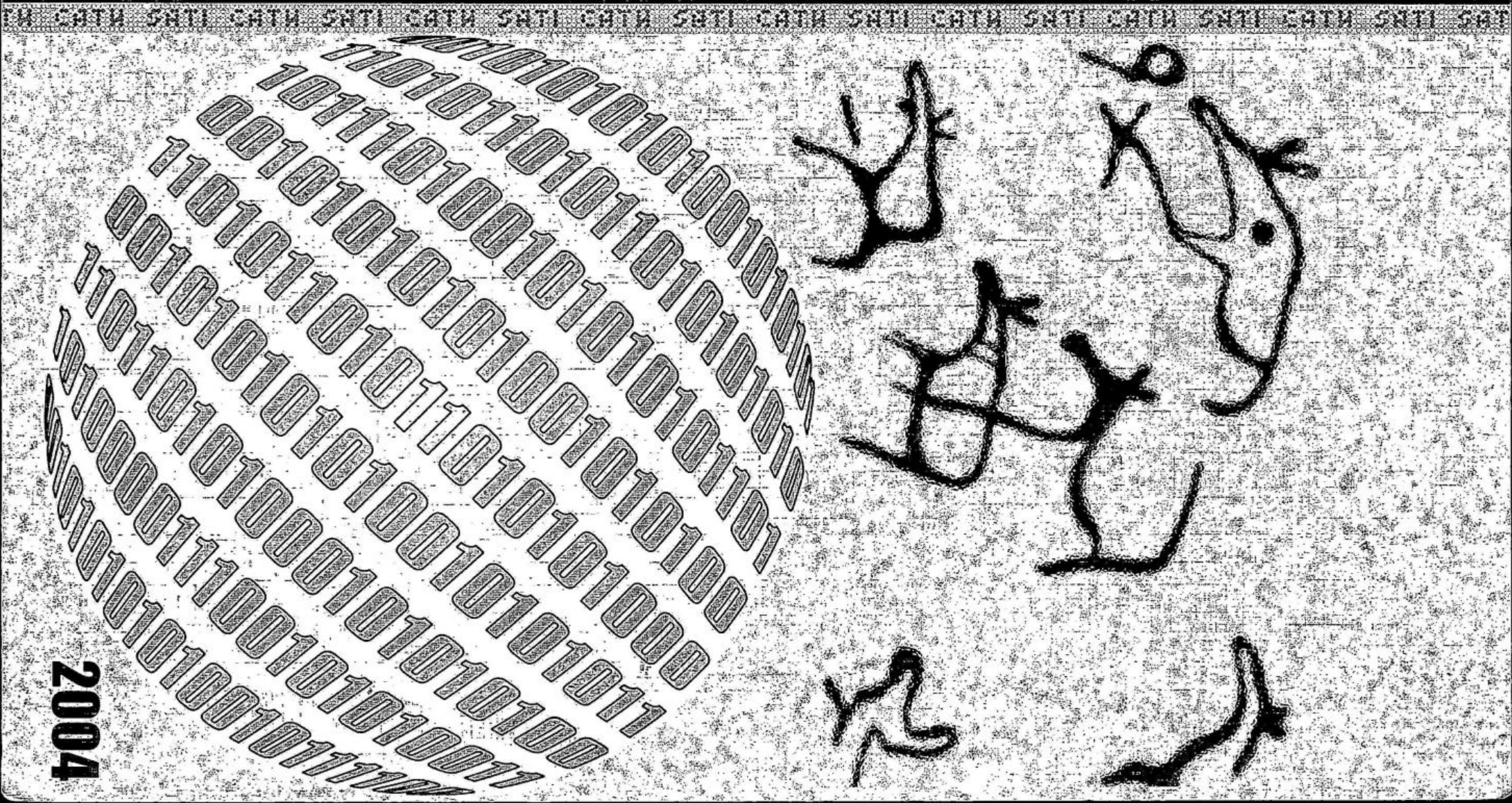
Информационные технологии в гуманитарных исследованиях



2004

РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ АРХЕОЛОГИИ И ЭТНОГРАФИИ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ

Выпуск 8

РАЗРАБОТКА НОВЫХ МЕТОДОВ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ПРЕДСТАВЛЕНИЯ И ОБРАБОТКИ АРХЕОЛОГИЧЕСКИХ И ЭТНОГРАФИЧЕСКИХ ДАННЫХ

Материалы научного отчета
по интеграционной программе СО РАН
за 2003-2004 гг. (проект № 149)

Ответственный редактор
академик РАЕН, доктор исторических наук *Ю. П. Холюшкин*

Новосибирск
2004

ББК 60
И 74

Авторы:
Марчук А.Г., Холюшкин Ю.П., Загорулько Ю.А., Воронин В.Т., Андреева О.А., Бердников Е.В., Боровикова О.И., Булгаков А.Н., Воробьев В.В., Загорулько Г.Б., Илларионов В.А., Ильиных М.Ю., Корнюхин Ю.Г., Костин В.С., Костов Ю.В., Нариньяни А.С., Нуртдинов А.Н., Сидорова Е.В.

Издание осуществлено при финансовой поддержке интеграционной программы № 149
Сибирского Отделения РАН

ISBN 5-94356-220-6

И 74 **Информационные технологии в гуманитарных исследованиях: Выпуск 8. Разработка новых методов и информационных технологий представления и обработки археологических и этнографических данных.** Материалы научного отчета по интеграционной программе СО РАН за 2003-2004 гг. (проект № 149). Новосибирск: Редакционно-издательский центр НГУ, 2004. с. 67

Настоящий выпуск представляет собой коллективную монографию с материалами научного отчета по итогам двухлетнего совместного исследования Института археологии и этнографии СО РАН, Института систем информатики СО РАН, Института искусственного интеллекта Минсвязи РФ и кафедры систем информатики НГУ. В монографии излагаются подходы к подготовке, созданию, обработке и представлению информации в археологии и этнографии. Выпуск рассчитан на археологов, историков, этнографов и на широкий круг исследователей, интересующихся информационными технологиями в гуманитарных исследованиях и образовании.

ISBN 5-94356-220-6

ББК 60

СОДЕРЖАНИЕ

ОТ РЕДАКТОРА

В настоящей монографии представлены материалы научного отчета о выполнении интеграционной междисциплинарной программы СО РАН "Разработка новых методов и информационных технологий представления и обработки археологических и этнографических данных" в 2003-2004 гг.

В монографию включены результаты исследований по разработке информационных ресурсов для археологии и этнографии, предназначенные для накопления, систематизации, сохранения и организации широкого доступа к информации о культуре, традициях и искусстве Северной Азии. Исследования и разработки выполнены совместно коллективами научных сотрудников Института археологии и этнографии СО РАН, Института систем информатики СО РАН, Института искусственного интеллекта Минсвязи РФ и кафедры систем информатики НГУ.

Соруководители проекта:
д.и.н. Холюшкин Ю.П.
д.ф.-м.н. Марчук А.Г.
к.т.н. Загорулько Ю.А.

Материалы отчета написаны и подготовлены к публикации авторским коллективом в составе:

Институт археологии и этнографии СО РАН:
Холюшкин Ю.П.
Воронин В.Т.
Костин В.С.
Воробьев В.В.,
Корнюхин Ю.Г.,
Нуртдинов
Бердников Е.В.
Илларионов В.А.
Ильиных М.Ю.,

Институт систем информатики СО РАН и Институт искусственного интеллекта Минсвязи РФ:
Марчук А.Г.
Загорулько Ю.А.
Костов Ю.В.
Булгаков А.Н.,
Сидорова Е.В.
Нариньяни А.С.
Боровикова О.И.
Загорулько Г.Б.
Андреева О.А.,
Сидорова Е.В.

Ю.Холюшкин

ВВЕДЕНИЕ

Обоснование необходимости проведения исследований

В настоящее время накоплен большой объем знаний и информационных ресурсов по археологии и этнографии. Однако эти данные мало систематизированы и рассредоточены по различным сайтам Интернет, библиотекам и архивам, что существенно ограничивает к ним доступ.

Большая часть архивных, опубликованных или представленных в Интернет количественных данных археологических и этнографических исследований, зафиксирована в плохо согласованной фрагментарной форме. Поэтому для систематизации этих информационных ресурсов и обеспечения удобного доступа к ним требуется их обобщение с помощью современных методов и инструментов представления, комплексной обработки и анализа, в первую очередь, путем адаптации и применения адекватных методов прикладной математики и информатики.

Другой актуальной проблемой доступа к накопленным данным и знаниям по археологии и этнографии является их оформление в виде информационных ресурсов в наиболее удобном представлении, обеспечивающем удаленный доступ к релевантному информационному контенту с поддержкой навигации в Интернет.

Для решения этой задачи необходимо разработать специализированный Интернет-портал, обеспечивающий содержательный доступ к информационным ресурсам по археологии и этнографии. Актуальность создания такого Интернет-портала определяется пониманием того, что российская наука, образование и культура испытывают в наши дни потребность в концентрации и обобщении накопленной информации по гуманитарным наукам и эффективном ее использовании. Удовлетворение этой потребности затрудняется тем, что в силу многоплановости и многоаспектности информационные ресурсы гуманитарного направления в пределах российского подпространства Интернет рассредоточены на удаленных страницах множества сайтов, основная тематическая направленность которых относится скорее к естественнонаучной, технической и технологической сфере, нежели к гуманитарной. Специализированный Интернет-портал и призван решить задачу сведения информационных ресурсов по археологии и этнографии в единое адресное пространство и обеспечить возможность открытого удобного доступа к ним.

Сложившиеся тенденции и современный уровень решения проблемы в стране и за рубежом

В настоящее время методы и инструменты представления, комплексной обработки и анализа данных, в первую очередь, методы прикладной математики и информатики активно применяются в социологии, экономике и инженерии. В мировой же и отечественной археолого-этнографической исследовательской практике традиционные статистические и новые методы применяются мало, вследствие чего накопленные знания, достижения и открытия в этой области пока еще недостаточно освоены для их внедрения в научно-познавательный и культурно-образовательный процесс.

В то же время следует заметить, что традиционные статистические методы, применяемые для обработки данных, не всегда дают должный эффект. Поэтому требуется разработка и применение новых современных подходов, основанных, в частности, на так называемых мягких вычислениях, а именно – на нейронных сетях и методах программирования в ограничениях.

Не решена до сих пор также задача сведения гуманитарных ресурсов, в частности, археологических и этнографических, в единое информационное пространство с обеспечением возможности открытого удобного доступа к ним.

Оценка проделанной работы в этом направлении в СО РАН

Развитие информационно-телекоммуникационной инфраструктуры СО РАН создало все предпосылки к развертыванию работ по созданию, накоплению и обработке информационных ресурсов Института археологии и этнографии СО РАН. Исследовательские коллективы Сибирского отделения РАН занимают ведущие позиции в развитии этого перспективного научно-практического направления разработок.

