

- Костин В.С., Постнов А.В., Хасанов С.А., Диконская Е.К., Кисарова В.П. Подготовка и генерация отчета по археологическим полевым работам средствами информационных технологий: анализ проблем и постановка задач // Информационные технологии в гуманитарных исследованиях. Вып. 15. Новосибирск, 2010: с. 60-65.
- Медведев Г.И., К проблеме морфологического анализа каменного инвентаря палеолитических и мезолитических ансамблей Восточной Сибири // Описание и анализ археологических источников. – Иркутск, 1981.
- Медведев Г.И., Алаев С.Н., Алаева Т.В., Опыт применения полевой фиксации карточки на палеолитических местонахождениях // Нач.-техн. конф. Археология, этнография Восточной Сибири – Иркутск, 1978,
- Тарасенко В.П. Некоторые проблемы формализации гуманитарных знаний (на примере археологии) // Информационные технологии в гуманитарных исследованиях. Вып. 2. Новосибирск, 2000: с. 84-87.
- Фролов А.В., Фролов Г.В. Электронные конференции в Интернет // Газета "Капитал", рубрика "Три килобайта", 1996.
- Холюшкин Ю.П., Воронин В.Т., Воробьев В.В., Бердников Е.В., Федоров С.А., Жилицкая Г.Ю., Грищенко А.А., Лузин А.В. Электронный каталог научной библиотеки Института археологии и этнографии СО РАН (археология и этнография) // Информационные технологии в гуманитарных исследованиях. Выпуск 6. - Новосибирск: Редакционно-издательский Центр НГУ, 2003, с. 81-85.
- Холюшкин Ю.П., Воронин В.Т., Воробьев В.В. Информационная система по подготовке годовых научных отчетов // Информационные технологии в гуманитарных исследованиях. Выпуск 7. - Новосибирск: Редакционно-издательский Центр НГУ, 2004, с. 72-74.
- Холюшкин Ю.П., Воронин В.Т. Сектор археологической теории и информатики: итоги десятилетия // Новосибирск: РИЦ НГУ, 2005, 26 с.
- Холюшкин Ю.П., Воронин В.Т., Ильиных М.Ю., Соловьева С.А. Разработка архитектуры системы музейного портала "История и культура Северной Азии и Дальнего Востока" // Информационные технологии в гуманитарных исследованиях. Вып.12. - Новосибирск: ЗАО РИЦ "Прайс-Курьер", 2008, с. 29-36.

Беленький К.Г.,
Витяев Е.Е.¹,
Костин В.С.,
Холюшкин Ю.П.

WEB-портал статистической обработки археологических данных

Статья посвящена проблеме создания статистического пакета, предназначенного для обработки археологической информации. По замыслу авторов такой инструментарий предоставляет археологам простые для понимания единообразные средства представления, сбора, хранения и обработки данных археологических памятников. Описываемый пакет позволяет применять специализированные, ориентированные на археологические задачи методы анализа данных. Пакет также открывает возможность решать задачи археологии специально подготовленными последовательностями методов анализа данных (стратегиями решения задач). Добавление новых стратегий и методов в состав пакета не требует изменений программного кода сайта. Это позволит в дальнейшем интегрировать в состав пакета множество методов, свободно распространяемых по лицензии Open Source.

Ключевые слова: Web-интерфейс, анализ данных, стратегии решения задач, статистический пакет, Drupal, PHP, MySQL, XML, нейронные сети.

Введение

В свое время еще И. Кант сформулировал мысль, которую вслед за ним повторяли и интерпретировали многие философы, о том, что любая отрасль знания с тем большим основанием может называться наукой, чем чаще и успешнее она использует математику в собственных целях.

В предыдущей статье была рассмотрена схема построения сетевой среды – носителя информационных потоков, возникающих во время и вокруг археологических исследований. Вариант заполнения одного фрагмента этой схемы будет описан далее. Этот фрагмент включает хранение и обработку массивов, состоящих из описаний археологических находок (их информационных копий). Такая обработка требует применения сложных вычислительных методов.

