ВНЕДРЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

УДК 004.6:631.42

ПРОБЛЕМЫ ПОЧВЕННОГО МОНИТОРИНГА АГРОЛАНДШАФТОВ: СТРУКТУРА И МОДЕЛЬ ДАННЫХ

В.С. Крыщенко, *д.б.н.*, **О.М. Голозубов**
Южный Федеральный Университет, e-mail:web@sarmat.ru

Рассматривается технология объектно-ориентированных баз данных, в которых объекты содержат как данные и метаданные, так и правила ввода и первичной обработки сырых данных, средства импорта-экспорта данных и средства построения интерфейса ввода данных.

Ключевые слова: реляционные базы данных, объектно-ориентированные базы данных, темпоральные базы данных, метаданные, почвенный мониторинг.

SOIL MONITORING' PROBLEMS OF AGROLANDSCAPE: THE STRUCTURE AND MODEL OF DATA

V.S. Kryschenko, O.M. Golozubov

There is being considered the technology of object-oriented databases, where the object include not only the data and metadata, but also the rules of input and primary processing of the raw data, data' import-export means and those interface' construction of data input.

Keywords: relative databases, object-oriented databases, temporal databases, metadata, soil monitoring.

В предыдущей статье [1], посвященной проблемам формирования базы данных для широкомасштабного почвенного мониторинга были определены элементарные информационные единицы – объекты базы данных (БД), а также указано место этих объектов в реляционной структуре данных. В настоящей работе более детально рассматривается структура и модель данных (метаданные) объекта «наблюдение» и вспомогательных объектов с учетом специфики задачи широкомасштабного мониторинга, а также определяются принципы объектно-ориентированного реляционного подхода к формированию БД.

В традиционных почвенных обследованиях для объекта наблюдения (точечного – почвенный профиль, почвенный образец, или имеющего площадь – почвенный контур, земельный участок) в БД вносится информация полевых и лабораторных исследований. Эти данные подразделяют на количественные или качественные и выделяют следующие шкалы измерений: именная или классификационная шкала измерений (например, названия видов растений, типов почв, название цвета почв и т.п.); порядковая шкала измерений; интервальная шкала измерений; относительная шкала измерений [2]. Измерения, выполненные в первых двух шкалах относятся к качественным признакам, в последних двух – к количественным.

Любая база данных содержит не только сами данные, для хранения которых она была создана, но и **метаданные,** информацию о типах данных, ограничениях, наложенных на значения данных, связях между данными. В практике применения почвенных баз данных для указанных шкал измерений используют такие основные типы данных, как целые и вещественные числа, текстовые поля с фиксированной и переменной длиной, логические (да или нет) данные, двоичные данные фиксированной или переменной длины, типы данных для хранения даты и времени [3]. Это, так называемые, простые типы данных, название которых происходит от определенного в теории реляционных баз данных простого домена (simple domain), как множества однотипных атомарных значений. Примеры простых типов данных:

- тип почвы, например, «чернозем» – текстовое поле переменной длины для именной (классификационной) шкалы измерений;
- нарушенность почвенного профиля, например, «слабая» в перечне значений от «отсутствует» до «очень сильная», текстовое поле переменной длины для порядковой шкалы измерений;
- степень кислотности – pH, например, 7,2, – вещественное число для относительной шкалы измерений;
- при определении гранулометрического состава почв содержание частиц размером от 0,01 мм до 0,001 мм, например, 35% – целое число для интервальной шкалы измерений.

В структурно-реляционном подходе база данных представляется [4] двумя частями: регулярной, состоящей из совокупности часто изменяющихся во времени отношений соответствующей степени, которая иногда называется экстенсионалом (extension), и нерегулярной части, состоящей из формул логики предикатов, которые являются относительно устойчивыми во времени (ее называют иногда интенсионалом (intension)). Другими словами на доменах определяются основные данные и вспомогательные (справочники), которые можно также рассматривать как **множество ограничений це-**

лостности (т.е. условий, которые определяют все допустимые экстенсионалы) и таким образом отделить эти понятия от изменчивости во времени. На практике это означает, что для указания, например, типа почвы создается две таблицы, в одной из которых (справочник) перечисляются возможные в соответствии с классификацией почв типы почв, а в другой (основные данные) указывается номер по этому справочнику для конкретного объекта – почвенного профиля. Таким образом, устанавливается отношение (реляция) типа «один ко многим» (рис. 1).