В ИАЭТ СО РАН на протяжении многих лет ведутся работы по систематизации накопленных знаний по археологии различных регионов Евразии, представленных в опубликованных источниках или размещенных в различных архивах и коллекциях. Для обработки и обобщения собранного материала привлекаются оригинальные и адаптированные современные методы анализа данных (выделение и

анализ связных областей, типологический анализ, факторный анализ, кластерный анализ, многомерное шкалирование, анализ устойчивости выделенной структуры данных и др.) и информатики.

В ИСИ СО РАН ведутся исследования, направленные на разработку средств представления знаний о предметных областях и релевантных им информационных ресурсах на основе популярного в настоящее время онтологического подхода. Разрабатывается подход к организации специализированных порталов знаний, которые должны обеспечивать удаленный доступ к определенному информационному контенту, в том числе, через Интернет. Оригинальность данного подхода состоит в том, что такие порталы знаний будут обеспечивать доступ не только к собственным информационным ресурсам, но и поддерживать навигацию по заранее размеченным ресурсам, размещенным в сети Интернет.

В ИСИ СО РАН и РосНИИ ИИ разрабатываются новые современные подходы к обработке неполных и неточных данных, в основе которых лежит применение нейронных сетей и методов программирования в ограничениях (в частности, метод недоопределенных вычислительных моделей).

Цели и предполагаемые результаты исследований

Целью проекта является разработка новых методов и информационных технологий представления и обработки археологических и этнографических данных.

Одной из первых задач проекта является разработка специализированного Интернет-портала, обеспечивающего содержательный доступ к информационным ресурсам по археологии и этнографии.

В основу этого портала знаний будет положена онтология, содержащая наряду с традиционным описанием предметной области соотнесенное с ним описание структуры и типологии соответствующих сетевых ресурсов.

Главным преимуществом данного подхода является то, что порталы знаний позволяют значительно сократить время обработки запроса пользователя и количество выдаваемых ресурсов за счет более точного определения степени их релевантности и хранения ссылок на них непосредственно на портале знаний. Причем сами ссылки автоматически накапливаются специальным модулем – коллекционером онтологической информации о ресурсах.

Другой частью проекта является разработка новых математических и статистических методов и технологий обработки археологических данных.

Для решения перечисленных задач в проекте выделены два блока.

1. Разработка информационных технологий представления и манипулирования данными по археологии и этнографии в среде Интернет.

2. Разработка новых методов и технологий представления и обработки археологических данных.

Предполагаемые результаты проекта – портал знаний по археологии и этнографии с ядром в виде информационных ресурсов с наполнением данными в систематизированной обобщенной форме.

Имеющаяся материально-техническая база, ее соответствие поставленным задачам

Техническое обеспечение: два сервера Pentium III 850 MHz, 1000 Mb RAM и Pentium III 500 MHz, 512 Mb RAM, с каналом доступа к Интернет.

Программное обеспечение: OS Linux, WWW-сервер Apache, Z39.50-сервер Zoopark, информационно-поисковая система Isite.

На WWW-сервере развернуты информационные ресурсы: база электронных словарей и энциклопедий (15 электронных баз данных), система классификации археологической науки, электронные издания (электронные журналы, обозрения, вестники, другие электронные периодические издания, монографии, сборники научных трудов, отдельные статьи, авторефераты диссертаций и др.), виртуальный музей "Древняя история, культура и искусство Северной Азии", электронный каталог научной библиотеки ИАЭТ СО РАН, гео-информационные системы и др.

Материально-техническая база полностью покрывает потребности в оборудовании, необходимом для выполнения проекта.

I ИНФОРМАЦИОННЫЕ РЕСУРСЫ В АРХЕОЛОГИИ И ЭТНОГРАФИИ

1. Интеллектуальный интернет-портал знаний для доступа к информационным ресурсам по археологии и этнографии

Для решения задачи сведения ресурсов, относящихся к одной области знаний в единое информационное пространство, обеспечения возможности открытого и удобного доступа к ним и поддержки их целостности нами предложена концепция специализированных Интернет-порталов знаний [Боровикова, Загорулько, 2002: 76-82]. На этой концепции основана разрабатываемая нами технология создания и сопровождения порталов знаний по гуманитарным наукам.

Информационную основу таких порталов знаний составляют онтологии [Gruber, 1993; Genesereth & Nilsson, 1987; Ushold, Gruninger, 1996; Ushold, King, 1995; Takeda, Takaai, & Nishida, 1998; Guarino, 1997: 293-310], включающие как описание науки и научной деятельности в целом, так и описание конкретной научной дисциплины и соотнесенное с ним описание структуры и типологии соответствующих хранилищ данных и сетевых ресурсов [Боровикова, Загорулько, 2002: 76-82; Жигалов, Загорулько, Нариньяни, Россеева, 2002: 29-71].

Благодаря предложенной структуризации системы знаний, когда явно выделяются предметно–независимые онтологии науки и научного знания, являющиеся общими для всех гуманитарных наук, портал знаний становится легко настраиваемым на выбранную предметную область. Так, при построении портала знаний для определенной гуманитарной дисциплины достаточно только построить ее онтологию и связать ее с предметно-независимыми онтологиями и соответствующими информационными ресурсами.

Для настройки портала на конкретного пользователя в состав информационной части портала включена модель пользователя, которая может постоянно уточняться и расширяться и тем самым всегда отражать актуальный "информационный портрет" пользователя.

В данной работе обсуждается основанный на предложенной концепции [Боровикова, Загорулько, 2002: 76-82] подход к разработке специализированного Интернет-портала, обеспечивающего содержательный доступ к систематизированным знаниям и информационным ресурсам по археологии и этнографии.

1.1. Назначение и основные функции портала знаний

Портал знаний представляет собой, с системной точки зрения, специализированную информационную систему, снабженную эргономичным пользовательским web-интерфейсом.

С точки зрения пользователя, портал является тематическим Интернет-ресурсом, обеспечивающим возможность поиска и просмотра информации в рамках заданной предметной области (гуманитарной дисциплины).

Как информационный ресурс портал:

– обеспечивает доступ к информации по различным аспектам и участникам научной деятельности, таким как: составляющие научной дисциплины (подразделы дисциплины, методы исследования, используемые термины и понятия), персоналии исследователей, информация по группам, сообществам, организациям, включенным в процесс исследования;

– позволяет интегрировать близкие по тематике ресурсы, представленные в Интернет, и локальной сети;

– предоставляет средства поиска интересующей пользователя информации в рамках всего информационного пространства портала;

– обеспечивает информационную поддержку пользователей ресурса (например, анонсирование разного рода событий и мероприятий);

– поддерживает гибкий пользовательский интерфейс, позволяющий учитывать предпочтения пользователя по работе с ресурсом и предоставляемыми сервисами.

1.2. Модель информационного наполнения портала

Информационную основу портала составляет онтология и соотнесенное с ней описание соответствующих сетевых ресурсов. Перед тем как описывать онтологию портала и его сетевые ресурсы, поясним, что мы понимаем под онтологией.

Понятие "онтология", заимствованное из философии, сейчас активно применяется в информатике и искусственном интеллекте. Напомним, что в философии онтология – это учение о бытии, о его категориях, формах и фундаментальных принципах. Мы полагаем, что одной из целей онтологии является описание и изучение сущностей, существующих в реальном мире и/или сознании человека.

Для систем искусственного интеллекта (ИИ) "существует" только то, что уже в них представлено или может быть представлено, поэтому в области ИИ самым распространенным определением онтологии является определение, данное в работе Томаса Грубера [Gruber, 1993]. Согласно нему, онтология является явной спецификацией концептуализации. Причем под концептуализацией понимается некоторая абстракция, т.е. упрощенное представление мира, построенное для определенной цели. Концептуализация включает все объекты, понятия и другие сущности, которые предполагаются существующими (и соответственно, учитываются) в рассматриваемой области, а также все значимые отношения между ними [Genesereth & Nilsson, 1987]. С этой точки зрения каждая база знаний, система ИИ или интеллектуальный агент явно или неявно фиксируются некоторой концептуализацией.

В работах М.Ашольда, М.Грюнингера и М.Кинга [Ushold, Gruninger, 1996; Ushold, King. 1995] подчеркивается, что онтология есть явное описание концептуализации или некоторой ее части. Она может иметь различные формы, но должна обязательно включать словарь терминов (понятий) и их определения. Отология также задает связи между понятиями, что в совокупности накладывает структуру на предметную область и ограничивает возможные интерпретации терминов. Кроме того, онтология фактически всегда является отражением распределенного понимания предметной области, с которым согласно некоторое сообщество специалистов или программных агентов. Такое соглашение способствует точной и эффективной передаче смысла, которое, в свою очередь, ведет к таким преимуществам, как повторное и распределенное использование онтологии.

Таким образом, в контексте ИИ основу онтологии составляет множество (словарь) представленных в ней терминов. В такой онтологии определения связывают имена сущностей предметной области (понятий, классов, атрибутов, отношений) с текстами на естественном языке, описывающими, что означают эти имена, и формальными аксиомами, ограничивающими интерпретацию и корректное использование терминов.

Заметим, что при таком подходе понятие онтологии сильно пересекается с уже принятым в информатике и лингвистике понятием тезауруса.

Действительно, тезаурус можно считать одним из видов онтологии, так как в нем тоже представлены понятия и отношения между ними. Однако, в связи с тем, что в тезаурусе особое внимание уделяется способам лексического представления понятий, а отношения в тезаурусе отражают, главным образом, лингвистические связи между понятиями, то его можно рассматривать как лингвистическую онтологию. Примером такой онтологии является известный словарный ресурс WordNet. Другими словами, тезаурус описывает предметную область с точки зрения ее представления лексическими единицами того или иного естественного языка, а онтология описывает семантику предметной области в терминах понятий и отношений между ними, отвлекаясь, насколько это возможно, от того, как они выражаются в языке.

Таким образом, онтология, прежде всего, обеспечивает согласованный словарь терминов для взаимодействия субъектов (например, людей, интеллектуальных агентов, программ и т.д.) в рамках некоторой предметной области. В связи с этим некоторые онтологии используются как словарь для спецификации базы знаний. На практике часто трудно провести четкие границы между онтологией и базой знаний, если они обе специфицированы на одном и том же языке. Различия могут быть в том, какая часть знаний является распределенной и согласованной, а какая более специфичной. В частности, эти различия проявляются в том, что в онтологии присутствуют только утверждения, не зависящие от конкретной ситуации, так как предполагается, что они всегда истинны для сообщества пользователей в силу согласованности значений используемого словаря. В то время как база знаний может включать факты и утверждения, связанные с определенной ситуацией, необходимые для решения задач или обеспечения ответов на произвольные вопросы, относящиеся к данной предметной области.