Исследователям не всегда доступны такие вычислительные ресурсы в силу их дороговизны и сложности освоения. Другим недостатком существующих пакетов анализа данных является отсутствие в них инструментария для решения предметно-ориентированных задач.

Отсюда возникает задача создания инструментария, который позволял бы:

- объединить данные полевых исследований, проводимые в различных районах Сибири, в единую базу данных;
- предоставлял простые в понимании средства для единообразного сбора, хранения, обработки и представления данных об археологических находках;

¹ Работа выполнена при финансовой поддержке гранта РФФИ № 11-07-00560-а, интеграционными проектами СО РАН № 47, 115, 119, а также Советом по грантам Президента РФ и государственной поддержке ведущих научных школ, проект НШ-3606.2010.1

- позволял через web интерфейс обрабатывать полученные данные специализированными, ориентированными на археологические данные методами анализа данных;
- в отличие от имеющихся пакетов, таких как Statistica, SPSS, и т.п., не ориентированных на конкретную проблемную область, предоставить пользователю возможность решать задачи археологии специально подготовленными последовательностями методов (стратегиями решения задач) анализа данных;
- предоставлял бы возможность автоматически добавлять новые методы без переделки сайта;
- предоставлял бы возможность автоматического включения новых задач археологии и последовательностей методов для их решения без модернизации сайта.

Все эти требования предъявляют достаточно жесткие требования к используемой технологии создания web интерфейса. Ниже приведено описание используемой технологии и порядка использования интерфейса.

Используемые технологии

Описанный инструментарий было решено реализовать в виде web-портала, который с помощью пользовательского web-интерфейса позволял бы археологам:

- хранить свои данные и иметь к ним доступ из любого места, где есть выход в интернет;
- делиться результатами исследований;
- обрабатывать данные на удаленном сервере, используя различные методы анализа данных.

Для реализации серверной части портала использовался язык PHP, выбор его был связан с тем, что этот язык поддерживается почти всеми web-серверами и является очень популярным и простым в понимании, что упрощает поддержку сайта.

В качестве базы данных была выбрана СУБД MySQL 5, удовлетворяющая всем требованиям для реализации системы хранения данных и другой служебной информации системы.

Для реализации клиентской части сайта использовался язык JavaScript и библиотеки jQuery и jQuery UI, созданные на его основе, которые позволяют разработать динамичный и удобный пользовательский интерфейс, использующий технологию AJAX для реализации динамически изменяющихся web-страниц.

Также была использована платформа (а по совместительству и система управления контентом) для написания сайтов Drupal 6, написанная на языке PHP. Эта платформа была выбрана из-за своей модульности и большим количеством модулей с открытым кодом, реализованных и поддерживаемых сторонними разработчиками, которые использовались для реализации целевой функциональности и типичных разделов web-портала, таких как новостная лента, карта сайта, обратная связь и т.п.

Внутреннее описание системы

Прежде чем перейти к описанию того, как пользоваться созданным web-приложением, хотелось бы осветить в общих чертах внутреннее представление данных анализа и методов, а также затронуть некоторые аспекты внутренней реализации портала, что, несомненно, поможет в понимании общих принципов использования инструментария.

Представление данных

Данные в системе организованы в «массивы данных», которые описывают таблицу объект-признак. Каждый массив данных имеет набор характеристик, предоставляющих дополнительную информацию о данных:

- Заголовок – название массива данных.
- Короткий заголовок – укороченное название, нужно для более удобного отображения массива данных в списке.
- Источник – текстовая информация о источнике данных, которую может предоставить владелец массива данных.
- Редактор – поле, хранящее информацию о последнем редакторе массива данных или описания массива данных.
- Время редактирования – время последнего редактирования массива данных.
- Количество объектов – количество наблюдений, представленных в исходной таблице объект-признак.
- Значение пропуска – значение по умолчанию кода пропуска во всех признаках массива.
- Изображения – графические изображения, которые могут дать дополнительную информацию о массиве данных.