Здесь необходимо подчеркнуть отличие этого примера от отношения между почвенным разрезом и почвенным образцом [4] (также «один ко многим»), поскольку в первом случае это отношение представляет собой технический прием реализации ограничения, а во втором это сущностная иерархия вложенности одного объекта в другой.

Для данных, измеряемых в количественной шкале измерений примером ограничения может быть задание **области допустимых значений**, например диапазон 0-14 для значения pH, или интервал 0-90% для показателя максимальной гигроскопической влажности.

В традиционных почвенных обследованиях, выполняемых в соответствии с единым стандартом или методикой [5], практика построения почвенных БД, имеющих классическую реляционную структуру и содержащих метаданные, нашла широкое применение.

Задачи современных широкомасштабных почвенных обследований имеют свою специфику и предъявляют новые требования к структурам и механизмам баз данных, чем и объясняется обращение почвоведов к концептуальным проблемам и терминологии БД, семантическому моделированию данных, онтологии и интеграции данных, обращению к проблемам «сущность-отношение» [4]. С другой стороны, специалисты в области БД отмечают недостатки поддержки для баз данных, описывающих естественные объекты, соответствующие *гипотезе открытого мира* [6, 7], в отличие от коммерческих СУБД, описывающих *замкнутый мир* искусственно созданных объектов. Попытаемся назвать основные специфические требования задачи широкомасштабного мониторинга агроландшафтов и способы их реализации за счет расширения базовой реляционной теории БД.

1. Многообразие применяемых методик анализа почв и классификаций. В проекте создания общероссийской почвенной базы данных [7] для такого показателя, как гранулометрический состав почвы, предусматривается применение более 10 различных методик подготовки почвенных образцов и их анализа. Применение современных технологий точного земледелия в России связано с использованием импортного оборудования (пробоотборников и экспресс-лабораторий), в программном обеспечении которых также используют разнообразные (в зависимости от страны-производителя) методики определения показателей. Такая множественность методик характерна для десятков показателей почвенного мониторинга.

Каждой методике соответствует показатель точности измерения и область допустимых значений, а иногда и единица измерения. Перевод показателя из одной системы в другую [8, 9] часто бывает невозможен. В общем случае, для некоторой пробы почв показатель может быть определен по нескольким методикам, что приводит к отношениям типа «многие ко многим» между справочником методик и списком показателей в БД. Чтобы не нарушать реляционную структуру БД для организации таких отношений используется промежуточная таблица «Показатель ***S***, измеренный по методике ***M*** для пробы ***P***».

Множественность классификаций также характерна для почвенного мониторинга. Включение архивных данных почвенных обследований, а также применение международных систем классификации приводит к использованию нескольких названий даже для типа почвы [7]. Здесь также возникают отношения типа «многие ко многим», поскольку один почвенный профиль или контур может быть определен по нескольким классификациям. Аналогичные отношения также возникают при введении семантического ограничения на список подтипов почвы, допустимых для данного типа почвы [10]. На рисунке 2 приведена реляционная структура, реализующая такие типы отношений.

2. Учет субъективности данных. Почвенные данные носят субъективный характер, и получаемые результаты во многом зависят от множества условий испытания и индивидуальности аналитика [2]. Связано это, как с несовершенством методик обследования (что может быть, в принципе исправлено), так и с открытостью и сложностью системы объектов почвенного мониторинга. При широкомасштабном сборе данных информацию часто комбинируют на основе разных источников, степень надежности которых различна. Следовательно, требуются методы для оценки достоверности полученной таким образом информации. Нужны также средства для опроса достоверности или *происхождения* (lineage) данных [6]. На практике это означает наличие списка