Таким образом, онтология обеспечивает понятийную структуру, каркас предметной области, а база знаний наполняет его конкретными знаниями, необходимыми при решении задач.

В работе Х.Такеды, М.Такаи и Т.Нишиды [Takeda, Takaai & Nishida, 1998] делается упор на то, что онтологии должны помочь в решении проблем, возникающих из– за того, что в разных областях существуют различные интерпретации одних и тех же терминов. В этой связи онтология рассматривается как соглашение о некоторой области интересов для достижения определенных целей.

Для установления соглашения о знаниях, представленных на некотором, в частности, логическом языке, по мнению N.Guariano [Guarino, 1997: 293-310], онтология должна характеризовать

концептуализацию, ограничивая возможные значения предикатов и функций. В его понимании, онтология – это логическая теория, аксиомы которой ограничивают интерпретации нелогических символов логического языка при его использовании для представления знаний.

Суммируя приведенные выше определения онтологии, можно сказать, что онтология представляет собой точное описание (модель) некоторой части мира применительно к конкретной области интересов.

Таким образом, онтология – это четверка вида <C, D, R, A>, где

C – множество понятий конкретной предметной или проблемной области;

D – множество определений понятий;

R – множество отношений (связей) между понятиями;

A – множество аксиом.

Таким образом, онтология представляет собой систему, описывающую структуру определенной проблемной области, и состоящую из множества классов понятий, связанных отношениями, их определений и аксиом, задающих ограничения на интерпретацию этих понятий в рамках данной проблемной области.

Онтология, как пример общего соглашения о семантике области, способствует установлению корректных связей между значениями элементов этой области, тем самым, создавая условия для их совместного использования.

По нашему мнению, применение онтологий для совместного использования знаний в такой распределенной и динамичной среде как Интернет, вполне обосновано. Здесь, если потребуется, онтологии, построенные разными сообществами и/или для разных предметных областей, могут быть тем или иным образом объединены в одну онтологию и использоваться, в частности, при построении запросов к информационно-поисковым системам.

Таким образом, на основе онтологий можно создать такую сетевую структуру, в которой пользователи могут быть обеспечены "абстрактным представлением" информационного пространства интересующей их предметной области.

Для достаточно полного и целостного представления пользователя о выбранной отрасли знаний онтология портала знаний объединяет следующие относительно независимые онтологии: 1) онтологию науки, 2) онтологию научного знания и 3) онтологию предметной области, описывающую конкретную гуманитарную дисциплину (археологию и этнографию).

Такое структурирование системы знаний в виде онтологий, большая часть которых является предметно независимыми, значительно упрощает настройку портала на выбранную область научных знаний.

Онтология науки основана на предложенной В.Беньяминсом и Д.Фензелом [Benjamins, Fensel, 1998] онтологии, служащей для описания научно-исследовательских проектов, и является ее развитием. В частности, она расширена набором понятий, характерных для гуманитарных наук. Онтология науки включает следующие классы понятий, относящиеся к организации научной деятельности:

Ученые. К этому классу относятся понятия, связанные с субъектами научной деятельности: исследователями, сотрудниками и членами организаций, исторически значимыми персонажами и другими людьми.

Организации. Понятия этого класса описывают различные организации, научные сообщества и ассоциации, институты, исследовательские группы и другие объединения.

События. К событиям относятся такие понятия, как собрания, семинары, конференции, исследовательские поездки и экспедиции.

Публикации. Этот класс служит для описания различного рода публикаций и материалов, представленных в печатном или электронном форматах (монографии, статьи, отчеты, труды конференций, периодические издания, фото- и видеоматериалы и др.).

Деятельность. В этот класс входят понятия, описывающие научно-организационную или научно-исследовательскую деятельность – проекты, программы и т.п.

Онтология научного знания содержит метапонятия, задающие структуры для описания рассматриваемой предметной области. К ним относятся:

Раздел науки. Этот класс позволяет структурировать науку, выделять в ней значимые разделы и подразделы.

Метод исследования. Данный класс служит для описания различных методов исследования, используемых в описываемой дисциплине.

Объект исследования. Понятия этого класса задают типизацию объектов исследования и структуры для их описания. В гуманитарных науках объектами исследования могут выступать как сам человек, общество или государство, так и различные объекты, созданные человеком в результате его деятельности.

Научный результат. К этому классу относятся такие понятия, как открытия, новые законы, теории и методы исследования. Обычно научные результаты находят свое отражение в публикациях.

Онтология предметной области описывает конкретную гуманитарную дисциплину в целом как раздел науки и включает формальное и неформальное описание понятий и отношений между ними. Эти понятия являются реализациями метапонятий онтологии научного знания. Так, если мы возьмем такую гуманитарную дисциплину как археология, то конкретными реализациями метапонятия раздел науки будут такие: археология, полевая археология и др. Причем эти понятия будут упорядочены в иерархию общее – частное и часть – целое. Для гуманитарных наук очень важны методы исследования и объекты исследования.

В частности, в археологии методам исследования будут соответствовать такие понятия как раскопки, разведка, а в качестве объектов исследования будут выступать памятники, орудия труда и другие артефакты.

Онтология предметной области опирается на словарь-тезаурус естественно-языковых терминов, описывающих ее значимую лексику. Существующие связи между терминами тезауруса и понятиями онтологии создают предпосылки для их совместного использования при поиске и обработке информации. Преимущества данного подхода обсуждаются А.С.Нариньяни [Нариньяни, 2002: 307-313].

На рис. 1 представлен эскиз общей схемы онтологии портала знаний для такой гуманитарной науки, как археология. Он включает онтологии науки и научного знания, а также соотнесенный с ними фрагмент онтологии археологии. На данной упрощенной схеме показаны не все связи, существующие между изображенными на ней понятиями. В частности, не указаны связи, отвечающие за информационное наполнение ресурсов (информационный ресурс "описывает" событие), связи с объектами исследования ("кто открыл", "в какой экспедиции открыт"), ассоциативные связи, не показана также иерархия разделов науки и научных направлений. Но в целом схема отражает основные понятия онтологии портала и связи между ними и является основой для построения полной модели.

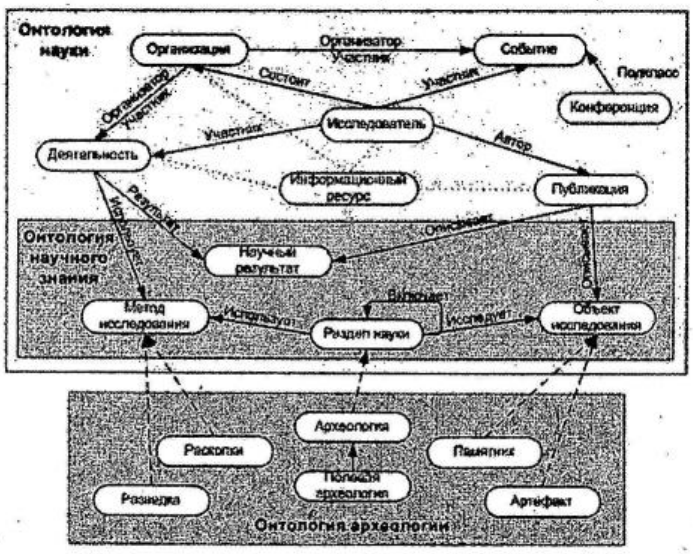


Рис. 1. Упрощенная схема онтологии портала.

Важным компонентом информационного наполнения портала является описание информационных ресурсов. Описание информационного Интернет-ресурса включает специфические атрибуты и связи, определяющие его взаимоотношения с элементами онтологии. Набор атрибутов и связей основан на стандарте Dublin Core [Жигалов, Загорулько, Нариньяни, Россеева. 2002: 29-71; Using Dublin Core] и включает такие элементы как:

- Название ресурса.
- Тематическая направленность ресурса. Указывает на тематику содержания ресурса и может связывать ресурс с объектами разных онтологических типов.

- Тип ресурса. Определяет тип ресурса (Интернет-сайт, база данных, отдельный документ) и формат представления данных.
- Язык содержания ресурса.
- Права доступа.

В информационном пространстве портала интегрируются следующие типы ресурсов:

- неструктурированные ресурсы – текстовое представление данных;
- слабоструктурированные ресурсы – например, xml-документы;
- структурированные ресурсы – внешние базы данных, к которым есть права доступа.

Описание предметной области портала основывается на системной классификации археологической науки, предложенной Ю.П.Холюшкиным и Е.Д.Гражданниковым в [Холюшкин, Гражданников, 2000: 58 с.] и развиваемой в настоящее время Ю.П.Холюшкиным.

Системная классификация состоит из фрагментов определенной универсальной структуры. Стандартный классификационный фрагмент может быть представлен в виде семантической карты (см. рис. 2), которая служит геометрической моделью фрагмента. Расположение элементов фрагмента определяется позиционной и ранговой координатами, соответствующим критериям первичности – вторичности, антиэнтропийности – энтропийности и общности – частности понятий. Каждое понятие может давать начало фрагменту более низкого яруса, для которого оно служит фоновым понятием, т.е. данный фрагмент охватывает площадку данного понятия, располагаясь под ней.

Таким образом, геометрической моделью классификационной системы может служить трехмерное классификационное пространство, осями которого служат позиционная, ранговая и ярусная координаты.

Внутри отдельного фрагмента существуют горизонтальные и вертикальные смысловые связи, делающие каждый классификационный фрагмент системой в том смысле, что это – целостное образование, содержащее информацию не только в отдельных элементах, но и в их упорядоченных сочетаниях. При описании понятия на более низком ярусе возникают межярусные связи, устанавливающие отношения между элементами разных фрагментов.

Следует заметить, что подробная детализация разделов науки и понятий, имеющая место в рассмотренной выше классификации, и наличие большого количества типов связей между ними затрудняют использование данной классификации в полном объеме для навигации по информационному пространству портала. Так, например, некоторые популярные тематические разделы археологии являются достаточно специализированными и поэтому расположены в глубине классификационной иерархии, и выход на них требует от пользователя наличия большой профессиональной подготовки. Поэтому, для упрощения навигации по порталу используется заданная традиционным образом онтология, построенная на основе полной системной классификации археологической науки.

Рис. 2. Фрагменты системной классификации, предложенной Ю.П.Холюшкиным и Е.Д.Гражданниковым

При разработке этой онтологии были выделены и представлены следующие аспекты классификации:

основные направления археологии и этнографии – классификация по теоретическим разделам научной дисциплины и объектам исследования;

классификация по временному признаку;

классификация по географическому признаку;

археологическая методология или научные подходы в археологии – классификация археологии по применяемым методам исследования.

1.3. Архитектура портала знаний

При разработке архитектуры портала учитывались такие требования со стороны пользователей портала, как его полнота с профессиональной точки зрения, удобство и простота использования.