Каждый массив данных состоит из набора признаков, которые хранят значения столбцов исходной таблицы объект-признак. Каждый признак также имеет набор характеристик, которые описывают соответствующий столбец данных:

- Информация – текст, с помощью которого владелец массива данных может описать признак.
- Шкала – шкала, в которой приведены значения признака. Может быть номинальной, порядковой и количественной.
- Количество уникальных значений.
- Общее количество значений.
- Среднее значение (имеет смысл только при количественной шкале признака).
- Стандартное отклонение (имеет смысл только при количественной шкале признака).

Перечисленные характеристики может заполнить создатель массива данных, и они будут доступны для просмотра другим пользователям сайта.

Методы, их добавление, запуск и результаты

Важным требованием к сервису является возможность добавления методов обработки данных без привлечения разработчика сайта, что позволяет сторонним разработчикам расширять спектр используемых на сайте методов.

Для удовлетворения этого требования была реализована возможность включения методов в систему путем загрузки исполняемого файла и файлов используемых библиотек. Естественно, наряду с самим приложением, разработчику нужно составить и загрузить файл описания входов и выходов исполняемого файла в формате XML, для чего был разработан язык на базе XML, с помощью которого описываются:

- Форма настройки метода:
 - Вид элементов формы, такие как текстовое поле, выбор из нескольких вариантов, галочка и элемент указания признаков в качестве входных данных.
 - Допустимые значения, вводимые в элементы формы, которые описываются:
 - Набором «интервалов» допустимых значений, которые могут быть фиксированными значениями, каждый конец интервала может быть открытым или закрытым и содержать плюс/минус бесконечность.
 - Набором допустимых шкал вводимых признаков.
 - Общим количеством значений в столбцах вводимых признаков.
 - Количеством значений, вводимых в элементы формы.
 - Типом значений, вводимых в элементы формы.
 - Реакцию на ввод пользователем значений в форму:
 - Активация/деактивация элементов формы для указания, какие еще поля нужно заполнить, а какие нет.
 - Корректировка введенных значений в форме для реализации помощи введения корректных значений.
 - Подсказки пользователям.
- Настраиваемые параметры, которые будут передаваться методу, значения которых определяет пользователь посредством формы.
- Способ интерпретации полученных результатов (добавление нового признака в массив данных, заполнение характеристики признака либо просто запись в отчете).

В итоге, для добавления нового метода в систему, администратору понадобится исполняемый файл с файлами библиотек и XML-файл описания метода. Для начального наполнения базы методов использовалась библиотека ALGLIB.

Запуск метода проводится в фоновом режиме и при помощи `cron` (демон-планировщик задач в UNIX-подобных операционных системах, использующийся для периодического выполнения заданий в заданное время) периодически проверяется, сгенерировался ли выходной файл с результатами, и если это так, то результаты сохраняются в систему. Для того, чтобы проверка выходного файла проводилась чаще, Если пользователь, который запустил метод, все еще находится на сайте, то проверка выходного файла проводится намного чаще, для того, чтобы обеспечить быструю реакцию системы на окончания работы метода.

Как говорилось ранее, результаты метода могут быть представлены тремя способами:

- Результат является новым признаком в массиве данных. В этом случае столбец значений признака сохраняется как дополнительный признак в массиве данных, с указанным в настройках файла метода заголовком и шкалой, указанной в описании результата. Также результат сохраняется в отчете запуска метода в виде строки с названием нового признака и его значениями.

- Результат является значением характеристики признака. В этом случае вычисленное значение характеристики сохраняется в описание признака, а также в отчет запуска метода в виде строки, включающее в себя результат.
- Результат является значением, которое должно сохраниться только в отчете запуска метода.

Для добавления способа интерпретации результирующего значения реализована система расширения функциональности добавления записей в отчет, которая позволяет разработчикам добавлять свои РНР-функции обработки результатов и необходимые JavaScript-функции для их отображения в отчете.