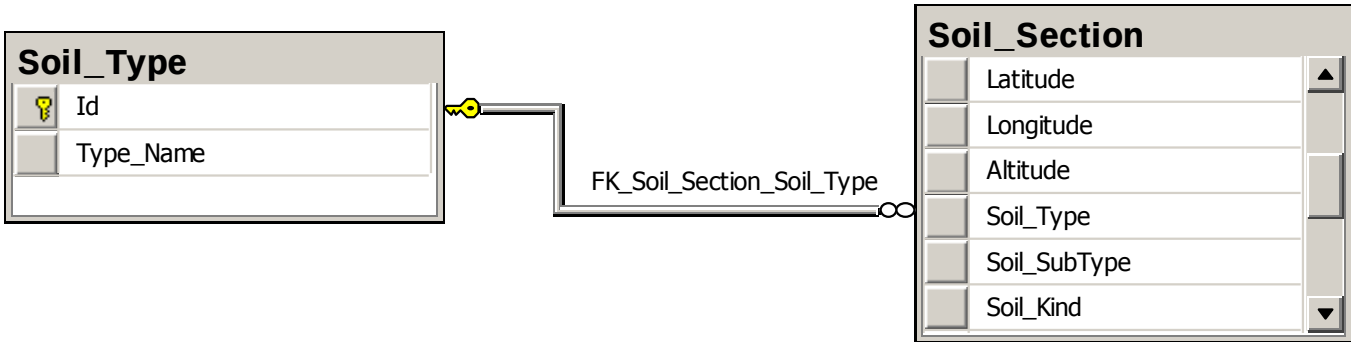

Рис. 1. Пример отношений «один ко многим» в таблицах Тип почвы и Разрез

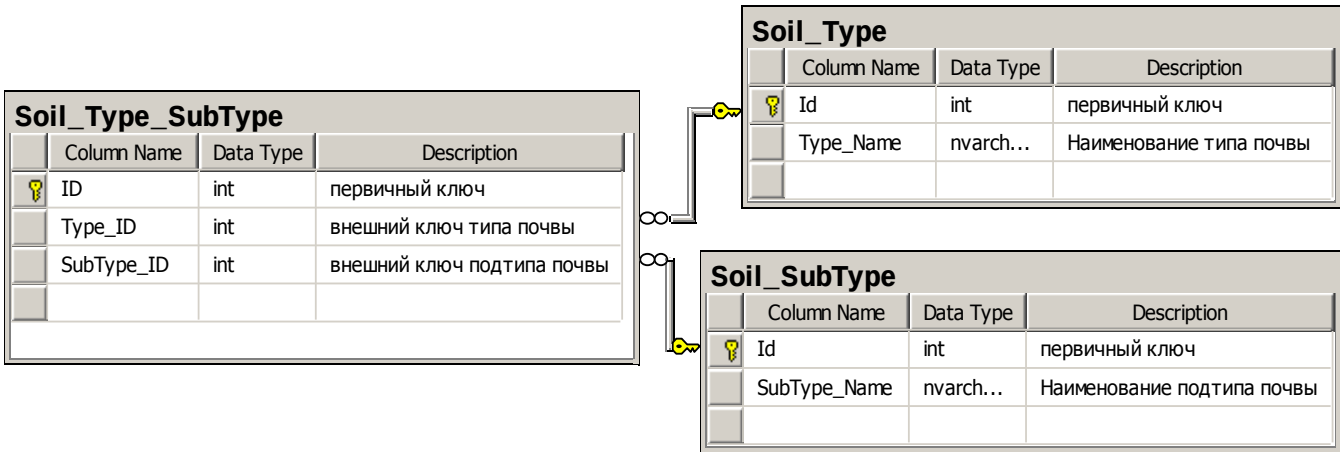

Рис. 2. Пример организации отношения «многие ко многим» для классификации типа и подтипа почвы