Рассмотрим основные компоненты и модули портала знаний (см. рис. 3).

База знаний объединяет тезаурус и онтологию портала.

Внутренняя база данных предназначена для хранения всей локальной информации, в частности, описаний ресурсов.

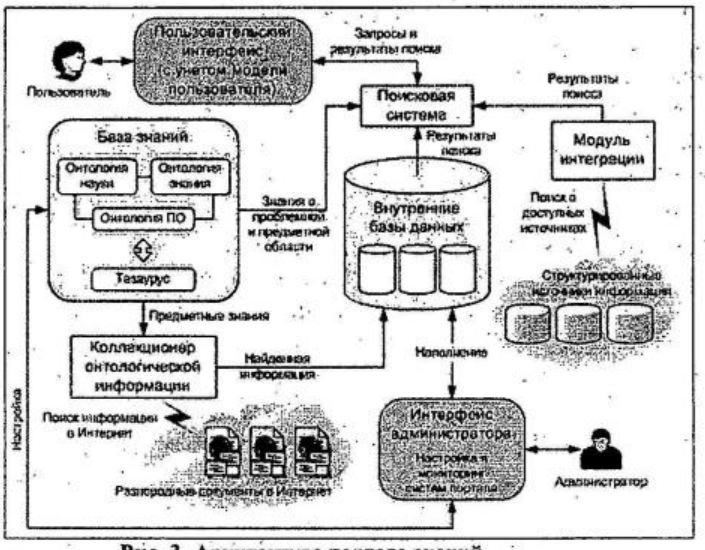


Рис. 3. Архитектура портала знаний.

Модуль интеграции знаний и данных, являющийся компонентом мультиагентной системы содержательного поиска во множестве информационных источников, описанной в работе С.В.Булгакова [Булгаков, 2003], служит для подключения новых информационных ресурсов (источников данных) и поддержки унифицированного доступа к ним. При подключении таких ресурсов устанавливается соответствие между онтологией портала и системой терминов ресурса.

В состав портала также входит подсистема извлечения знаний и данных из сети Интернет – коллекционер онтологической информации о ресурсах. Входящие в ее состав специальные информационные агенты осуществляют поиск и сбор необходимой информации, которая накапливается во внутренней базе данных портала. В состав подсистемы входит также набор конверторов, преобразующих информацию из оригинального представления во внутренний формат представления данных (знаний). Таким образом, происходит автоматическое пополнение портала.

Пользовательский интерфейс предоставляет удаленный содержательный доступ и навигацию по внутренней базе данных и базе знаний портала, а также по информационным ресурсам, проиндексированным в процессе его функционирования.

Для настройки портала на конкретного пользователя или группу пользователей в его состав включена ***информационная модель пользователя***. В частности, модель пользователя содержит его тематические предпочтения, список дополнительно подключаемых/отключаемых ресурсов, способ визуализации страниц и др. Заметим, что модель пользователя уточняется и расширяется при каждом входе пользователя в портал, благодаря чему она всегда отражает его актуальный "информационный портрет".

Интерфейс администратора служит для настройки портала, пополнения и модификации базы данных и онтологий.

Подсистема поиска информации предоставляет пользователю возможность задания запроса не только по ключевым словам, но и в терминах предметной области.

Основными элементами поискового запроса, заданного в терминах ПО, являются:

– ***понятия***, являющиеся элементами онтологии.

– ***ограничения***, которым должны удовлетворять найденные данные. Ограничения могут быть заданы в виде поискового шаблона определенного вида и/или логическими выражениями над значениями атрибутов понятий ПО.

Сформулированный таким образом запрос представляется как фрагмент онтологии с дополнительными ограничениями. Этот запрос преобразуется в один или несколько запросов к внутренней базе данных портала и/или к подключенным к порталу внешним структурированным источникам данных (СИД).

Для обеспечения поиска в СИД во внутренней базе данных в унифицированном виде хранятся описания схем данных внешних источников. Связывание схемы данных СИД с онтологией портала выполняется экспертом-настройщиком.

Сформулированный таким образом запрос представляется как фрагмент онтологии с дополнительными ограничениями. Этот запрос преобразуется в один или несколько запросов к внутренней базе данных портала и/или к подключенным к порталу внешним структурированным источникам данных (СИД).

Для обеспечения поиска в СИД во внутренней базе данных в унифицированном виде хранятся описания схем данных внешних источников. Связывание схемы данных СИД с онтологией портала выполняется экспертом-настройщиком.

Поиск в неструктурированных источниках данных основан на использовании содержательных индексов, хранящихся во внутренней базе данных. Индекс источника данных строится на основе онтологии либо автоматически (коллекционером онтологической информации о ресурсах), либо вручную (экспертом).

Важным достоинством портала является то, что он обеспечивает доступ не только к собственным информационным ресурсам, но и поддерживает эффективную навигацию по релевантным ресурсам сети Интернет, проиндексированным в процессе его функционирования.

К настоящему времени разработана архитектура портала, онтологии науки и научного знания. Завершается разработка начальной версии онтологии археологии и этнографии и соответствующего ей словаря-тезауруса. Ближайшими задачами проекта являются:

– разработка web-интерфейсов пользователя и администратора системы;
– разработка коллекционера онтологической информации;
– автоматическое индексирование Интернет-ресурсов по археологии и этнографии.

2. Информационная система "Системная археология"

В настоящее время в археологии накопилось большое число классификаций, цели и формы которых различны. Общим для большинства классификационных построений является наличие простых цепочных, диадных или триадных структур. Применение таких классификаций никогда не дает понимания действующих как в прошлом, так и в настоящем процессов познания. Недаром Ф. Плог отметил, что "единственной типологией, которая, в конечном счете, оказалась полезной, является периодическая система элементов Менделеева, поскольку она используется не потому, что физики и химики по договоренности решили ее использовать. Она используется, потому что она работает. Она предложила новое понимание структуры элементов, причин их поведения по отношению друг к другу" [Plog, 1973: 653].

Подобное высказывание свидетельствует о том, что осознание необходимости создания системных классификаций в археологии является назревшей задачей. Основными критериями этих систем должны быть:

– упорядоченность;
– периодичность классификаций;
– структурированность;
– теоретическая обоснованность.

В институте археологии и этнографии СО РАН была создана система, опирающаяся на использование, рассмотренных выше требований к "хорошей" системе [Холюшкин, Гражданников, 2000]. В ее основе лежит построение классификационных фрагментов (интеллектуальных карт) с применением законов диалектики для выявления системных связей между понятиями. Классификационные фрагменты обладают следующими системными свойствами:

– однозначностью;
– координатной картографичностью;
– системной историчностью;
– прогностической силой;
– подфоновой полнотой.

Свойство однозначности обусловлено тем, что отдельные значения многозначных слов занимают разные места на интеллектуальных картах.

Координатная картографичность связана с понятийной когерентностью фрагмента, т.е. смысловым соответствием в горизонтальных рядах наук.

Системная историчность проявляется в том, что горизонтальные ряды разделов археологической науки повторяют историю археологии и этапы археологического исследования.

Прогностическая сила интеллектуальной карты вытекает из всеобщего периодического закона, из которого для нас важны две серии прогнозов (ожидания результатов проявления некоторой закономерности):

– прогноз на основе феномена дупликации научных дисциплин;

– прогнозы на основе прогностической линии, которая делит классификационный фрагмент (интеллектуальную карту) на левую (базисную) и на правую (прогнозную) части. Например, такой прогноз для разделов археологии показывает, что ведущую роль на протяжении ближайших десятилетий будет играть мировая археология на базе технологической и реконструктивной археологии.

Подфоновая полнота заключается в том, что каждая карта содержит набор разделов данной области науки, полностью охватывающих ее.

Однако построенная классификация археологической науки носила описательный характер и не имела программной реализации.

В связи с этим была поставлена задача создания программного средства для представления классификационных фрагментов в Интернет. Конкретным решением этой задачи стала реализация проекта создания информационной системы "Системная классификация археологической науки".

Веб-интерфейс информационной системы используется для просмотра и редактирования системной классификации археологической науки. Формально система представлена в виде набора таблиц базы данных:

– понятия и их описания;

– фрагменты с указанием названий и родительского фрагмента;

– структура фрагментов.

Для работы с этими таблицами разработан набор php-скриптов, выдающих соответствующую информацию браузеру пользователя.

Вверху и внизу на каждой страничке информационной системы приведены общие ссылки: [*Главная*] (на основную страницу, содержащую вводную информацию), [*Фрагменты*], [*Понятия*], [*Граф*], [*Вход*] (на страничку авторизованного доступа по имени пользователя и паролю для возможности дальнейшего редактирования системной классификации), [*Поиск*]. В случае авторизованного доступа в систему к этим ссылкам добавляются следующие: [*Создать фрагмент*], [*Создать понятие*], [*Редактировать граф*], [*Выход*] (выход из режима авторизованного доступа; отображается вместо ссылки [*Вход*]) (рис. 4).

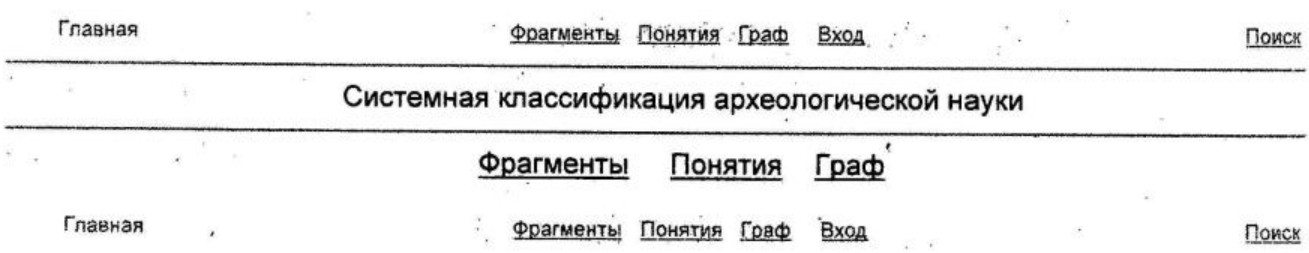


Рис. 4. Общие ссылки.

Ссылка на фрагменты приводит на упорядоченный по алфавиту список фрагментов классификации, который может служить отправной точкой в поиске конкретного фрагмента (рис. 5).

Ссылка с названия фрагмента приводит к страничке его просмотра. Если в базе данных существует понятие с таким именем (что равносильно информации о том, что у данного фрагмента есть родитель), то также будет приведена ссылка [*Понятие*], ведущая к его просмотру. В случае авторизованного доступа дополнительно отображаются ссылки [*Редактировать*], [*Структура*], [*Удалить*], [*Создать потомка*], [*Редактировать понятия*], подробнее о которых будет сказано ниже.