Пользовательский интерфейс

Важным требованием к сервису являются простота использования конечным пользователем, то есть простота понимания им пользовательского интерфейса. Для реализации отображения и управления данными и методами был разработан модуль, который был назван Taxonomy Content Browser. Этот модуль использует технологию AJAX для быстрого и удобного просмотра категоризированного содержания без перезагрузки страницы, что позволило расположить части интерфейса, относящиеся к управлению данными и запуску методов, на одной web-странице.

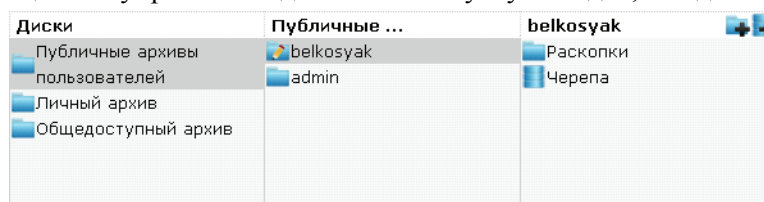


Рис. 1. Личный открытый архив пользователя belkosyak.

Управление данными

В системе существует четыре раздела (диска), к которым относится конкретный массив данных:

- Личный архив. В этом разделе пользователь может создавать массивы данных и папки, до которых он имеет полный доступ. Другим пользователям системы не разрешается просматривать/редактировать массивы данных других пользователей. Анонимные пользователи системы не имеют доступ до этого раздела.
- Личный открытый архив. Для каждого пользователя в этом разделе существует папка, в которой пользователь может создавать массивы данных и папки, до которых он имеет полный доступ (рис.1). Другим пользователям системы и анонимным пользователям разрешается просматривать и копировать массивы данных других пользователей, помещенных в этот раздел.
- Общедоступный архив. В этом разделе пользователь может создавать массивы данных и папки, до которых он имеет полный доступ. Другим пользователям системы и анонимным пользователям разрешается просматривать и копировать массивы данных других пользователей, помещенных в этот раздел.
- Временный архив. В этом разделе анонимные пользователи могут создавать массивы данных и папки, которые сохраняются там в течении HTTP-сессии (рис.2). Созданные массивы данных и папки могут изменять только их создатели, другие пользователи системы не имеют к ним доступ. Зарегистрированные пользователи системы не имеют доступ до этого раздела.

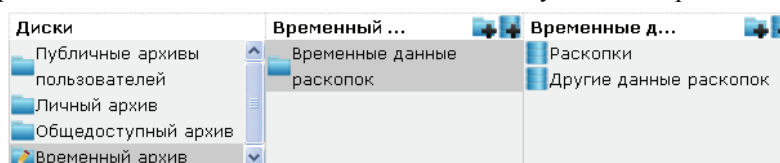


Рис. 2. Временный архив анонимного пользователя.

Управление данными осуществляется с помощью реализованного контекстного меню, которое позволяет выполнить следующие операции (рис.3):

- Добавление массива данных. Данные добавляются путем загрузки CSV-файла, либо вручную (скопировав, например, из Excel).
- Добавление папки в текущий раздел/папку.
- Удаление признака/массива данных/папки.
- Редактирование значений массива данных.
- Скачивание массива данных в формате CSV.
- Просмотр описания массива данных и его редактирование.
- Редактирование признака.
- Переименование признака/массива данных/папки.

- Копирование/вставка признаков/массивов данных.

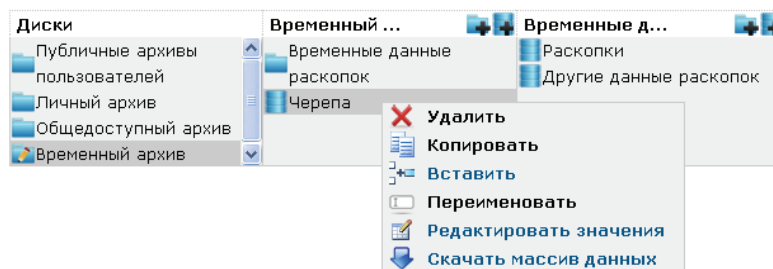


Рис. 3. Контекстное меню операций с массивом данных.