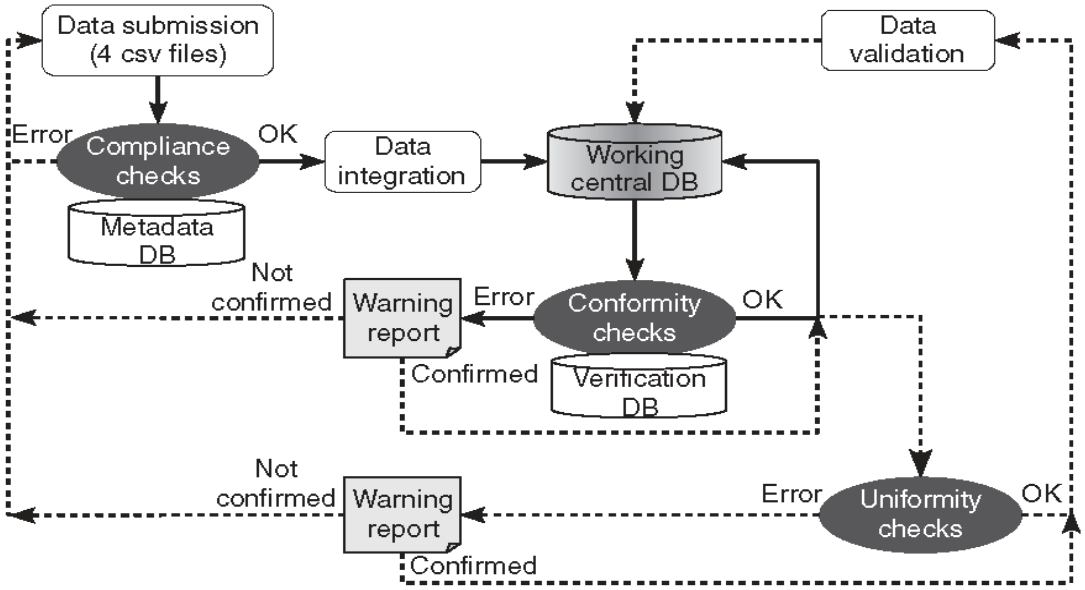

Рис. 3. Функциональная схема трехшагового процесса проверки данных на единообразие (uniformity), соответствие (conformity) и согласованность (compliance). Сплошными линиями процессы выполняющиеся автоматически, пунктирными – с привлечением эксперта.

поставщиков информации и формирование матрицы размерности ***M*** x ***N*** (где ***M*** – это количество показателей мониторинга, а ***N*** – количество поставщиков информации), в ячейках которой стоит коэффициент доверия к получаемой по данному показателю информации от данного поставщика. Например, данные по урожайности культур, получаемые от хозяйств, могут быть менее точными, нежели данные дистанционного зондирования, а данные по агрохимическим мероприятиям будут более точными, нежели показатели Агрохимцентров. В некоторых СУБД динамически определяется уровень доверия к данным, которые подвергаются «очистке» как в автоматическом режиме (по соответствию метаданным), так и в супервизорном режиме с привлечением экспертных (а значит, субъективных) знаний. Функциональная схема трехшагового процесса верификации данных в европейском проекте **Biosoil** представлена на рисунке 3.

3. Временно́й характер вводимых данных. Обычные БД хранят мгновенный снимок модели предметной области. Любое изменение в момент времени *t* некоторого объекта приводит к недоступности состояния этого объекта в предыдущий момент времени.

В задачах мониторинга в зависимости от специфики показателей *время съема показателя, время составления отчета и время ввода показателя в БД* играют существенную роль. Например, отчет одного источника, поступивший в момент времени *t1*, о внесении азотистых удобрений на земельном участке в момент времени *t2*, может существенно повлиять на оценку значения показателя общего содержания азота в % образца почвы, взятого в момент времени *t3*, обработанного в лаборатории в момент времени *t4,* и введенного в БД в момент времени *t5*. Оценка данного показателя также может измениться при наличии данных о культуре произрастания, а также температуре, осадках в период [*t1, t3*]. Другим

примером временны*х* соотношений служит изменение основных справочников. Так, административные решения, приводящие к изменению кодов ОКАТО хозяйства, должны иметь срок действия, задаваемый интервалом [*t1, t2*] *времени начала и окончания* действия кода, а временной объект должен содержать указатель на своего предшественника, т.е. на объект, предшествующий значению *t1*. Аналогичные временные параметры должны указываться и при введении в классификацию нового подтипа почв, или изменении названия методики и т.п.

Существует отдельное направление исследований и разработок в области темпоральных БД (то есть, базы данных, в которых поддерживаются исторические данные, а запросы могут содержать временные характеристики интересующих объектов). В этой области исследуют вопросы моделирования данных, языки запросов, организацию данных во внешней памяти и т.д. Основной тезис темпоральных систем [5] состоит в том, что для любого объекта данных, созданного в момент времени *t1* и уничтоженного в момент времени *t2*, в БД сохраняются (и доступны пользователям) все его состояния во временном интервале [*t1, t2*].

Исследования и построения прототипов темпоральных СУБД обычно выполняют на основе некоторой реляционной СУБД. Это не лучший способ реализации с точки зрения эффективности, но он прост и позволяет проводить достаточно глубокие исследования. Возможна выборка информации, хранившейся в базе данных в указанное время, в указанном временном интервале и т.д. В некоторых базах данных важную роль играет упорядоченность событий во времени, и альтернативное следование событий. Обеспечение регистрации этого упорядочения на уровне типа метаданных представляет собой шаг в направлении *поддержки сценариев* (script) для более крупных, нежели элементарный объект семантических единиц.

4. Потребность в экономически эффективном массовом вводе данных. В СУБД, рассчитанных на массовый ввод данных (тысячи и десятки тысяч поставщиков информации, десятки видов форм для заполнения) через Интернет, большую роль играет экономическая эффективность и скорость ввода данных [9].