Ссылка на упорядоченный по алфавиту список понятий, существующих в базе данных, аналогична списку фрагментов. Здесь ссылка с названия понятия приводит к страничке его просмотра, а если в базе

существует фрагмент, раскрывающий данное понятие, то будет указана ссылка на его просмотр – [*Фрагмент*]. При авторизованном доступе добавляются ссылки [*Редактировать*], [*Удалить*].

Археологическая информатика [Понятие]
Археологические методы [Понятие]
Археология [Понятие]
Биологическая археология [Понятие]
Естественная история [Понятие]
Историография археологической науки [Понятие]
Классификационная археология [Понятие]
Методологические основания археологии [Понятие]
Общая археологическая методология [Понятие]
Общая археология [Понятие]
Основания археологии (общая археология) [Понятие]
Полевая археология [Понятие]
Принципы полевой археологии [Понятие]
Реконструктивная археология [Понятие]
Социально–экономическая археология [Понятие]
Теоретическая археология [Понятие]
Теория развития археологического знания [Понятие]
Типологическая классификация [Понятие]
Философская археология (философские основания археологии) [Понятие]

Рис. 5. Фрагменты.

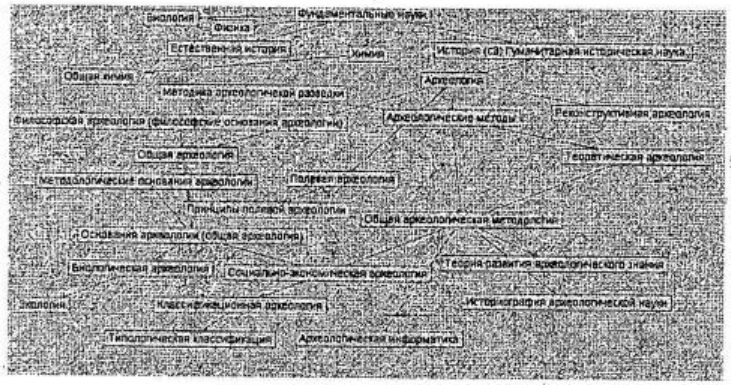


Рис. 6. Граф фрагментов системной классификации.

Поскольку классификация представляет иерархию связанных элементов – классификационных фрагментов, то её можно представить в виде графа, вершинами которого являются схематические прямоугольники, заключающие в себе название фрагмента. На страничке, открывающейся по этой ссылке, графическое представление графа обеспечивается с помощью java-апплета, расположенного в верхней части окна. Выделение вершины в графе сопровождается более детальным отображением фрагмента в нижней части окна, что позволяет в совокупности увидеть не только связи между фрагментами на графе, но и понятия, составляющие каждый фрагмент. Этот вариант просмотра классификационного фрагмента является сокращённой версией полного варианта просмотра, о котором будет сказано в дальнейшем. В случае, когда ни одна вершина в графе не выделена, в нижней части окна отобразится список всех фрагментов (рис. 6).

Страничка с формой поиска фрагментов и понятий позволяет производить поиск по ключевым словам. Поиск осуществляется либо по фрагментам, либо по понятиям. Поиск фрагментов производится по названиям и/или по понятиям, которые их составляют. Поиск понятий осуществляется по названиям и/или по их описаниям. Результатом поиска служит упорядоченный по алфавиту список фрагментов (понятий) (рис. 7).

Отдельный классификационный фрагмент представляется в виде прямоугольника – схемы, позволяющей получить информацию об элементах (понятиях), составляющих фрагмент, и о горизонтальных и вертикальных смысловых связях между ними.

Стандартный классификационный фрагмент содержит 18 элементов. Каждое понятие, входящее в фрагмент, может быть раскрыто в другом классификационном фрагменте, который тем самым становится потомком данного фрагмента.

Таким образом, число потомков одного стандартного фрагмента может достигать 18 (рис. 8).

Страничка просмотра классификационного фрагмента демонстрирует сам фрагмент (рис. 8) и описание входящих в него понятий (рис. 9), а также указание (ссылку) на родительский фрагмент, если он существует, на дочерние фрагменты (ссылки от самих понятий), а также ссылку на страничку с графом, на котором будет выделено местоположение данного фрагмента в общей иерархии.

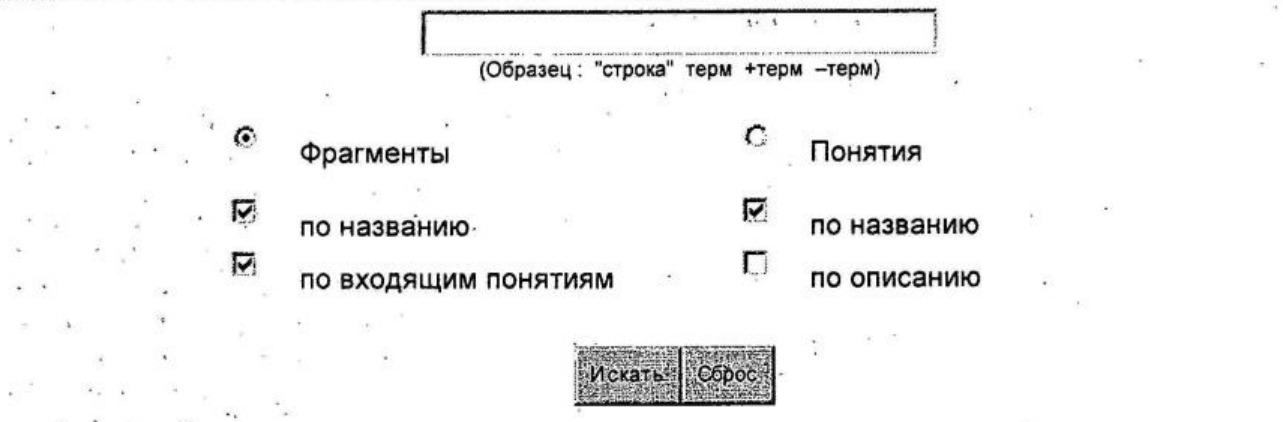


Рис. 7. Ссылка на страничку с формой поиска фрагментов и понятий.

Такой просмотр позволяет перемещаться по иерархии фрагментов на один шаг вверх или вниз и при этом даёт максимум информации, касающейся данного фрагмента. Расположенная в элементарной ячейке ссылка [*Понятие*] позволяет быстро перейти от названия данного понятия к его описанию, расположенного ниже фрагмента на той же страничке.

Общая археологическая методология [Понятие]				
Эмпирическая археология [Понятие] [Создать потомка]			Теоретическая археология [Понятие]	
Описательная археология [Понятие] [Создать потомка]		Компаративная археология [Понятие] [Создать потомка]		Экспериментальная археология [Понятие] [Создать потомка]
Археологические методики [Понятие] [Создать потомка]				
Методика археологичекой разведки [Понятие]	Методика раскопок [Понятие] [Создать потомка]	Методика датировки [Понятие] [Создать потомка]	Методика археологических построений [Понятие] [Создать потомка]	Методика археологических интерпретаций [Понятие] [Создать потомка]

[Предыдущий] [Просмотр] [Понятие] [В списке] [Следующий]

Рис. 8. Стандартный классификационный фрагмент.

Общая археологическая методология	Под общей археологической методологией понимается подраздел археологии, целью которого является обобщение опыта использования всех известных археологических методов.
Эмпирическая археология	Эмпирическая археология – методы получения информации в процессе полевых исследований и непосредственного изучения археологических источников.
Теоретическая археология	Теоретическая археология – методы получения информации путем поиска и анализа закономерностей в эмпирических данных. Согласно Ж.-К. Гардену, теоретическая археология – это "анализ приемов научных рассуждений в археологии" (Гарден, 1983:35).

Рис. 9. Описание входящих в классификационный фрагмент понятий.

Ссылки, находящиеся непосредственно под схематическим изображением, предоставляют различные возможности работы с фрагментом: переход к той страничке списка фрагментов, на которой расположен данный фрагмент (ссылка [*В списке*]), а также к предыдущему или следующему фрагментам списка, минуя переход к самому списку. Ссылка [*Понятие*], относящаяся к фрагменту, позволяет перейти к страничке с описанием понятия, которое дало название данному фрагменту. Если какая-либо

из ссылок заблокирована, это означает, что это действие недоступно. Например, если фрагмент является последним в списке, то ссылка [*Следующий*] будет загашена. В случае авторизованного доступа к системе дополнительно будут отображены ссылки: [*Редактировать*], [*Структура*], [*Удалить*], [*Создать потомка*], [*Редактировать понятия*].

На страничке просмотра понятия приведено его полное описание. Аналогично тому, как это сделано для фрагментов, здесь есть ссылки [*В списке*], [*Предыдущее*], [*Следующее*], [*Фрагмент*], а в случае авторизованного доступа – [*Редактировать*], [*Удалить*].

Права доступа на редактирование предоставляются только авторизованному пользователю. К редактированию относятся создание, редактирование и удаление понятий, создание, редактирование и удаление фрагментов, редактирование структуры фрагмента, редактирование взаимного расположения фрагментов на графе классификации.

Прежде чем создать фрагмент, необходимо внести в базу данных понятия, присутствующие в данном фрагменте. На страничку создания понятия ведёт одна из общих ссылок, расположенных вверху или внизу страницы – [*Создать понятие*]. Пользователю предлагается заполнить название понятия и его описание (может быть пустым). Как название понятия, так и его описание могут быть отредактированы в дальнейшем. Системой отслеживается уникальность названий понятий. Пользователь также может выбрать страничку, которую следует загрузить после того, как понятие будет сохранено в базе данных: можно перейти к списку понятий, к созданию нового понятия, к просмотру или редактированию только что сохранённого понятия (с целью контроля). По умолчанию выбирается пункт перехода к созданию нового понятия – для удобства потокового создания понятий; а при редактировании – пункт перехода к списку понятий. При этом будет отображена та страничка списка, на которой присутствует сохраняемое понятие. При редактировании понятия можно либо выбирать из списка, либо редактировать все понятия, относящиеся к одному фрагменту. Для этого нужно пойти по ссылке [*Редактировать понятия*] для нужного фрагмента (в списке фрагментов или под схематическим прямоугольником). Открывшаяся страничка будет похожа на страничку просмотра фрагмента за тем исключением, что названия и описания всех понятий фрагмента (пустая ячейка не содержит в себе понятия) будут доступны для редактирования.