Использование методов

Методы организованы в категории/подкатегории, для каждой из категорий, подкатегорий и методов администратор может добавить описание для ознакомления пользователей с этой категорией/подкатегорией/методом.

При выборе пользователем определенного метода, ему становится видна форма запуска метода, которую он может заполнить и запустить выполнение метода. Заполнение полей, которые являются полями указания признаков, использующихся в качестве входных данных для запуска метода, происходит с помощью подхода Drag&Drop, т.е. пользователь попросту «перетаскивает» признак из навигатора по данным (рис.4).

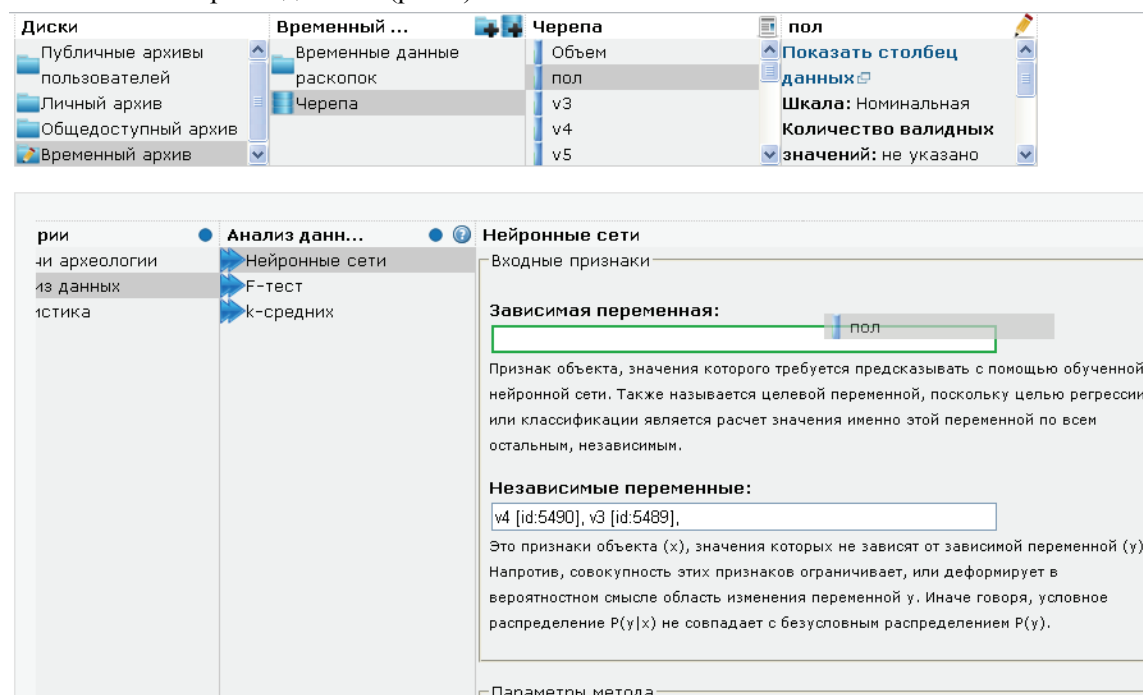


Рис. 4. Применения подхода Drag&Drop к заполнению формы метода.

В случае прохождения валидации элементов формы становится активной кнопка запуска метода, в противном случае кнопка запуска метода деактивируется и выводятся сообщения с указанием, какие поля формы заполнены неправильно.