Поставщик информации должен заполнять часто вручную формы или таблицы с десятками полей и пересылать их в БД. Если этот процесс происходит одномоментно в оперативном (on-line) режиме, то СУБД обеспечивает корректность справочников, проводит проверку достоверности данных и сообщает пользователю результат. А использование веб-интерфейса позволяет обеспечить единообразие форм и их своевременное обновление для всех пользователей независимо от клиентской платформы. Однако часто необходимы режимы автономного (off-line) заполнения форм.

В разнообразных приемах, используемых для решения этой задачи можно обнаружить одну тенденцию – перенос части программного кода «ближе» к данным – в форму, заполняемую пользователем. Например, данные от фермеров, поступающие в почвенную БД, налоговые декларации в США заполняют на основе PDF-форм, стандарта фирмы Adobe для обмена информацией. Этот стандарт позволяет осуществлять первичную проверку данных на область допустимых значений, осуществляет проверку типа данных (например, ввод текстовой информации в числовое поле), гарантирует защищенность данных от модификаций в процессе пересылки и многое другое. Стандарты, применяемые в Минсельхозе РФ для сбора форм МОП и ДДЗ, основаны на использовании в таблицах EXCEL макросов для проверки правильности и полноты заполнения форм. В формате XML, применяемом в МНС РФ для заполнения нескольких сотен форм налоговой отчетности, схема данных включает в себя области допустимых значений, списки (справочники) для выбора и даже форматы ввода текстовой информации. В перечисленных примерах ***объект*** (наблюдение фермера или Агрохимцентра, отчет налогоплательщика или юридического лица) содержит в себе не только сами данные, но и элементы программного кода в одном из стандартных международных форматов. В базу данных объектная информация может поступать как совместно с кодом (для случая EXCEL или PDF) в едином блоке, либо в разделенном (XML) виде. СУБД может оперировать с объектом как с единым целым, но может также получать доступ к отдельным элементам объекта.

5. Применение ГИС и новые типы данных. В ГИС основные показатели почвенного мониторинга называют атрибутивной информацией. В задачах ЦПК (цифрового почвенного картографирования) основной вес данных ложится на геоинформационные структуры данных (для каждого слоя топографические данные – классы, первичные и вторичные топографические атрибуты), и их обработку для исследования пространственного распределения характеристик и пространственной экстраполяции. Атрибутивные данные прилагают к картографическим структурам [3].

Многообразие программных реализаций и версий ГИС привело к необходимости выработки стандартов обмена геоинформацией; международные организации постоянно совершенствуют и уточняют эти стандарты. Выполненные на основе XML стандарты (GML, KML, XML) позволяют включать в себя не только собственно географические данные и атрибутивную информацию, но также правила и код для манипулирования этими объектами. Фрагмент файла стандарта KML 2.2 для экспорта данных о почвенном контуре из ArcGIS 9.3 в GoogleEarth 4.5 представляет не только сам базовый объект, но и атрибутивные данные, основные справочники (тип, подтип, род, вид почвы и др.) и правила обработки этих данных.

В традиционных ГИС почвенные данные представляют контурами с «жесткими» границами, задаваемыми многоугольниками или кривыми, и являются *векторными* географически привязанными данными.

Автоматизированный сбор данных, будь то данные по урожайности, полученные от уборочной техники, оснащенной специальным оборудованием, или цифровая модель высот, полученная из космоснимка, или данные проб почв от автоматизированного пробоотборника, приводит к образованию нового типа *растровых* географически привязанных данных для обладающего площадью объекта. При обмене геоинформацией такой растровый объект может быть либо включен в состав файла обмена в виде бинарного поля переменной длины, либо в виде прикрепленного файла. Сочетание векторных и растровых данных для одного объекта часто определяется как *гибридные* данные.

Поскольку в задачах широкомасштабного мониторинга при сборе информации от различных ведомств и организаций неизбежно будут использовать ГИС различных производителей и версий, то одна из основных задач при интеграции данных – это создание и обновление нового типа метаданных – перечня стандартов, их версий, кодов для проверки соответствия и преобразования стандартов. Для случая XML-подобных стандартов в качестве метаданных может использоваться *схема* (XSD-файл). Примером масштабного использования XML/XSD стандартов является всероссийская система сдачи налоговой отчетности (www.nalog.ru) и другие проекты государственного масштаба.

Заключение. Структуры данных и метаданных в БД почвенного мониторинга имеют более сложную организацию по сравнению с обычными почвенными БД, которая успешно может быть описана в рамках объектно-реляционных структур:

- Ввод субъективных данных в систему сопровождается информацией об источнике и происхождении данных, а сами данные могут проходить один или несколько этапов «очистки».