На страничку создания фрагмента ведут ссылки: [*Создать фрагмент*] (одна из общих ссылок) и [*Создать потомка*]. Запустится java-апплет и в отдельном окошке появится прямоугольник пустого стандартного фрагмента. В первом случае необходимо указать родителя фрагмента и раскрываемое понятие (нажав на кнопку *"Установить родителя"*, в открывшемся окошке выбрать из списка фрагмент, затем выделить ячейку с понятием и нажать кнопку *"Принять"*), либо просто написать название фрагмента, если у него нет родителя. Названия фрагментов уникальны. Во втором случае если ссылка [*Создать потомка*] вела от конкретного понятия фрагмента, то у вновь создаваемого фрагмента уже будет прописан родитель и выбранное понятие; если же ссылка [*Создать потомка*] для фрагмента вела из списка фрагментов или из группы ссылок, расположенных под схематическим прямоугольником, то сразу будет открыто окошко *"Выбор родителя"*, в списке будет выделен данный фрагмент, и пользователю остаётся только указать конкретное понятие (при желании он может выбрать и другого родителя). Заполнение фрагмента происходит следующим образом: необходимо выделить ячейку, нажать на кнопку *"Установить понятие"*, выбрать из списка понятий нужное и нажать кнопку *"Принять"*. Внизу окошка выбора понятия для удобства приводится описание выделенного понятия. Чтобы сохранить фрагмент в базе данных, нужно нажать кнопку *"Принять"*. Для редактирования фрагмента открывается такое же окно. В каждом фрагменте можно заполнить пустые ячейки, либо заменить одно понятие в ячейке на другое, либо вообще очистить ячейку; можно установить нового родителя и раскрываемое понятие, либо очистить родителя.

Если вновь создаваемый фрагмент не является стандартным, можно изменить его структуру, выбрав для фрагмента ссылку [*Структура*] в общем списке или под схематическим изображением. Здесь предоставляется возможность добавить или удалить ряд или ячейку, разделить выбранную ячейку пополам или объединить уже разделённую. Визуальные ряды имеют некоторое логическое преимущество перед ячейками. Если в ряду всего одна ячейка, то удаление её равносильно удалению ряда, поэтому удалить её можно, только выбрав пункт: *"Удалить выбранный ряд"*. Если в ряду все ячейки, кроме одной, разделены пополам, то оставшуюся ячейку разделить нельзя, поскольку тем самым получится уже два ряда. При объединении ячейки, у которой были указаны понятия в верхней и нижней половинках, в итоге останется верхнее понятие. Если же понятие указано только в одной из половинок, то оно останется и в объединённой ячейке. Чтобы действие над ячейкой или рядом вступило в силу, нужно нажать кнопку *"Сохранить"*.

При удалении элементов классификации пользователь должен подтвердить своё намерение удалить данное понятие или фрагмент. На страничке удаления понятие будет приведено вместе с описанием, а фрагмент представлен в сокращённом схематическом виде, позволяющем увидеть составляющие его понятия. При этом составные понятия фрагмента, являющиеся ссылками, будут сигнализировать о том, что данный фрагмент является родителем для других. Это, однако, не препятствует удалению: у всех дочерних фрагментов будет убрана ссылка на родителя, а названия фрагментов останутся теми же. В дальнейшем, при попытке создать потомка для другого фрагмента (Ф1), используя то же самое понятие, которое является названием какого-либо фрагмента без родителя (Ф2), пользователю будет сообщено, что фрагмент с таким именем уже существует. В этом случае нужно либо удалить существующий фрагмент (Ф2), если он не соответствует классификации, либо установить ему родителем фрагмент (Ф1) и выбрав соответствующее понятие.

При авторизованном доступе ссылки [*Предыдущий*] и [*Следующий*] несколько меняют своё поведение: они ведут на страничку аналогичного содержания относительно текущей, но заполненной для предыдущего или следующего фрагмента из списка, соответственно. Например, со странички редактирования понятия по ссылке [*Следующее*] пользователь переходит к редактированию следующего понятия в списке. Также становится ссылкой надпись [*Просмотр*] – на страницах редактирования.

Редактирование графа позволяет создавать, редактировать и удалять фрагменты, а также задавать взаимное расположение фрагментов – вершин в графе, одновременно имея в поле зрения всю иерархию и связи между существующими фрагментами.

3. База данных по фауне палеолита Северной Азии

Создаваемая база данных является подзадачей научной программы сектора археологической теории и информатики Института археологии и этнографии СО РАН по созданию интегрированной базы данных по археологии и этнографии и смежным вопросам гуманитарной науки и образования.

Входными данными служит информация о находках плейстоценовых животных из палеолитических комплексов Северной Азии, содержащая следующие данные: название памятника, слой залегания, степень распространенности фауны и их количество.

Выходными данными является база данных, выводимая в виде множества гипертекстовых документов. В итоговых гипертекстовых документах представлены следующие данные:

LayerName – название слоя
SiteName – название памятника
EngName – название по латыни
RuName – название русское
Persons – количество особей
Bones – количество найденных костей
Presence – присутствие

Предметом проектирования стала логическая структура базы данных и средства просмотра. Для проектирования структуры базы данных использовалось специализированное программное средство ERWin Data Modeler v.4.1 для проектирования реляционных баз данных. В результате проектирования структуры базы данных (рис. 10).

Для создания базы данных использован язык SQL(XSQL). База данных имеет следующие таблицы и отображения(views) :

ANIMAL – таблица с данными о фауне (см. табл. 1).
ANIMALID – первичный ключ
ANIMALFOUND – таблица с находками. (см. табл. 2).
ANIMALID – внешний ключ, который ссылается на таблицу **ANIMAL**
ANIMALID, LAYERID, SITEID – составной первичный ключ
LAYERID, SITEID – составной внешний ключ на таблицу LAYER

Также для работы с базой, кроме таблиц, используются отображения(**View**). Это требуется для выборки данных из нескольких таблиц. **ANIMALVIEW** – отображение, содержащее информацию о животных (см. табл. 3).

ANIMALFOUNDVIEW – отображение, содержащее информацию о фаунистических находках. (см. табл. 4).

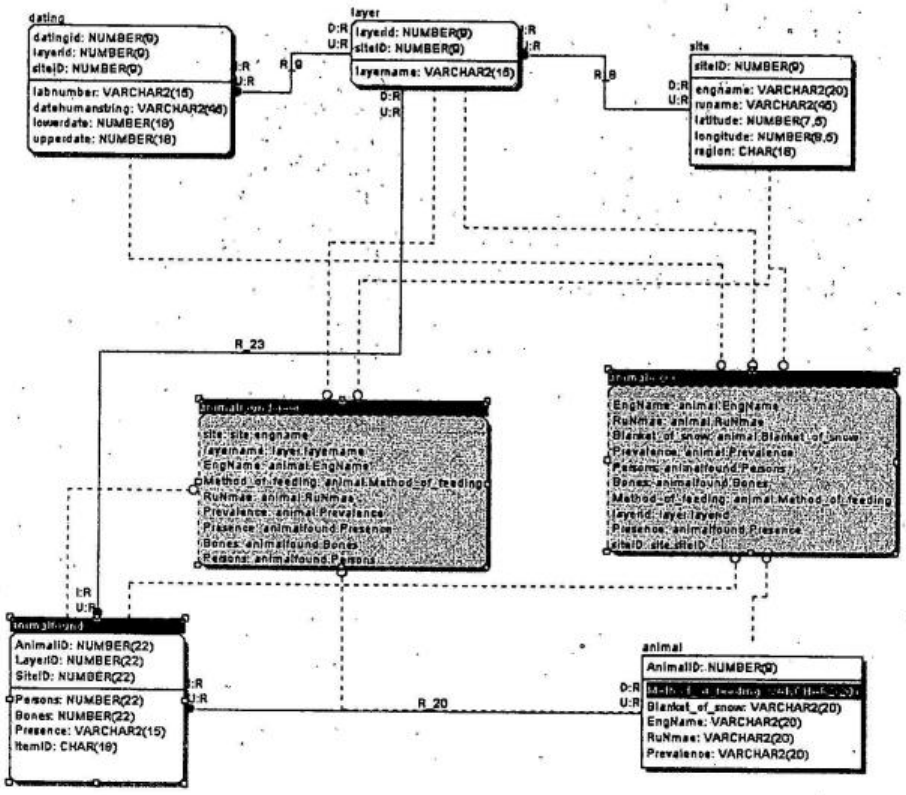


Рис 10. Схема базы данных

Таблица 1. **ANIMAL**

ANIMALID	NUMBER(22) NOT NULL
PREVALENCE	VARCHAR2(22) NULL
METOD_OF_FEEDING	VARCHAR2(22) NULL
BLANKET_OF_SNOW	VARCHAR2(22) NULL
ENGNAME	VARCHAR2(50) NULL
RUNAME	VARCHAR2(50) NULL

Таблица 2. **ANIMALFOUND**

ANIMALID	NUMBER(22) NOT NULL
PERSONS	NUMBER(22) NULL
BONES	NUMBER(22) NULL
LAYERID	NUMBER(22) NOT NULL
SITEID	NUMBER(22) NOT NULL
PRESENCE	VARCHAR2(15) NULL

Таблица 3. Отображение 3. **ANIMALFOUND**

ANIMAL.ENGNAME	VARCHAR2(50) NULL
ANIMAL.RUNAME	VARCHAR2(50) NULL
ANIMAL.PREVALENCE	VARCHAR2(22) NULL
ANIMAL.BLANKET_OF_SNOW	VARCHAR2(22) NULL
ANIMAL.METHOD_OF_FEEDING	VARCHAR2(22) NULL
ANIMALFOUND.PERSONS	NUMBER(22) NULL
ANIMALFOUND.BONES	NUMBER(22) NULL
ANIMALFOUND.LAYERID	NUMBER(22) NOT NULL
ANIMALFOUND.SITEID	NUMBER(22) NOT NULL
ANIMALFOUND.PRESENCE	VARCHAR2(15) NULL

База данных хранится на сервере **Oracle8i**.

Алгоритм функционирования программного средства приведен на рис. 11, схема реализации поиска по памятнику в базе данных – на рис. 12.

Поиск, преобразование и вывод необходимой информации пользователю осуществляется с помощью языков **HTML**, **XML**, **XSL**.

Таблица 4. Отображение 4. **ANIMALFOUND**

LAYER.LAYERNAME	VARCHAR2(22) NULL
SITE.ENGNAME	VARCHAR2(22) NULL
ANIMAL.ENGNAME	VARCHAR2(22) NULL
ANIMAL.RUNAME	VARCHAR2(22) NULL
ANIMAL.METHOD_OF_FEEDING	VARCHAR2(22) NULL
ANIMALFOUND.PERSONS	NUMBER(22) NULL
ANIMALFOUND.BONES	NUMBER(22) NULL
ANIMAL.BLANKET_OF_SNOW	VARCHAR2(22) NULL
ANIMALFOUND.SITEID	NUMBER(22) NOT NULL
ANIMALFOUND.PRESENCE	VARCHAR2(15) NULL

Доступ пользователя к базе данных осуществляется с помощью IE (Internet Explorer). Посредством HTML-интерфейса он может просмотреть БД и выполнять поиск нужной информации.

