Это делает практически легко реализуемой подсистему сбора данных для электронных атласов.

В перспективе повышение эффективности описанной агентной системы возможно путём дополнительной автоматизации процесса поиска новых источников, в частности, путём анализа известных гипертекстов, выделения ключевых слов и поиска новых источников и сайтов, «похожих» на уже известные.

## Список литературы


1. Бусыгин Б.С. Создание электронного атласа устойчивого развития регионов Украины / Б.С. Бусыгин, А.В. Качанов, Л.В. Сарычева // Ученые записки Таврического национального университета им. В.И.Вернадского. Том 18(57) №1 Серия "География", Симферополь, ТНУ, 2005. -С. 9-15.
2. Краюхин А.Н. Роль и место тематической картографии в системе государственных информационных ресурсов. А.Н.Краюхин // Тематическое картографирование для создания инфраструктур пространственных данных / Материалы IX научной конференции по тематической картографии (Иркутск, 9-12 ноября 2010 г.). – Иркутск: Изд-во Института географии им. В.Б. Сочавы СО РАН, 2010. – В 2-х т. – Т. 1. – С. 5-7
3. Путренко В.В. Світовий досвід організації тематичної інформації у інфраструктурах геопросторових даних / В.В.Путренко // Розвиток тематичної складової інфраструктури геопросторових даних в Україні: Зб. наук. праць. – К., 2011. – С. 133-138
4. Бешенцев А.Н. Картографирование инфраструктур пространственных данных. / А.Н. Бешенцев // Тематическое картографирование для создания инфраструктур пространственных данных / Материалы IX научной конференции по тематической картографии (Иркутск, 9-12 ноября 2010 г.). – Иркутск: Изд-во Института географии им. В.Б. Сочавы СО РАН, 2010. – В 2-х т. – Т. 1. – С. 25-28
5. Карпінський Ю.О. Від інфраструктури картографічного виробництва до інфраструктури геопросторових даних / Ю.О. Карпінський, А.А. Лященко // Розвиток тематичної складової інфраструктури геопросторових даних в Україні: Зб. наук. праць, Інститут географії НАН України, - К., 2011, - С. 39-61.
6. Grobelny P. Results of research on method for intelligent composing thematic maps in the field of Web GIS. / P.Grobelny, A.Pieczyński //Lecture Notes in Artificial Intelligence: Computational Collective Intelligence, Technologies and Applications, LNAI 6922, pp. 264—274. Springer, Berlin, Heidelberg, 2011



*Рекомендовано до публікації д.т.н. Слєсарєвим В.В*
*Надійшла до редакції 26.02.2013*