- В системах баз данных традиционно не поддерживается необходимый тип данных – N-мерный массив. Стандартные языки запросов к реляционной БД типа SQL не поддерживают обработку таких массивов, и тем более обработку объекта представленного гибридными данными (N-мерный массив, векторные данные, атрибутивная информация).

- Темпоральные БД также требуют обработки объектов – моментальных снимков состояния объекта в момент времени; темпоральные объекты должны включать в себя программный код поддержки сценариев.

- Появление новых структур данных – объектов, объединяющих в себе основные типы данных, двоичные данные или файлы, и код для проверки и манипулирования этими данными, привело к появлению термина «*объектно-ориентированные базы данных*» по аналогии с «объектно-ориентированным программированием», существующим не один десяток лет.

Разделение данных и программ является искусственным – никто не может увидеть данные без использования программ, и большинство программ управляется данными. Поэтому парадоксален факт о том, что сообщество управления данными уже 40 лет пытается достичь нечто, называемое независимостью данных – явное отделение программ от данных. В системах БД обеспечиваются два вида независимости данных, называемые физической независимостью данных и логической независимостью данных. Имеется конвергенция файловых систем, систем БД и языков программирования. В расширяемых системах БД используют объектно-ориентированные приемы из языков программирования, позволяющие определять сложные объекты как естественные типы БД. Файлы (или расширенные файлы, подобные XML) становятся тогда частью БД и получают преимущества от параллельного поиска и управления метаданными.

Литература

1. Крыщенко В.С., Голозубов О.М., Овчаренко М.М., Темников В.Н. База данных широкомасштабного почвенно-экологического мониторинга агроландшафтов: реляционный подход // Агрохимический вестник, 2010, № 1. – С. 12-16.

2. Дмитриев Е.А. Математическая статистика в почвоведении: Учебник. – М.: Изд-во МГУ, 1995. 320 с.: ил. 13ВК 5-211-02930-5.

3. Крыщенко В.С., Голозубов О.М., Колесов В.В., Рыбянец Т.В. База данных состава и свойств почв. Ростов-на-Дону: Изд-во РСЭИ, 2008.

4. Кузнецов С.Д. Основы современных баз данных, информационно-аналитические материалы Центра Информационных Технологий // www.citforum.ru/database/osbd/contents.shtml (дата обращения: 1.02.2010).

5. Общесоюзная инструкция по почвенным обследованиям и составлению крупномасштабных почвенных карт землепользования. – М.: Колос, 1973 г. 96 с.

6. Базы данных: достижения и перспективы на пороге 21-го столетия. Под ред. Ави Зильбершатца, Майка Стоунбрейкера и Джеффа Ульмана // Системы Управления Базами Данных № 3. 1996, издательский дом «Открытые системы». Новая редакция: Кузнецов С.Д., 2009 г.

7. Рожков В.А., Алябина И.О., Колесников В.М., Молчанов Э.Н., Столбовой В.С., Шоба С.А. Почвенно-географическая база данных России // Почвоведение, 2010, № 1, – С. 3-6.

8. Теория и методы физики почв / Под ред. Шеина Е.В., Карпачевского Л.О. – М.: Гриф и К, 2007. 616 с.

9. Шеин Е.В. Гранулометрический состав почв: проблемы методов исследования, интерпретация результатов и классификаций // Почвоведение, 2009, № 3. – С. 309-317.

10. Рожков В.А. Методы оценки информативности почвенных признаков и классификаций. Презентации докладов конференции «ЭКОМАТМОД-2009». lem.edu.mhost.ru/doc/presentations/Rozhkov.pdf (дата обращения: 14.09.2009).

ИНФОРМАЦИЯ

20 апреля 2009 г. АНО «Редакция «Химия в сельском хозяйстве» и ООО «РУНЭБ» заключили договор на распространение электронных копий статей, опубликованных в журнале «Агрохимический вестник», для наполнения базы данных проекта Российского индекса научного цитирования (РИНЦ), представленной в виде научного информационного ресурса сети Интернет, который можно будет приобрести на сайте **www.elibrary.ru**.

Уважаемые авторы, в связи с вышеизложенным, при направлении статей в редакцию просим Вас указывать в сопроводительном письме о согласии на передачу опубликованных работ для размещения в сети Интернет. Отсутствие данного условия будет служить основанием для отказа в публикации в журнале.