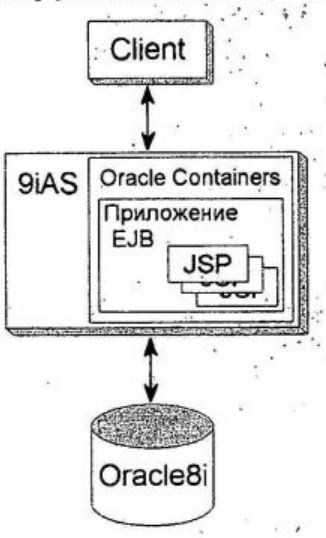


Рис. 11. Алгоритм функционирования программного средства.

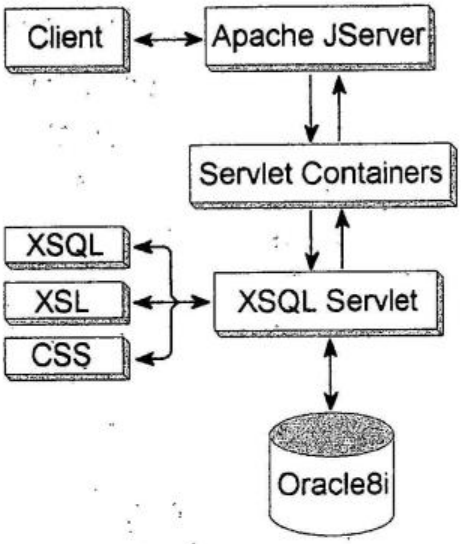


Рис. 12. Схема реализации поиска по памятнику в базе данных.

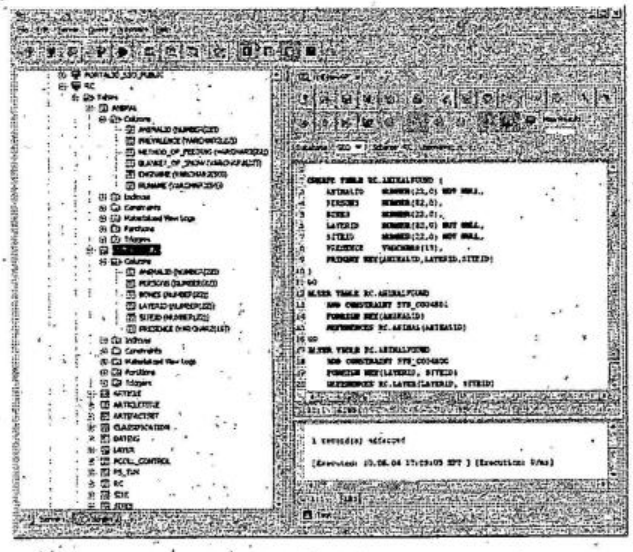


Рис 13. Aqua Data Studio 3.5.

Кроме того, при создании и работе с базой данных использовались программные средства:
Windows 2000 Advanced Server,
ОС Win2000 Professional,
Aqua Data Studio v.3.5,
ER Win Data Modeler 4.1,
Oracle9iDS Containers for J2EE – сервер приложений,

JDeveloper9i – среда разработки,
сервлет и язык XSQL,
стилевые таблицы XSL, CSS.

Наиболее интенсивно использовались средства Aqua Data Studio v.3.5, с помощью которых осуществлялось наполнение, редактирование и сопровождение. (см. рис. 13).

Aqua Data Studio позволяет создавать, управлять и поддерживать реляционные базы данных на серверах Oracle 8i/9i, IBM DB2, Informix Dynamic Server, Sybase Adaptive Server, Sybase Anywhere, Microsoft SQL Server, а также MySQL и PostgreSQL. В версию 3.5 входит графический редактор таблиц, предоставляющий удобную возможность для их редактирования, а также усовершенствованные средства для автоматизации SQL-запросов.

Браузер схемы также позволяет визуально редактировать любой объект схемы с помощью графического отображения проекта. Визуальное редактирование поддерживает таблицы, индексы, процедуры, типы данных и другие объекты схемы. Визуальный редактор также обеспечивает предварительный просмотр SQL всех команд, которые будут выполнены. Проектирование структуры базы данных осуществлялось с помощью ER Win Data Modeler 4.1 (см. рис. 14).

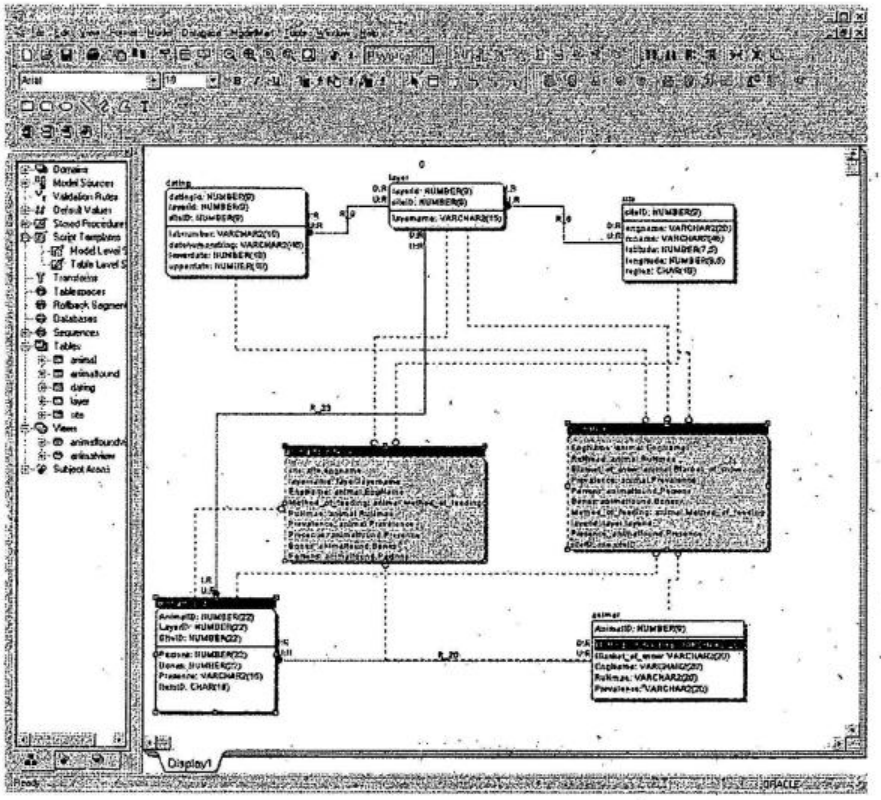


Рис. 14. AllFusion Erwin Data Modeler 4.1

Поскольку ERwin Data Modeler поддерживает работу с БД на физическом уровне, учитывая особенности каждой конкретной СУБД. Разработчики с помощью ERwin Data Modeler могут сначала, используя визуальные средства, описать схему БД, а затем автоматически сгенерировать файлы данных для выбранной реляционной СУБД (прямое проектирование). Автоматически генерируются также триггеры, обеспечивающие ссылочную целостность БД. ERwin Data Modeler поддерживает нотации проектирования данных IDEF1x, IE и Dimensional. На основе модели данных предоставляется возможность создавать отчеты, которые позволяют существенно упростить процесс документирования технического проекта. ERwin поддерживает прямое и обратное проектирование 20 типов баз данных различных производителей, от настольных до реляционных СУБД и специализированных СУБД, предназначенных для создания хранилищ данных.

Поиск в базе данных разрабатывался с помощью программного средства проектирования JDeveloper 9i (см. рис. 15). JDeveloper предоставляет единую интегрированную среду разработки для Java. Для коллективов разработчиков, ориентирующихся на командный метод ведения проектов, в JDeveloper имеется интерфейс к единому хранилищу метаданных, где разработчики могут хранить всю информацию о проекте, об объектах (исходные тексты программ, исполняемые модули, документацию).

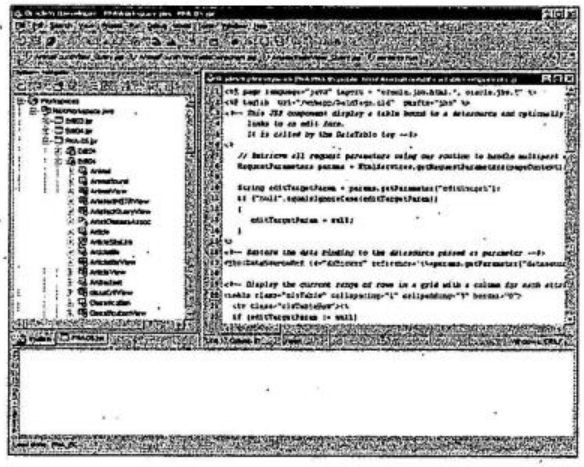


Рис. 15. JDeveloper9i.

В состав Oracle **JDeveloper** включены JavaBeans – компоненты с аналитическими функциями. Например, в Presentation Beans реализованы функции визуализации данных (графики и диаграммы), в Data Query Beans – построения сложных запросов, а в Analytic Beans – аналитических вычислений. Средствами Oracle9i Developer Suite эти компоненты можно интегрировать в любое Java-приложение и легко реализовать в нем сложные аналитические вычисления и запросы.

Используемые программно-технические средства состоят из двух частей: серверная часть и терминальная.

На сервере в директории "pna" находятся исполняемые скрипты. Скрипты написаны на языках:

SQL – просмотр БД при отладке (создание запросов).

XSQL – выборка из базы данных

XML – вспомогательный промежуточный формат

XSL; CSS – язык стилевых таблиц

JSP – поисковые формы запрограммированы на языке JSP(Java Server Page)

Программно реализована база данных, средства поиска и просмотра.

Работа с базой данных осуществляется через веб-интерфейс. Вывод необходимой информации осуществляется двумя способами. Во-первых, выводится информация о всех найденных останков ФАУНЫ на конкретном памятнике. С помощью выпадающего меню пользователь выбирает памятник. В результате исполнения запроса (см. рис. 16, 17) на экран выводится информация о фаунистических находках на этом памятнике и другие археологические данные (датировки, данные по биотопам) (см. рис. 18). Во-вторых, реализован поиск по необходимым данным.