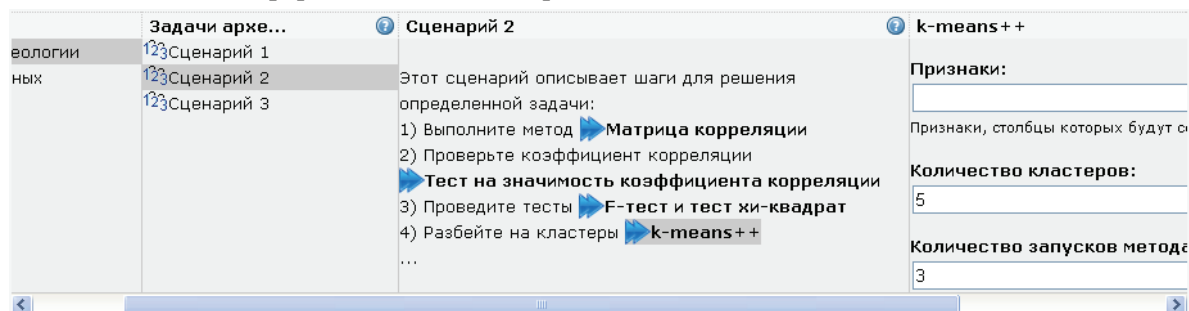


Рис. 5. Использование сценария.

Также из методов администратор сайта может конструировать сценарии решения определенных задач предметной области (в нашем случае археологии). Сценарий представляет из

себя текст описания последовательности шагов для достижения цели, а также вставленными в текст ссылками на методы. При нажатии на такую ссылку появляется форма настройки метода с предзаполненными значениями элементов формы, отвечающих за определенные параметры метода (рис. 5).

После того, как выполнится метод, система разбирает полученный на выходе у исполняемого файла описание результатов, подготовленного в виде XML. Когда текст отчета готов, система сохраняет отчет в подпапку массива данных, причем отчеты в рамках одного массива данных организованы по названиям метода, датам и времени получения результатов работы метода (рис.6). Причем текст отчета также может быть HTML кодом, который может использоваться для визуализации выходных данных метода.

Черепа123	[Reports]	Среднее и с...	24.05.11	09:50:21
[Reports]	Нейронные сети	20.05.11	09:20:51	Среднее = 1421.11428
Объем	Среднее и	24.05.11	09:47:21	=====
пол	стандартное		09:49:51	Стандартное отклоне
v4	отклонение		09:50:21	169.319012
v5				

Рис. 6. Организация отчетов в массиве данных.

Общая информация о структуре сайта

Сайт состоит из шести разделов:

- Главная страница. На этой странице находится описание сайта и последние новости.
- Анализ. На этой странице находятся навигаторы по данным и методам.
- Новости. Список урезанных текстов новостей со ссылками на страницы, на которых отражается полный текст новости.
- Описание методов. Список урезанных описаний методов со ссылками на страницы, на которых отражается полный текст описания.
- Карта сайта. Раздел со ссылками на основные разделы сайта.
- Обратная связь. Страница для отправки письма администратору сайта.

Также реализовано представление новостей в формате RSS-ленты. Интерфейс сайта реализован на двух языках: английском и русском.

Заключение

В результате проведенной работы был реализован сайт, который позволяет загружать данные, управлять ими и скачивать их с сервера. Разработана система доступа к данным, позволяющая пользователям иметь личный закрытый и открытый архив данных и размещать данные для общественного использования. Реализована возможность добавления методов без участия разработчика сайта, для чего был разработан язык описания формы настройки метода на базе XML и набор классов на языке C++, предоставляющих средства для написания методов, пригодных для встраивания в систему. Также система предоставляет возможность группировать методы для решения определенных задач археологии и предоставляет средства для их описания.

Созданная система позволяет решать проблемы, присущие существующим пакетам обработки данных и предоставляет возможность расширения новыми возможностями для разработчиков.

ЛИТЕРАТУРА

- Холушкин Ю.П. К вопросу об оценке характеристик археологического научного знания // Окно в неведомый мир: Сб. статей к столетию со дня рождения академика А.П. Окладникова- Новосибирск: Изд-во Ин-та археологии и этнографии СО РАН, 2008.; с. 96-103.
- Холушкин Ю.П., Воронин В.Т., Костин В.С. Концептуальные подходы к созданию on-line статистического пакета анализа археологической информации с элементами картографии на сайте "Sibirica" // Информационные технологии в гуманитарных исследованиях / Ин-т археологии и этнографии СО РАН. - 2008. - Вып. 12. - С. 50-53.
- Материалы сайта drupal.org
- Материалы сайта jquery.com
- Материалы сайта alglib.sources.ru