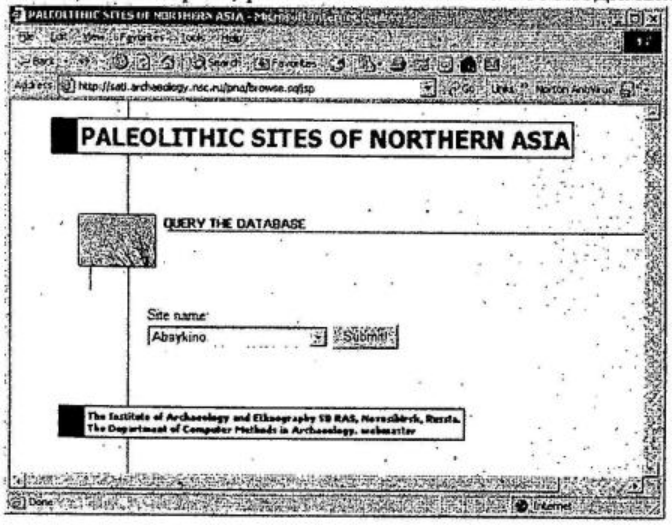


Рис. 16. Запрос данных по находкам по названию памятника.

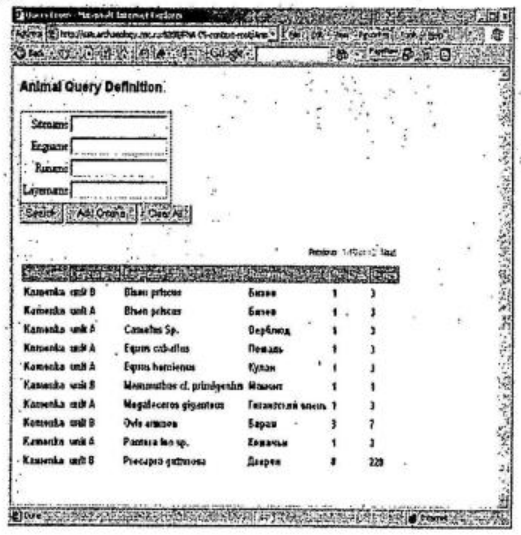


Рис. 17. Интерфейс поиска в базе данных.

В результате проделанной работы спроектирована база данных на языке SQL и реализована на Oracle 8i. Реализован поиск по необходимым параметрам. В ходе проекта изучены новые программные средства

и получен реальный опыт работы с новым программным обеспечением: Aqua Data Studio, AllFusion Erwin Data Modeler, JDeveloper9i.

База данных будет востребована сотрудниками Института Археологии и Этнографии СО РАН, и представителями других смежных наук. Круг пользователей охватывает сотрудников научно-исследовательских учреждений и подразделений гуманитарного профиля Сибирского отделения Российской академии наук и студентов гуманитарных кафедр и отделений колледжей и университетов сибирского региона, а также школьных учителей и старшеклассников, имеющих доступ в Интернет. Она включает широкий круг задач археологии палеолита Северной Азии и опирается на огромный объем накопленных археологических данных, большое количество разных видов информации (история исследования, топографические планы, стратиграфия, фаунистические остатки, данные споропыльцевого анализа, абсолютные датировки, категории артефактов и многое другое).

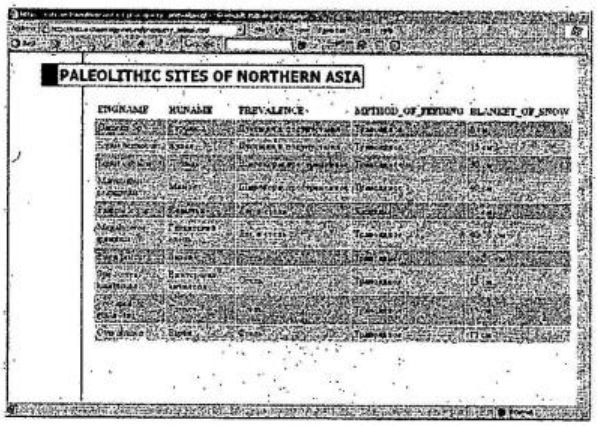


Рис. 18. Способ отображения находок фауны по биотопам

4. База данных по духовной культуре угорских народов Западной Сибири

Целью данного проекта являлась разработка базы данных по духовной культуре обских угров, позволяющей структурировано хранить археолого-этнографические данные, манипулировать и осуществлять быстрый доступ к ним, с удобным и интуитивно понятным пользователю интерфейсом.

База данных содержит информацию о жертвенных покрывалах, культовых местах, атрибутах медвежьего праздника, колчанах, а так же предметах из бронзы и серебра.

При проектировании разработаны алгоритмы манипуляции данными, а также система разграничения доступа для различных групп пользователей.

Спроектирован интерфейс, эффективный и интуитивно понятный пользователю с любой квалификацией.

Программно реализованы:

база данных с этнографическими данными;

система поиска, запросов на модификацию записей об этнографических экспонатах;

база данных пользователей системы;

контроль уровня доступа пользователей и запросы на модификацию записей о пользователях;

удобный и наглядный интерфейс.

На основе применения XML (расширяемый язык разметки) построена гибкая и легко переносимая структура данных. Эту структуру с легкостью можно использовать в различных приложениях (от офисных программ до крупных систем управления базами данных), а также объединять с другими этнографическими системами и базами данных.

Для базы данных создан комплекс программ по манипулированию, обработке и визуализации этнографических данных и данных о пользователях на языке PHP.

Для эффективной визуализации данных, использовались методы, обеспечивающие возможность быстрого изменения дизайна:

HTML (гипертекстовый язык разметки);

CSS (каскадные таблицы стилей).

Для быстрого и эффективного поиска заданной информации использовался язык регулярных выражений (Regular Expressions).

Применение современных кроссплатформенных технологий обеспечило системе стабильное функционирование под управлением различных операционных систем Windows и UNIX семейств.

Разработка выполнялась в рамках проекта по создания геоинформационной системы ИАЭТ СО РАН, предназначенной для исследователей-этнографов, а также для других интересующихся пользователей.

ГИС-системы в этнографии являются мощным комплексным информационным ресурсом, объединяющим разнородные этнографические данные в наиболее естественной для пользователя форме представления. Значение подобных ГИС в гуманитарном образовании и культуре возросло на рубеже XX и XXI веков, когда, с одной стороны, в среде исследователей, преподавателей, студентов и всех интересующихся мифологическими и религиозными представлениями народов Западной Сибири, резко обозначились потребности в интерактивных комплексных формах представления данных (Интернет, сетевых ГИС, гипертекстовых баз данных) и, с другой, все крупные этнографические, археологические и музейные центры, высшие учебные заведения и культурно-просветительские учреждения приобрели и стали развивать собственные web-узлы, с помощью которых пользователи могут получать доступ к информации, размещенной в Интернет.

Однако для этих и других категорий пользователей Интернет реально пока еще мало доступных по WWW баз данных в виде полноценных ГИС-систем не только по этнографии, но и по другим гуманитарным направлениям.

Источниками информации для внесения в базу данных по духовной культуре обских угров и размещения их описания на страницах специализированного сайта являются результаты этнографических исследований и разработок, представленные в монографиях, статьях, обзорах, каталогах и т.д., в том числе малодоступных изданиях XVIII-XIX вв. На картах ГИС эти материалы привязаны к выделенным ареалам форм религиозно-мифологических представлений.

С помощью современных информационных технологий были решены следующие задачи:

– построена удобная структура представления данных об экспонатах (об их местонахождении, дате создания и изготовления, материала и способа изготовления, внешнем виде, составляющих элементах (и их описании) и др.),

– отработаны процедуры хранения информации и удобного доступа к ней;

– подготовлены этнографо-типологические таблицы и классификации;

– разработаны процедуры быстрого поиска;

– созданы процедуры добавления записей о новых экспонатах, изменение и удаление записей при получении новых данных об экспонатах (из публикаций или новых экспедиционных исследований);

– разработаны схемы удобного визуального представления информации и способы ее модификации.

Комплексное решение этих задач воплотилось в создании глобально-доступной базы данных, обладающей рядом необходимых пользователю возможностей:

– доступ к базе данных через сеть Internet. Это означает, что работать с информацией может пользователь из любой точки мира. Но так как база данных общедоступна, доступ к информации, представленной в базе данных, необходимо разграничить – разделить всех пользователей системы на группы и предоставить им определённый уровень привилегий:

– доступ для обычного пользователя, которому можно только читать информацию, не производя никаких действий с ней;

– доступ для этнографа (оператора базы данных), который может читать информацию и производить все манипуляции с данными об экспонатах;

– доступ для администратора базы данных, который может читать информацию, производить любые манипуляции с данными об экспонатах, а также управлять уровнем доступа пользователей.

– удобная структура для облегчения понимания и возможности увидеть всю возможную информацию в компактной и структурированной форме:

– таблицы в базе данных – группы экспонатов;

– доступ к специальным полям, фиксирующим критерии и свойства определенной группы экспонатов для каждой из таблиц.

– доступ к записи в определенном поле таблицы, фиксирующем данные о конкретном экспонате.

Классификации и типологии обеспечивают этнографам быстрый поиск нужных экспонатов.

Методы манипуляции данными (добавление, удаление, изменение, выборки) обеспечат удобный ввод информации об экспонатах, удаление или изменение. Их применение также целесообразно при

исправлении ошибок ввода и устранении неточностей (при получении новых данных об экспонатах), а также для выборки (запросов в БД по определенным критериям – классификационным группировкам).

Эти методы обеспечивают:

– реализацию поиска по всем таблицам в базе данных даст возможность найти экспонаты по различным критериям и вопросам;

– простой интерфейс, доставляющий базе данных удобство эксплуатации для различным категорий пользователей.

Интерфейс включает следующие алгоритмы:

отображение данных,

отображение классификационных группировок,

редактирование данных, включающее добавление, изменение и удаление записей,

поиск;

систему разграничения прав доступа для обычных пользователей, этнографов и администратора.

Для создания базы данных было выполнены следующие разработки:

В начальном варианте системы:

разработана и реализована база данных экспонатов с шестью таблицами и 200 экспонатами в формате XML;

осуществлен дизайн страниц средствами каскадных страниц стилей – CSS, навигация реализовать средствами HTML;

выбрана информация из базы данных и отображена на HTML-страницах при помощи стандарта DOM XML;

реализована система поиска и запросов на языках PHP и RegExp;

реализована классификация и выборки к ней – PHP;

реализованы запросы на добавление, изменения и удаление записей об экспонатах (с формами для ввода, редактирования и удаления информации – HTML) для базы данных – PHP, DOM XML;

разработана и реализована база данных пользователей системы (этнографами и администратором) – XML;

реализована проверка уровня доступа пользователей – HTML, DOM XML, PHP, RegExp;

реализованы запросы на создание и удаление пользователей – HTML, DOM XML, PHP.

Входные данные.

Запросы к главным файлам-сценариям:

1) на отображение нужной таблицы или формы – запрашивается выбранная таблица (покрывала, святилища, маски и медведи, колчаны, бронза, серебро) и выбранный вид действия над ней (показать все экспонаты, показать классификации, найти нужные экспонаты, добавить, изменить, удалить запись);

2) на отображение нужной формы – запрашивается временный идентификатор (пользователя или администратора) и вид действия над пользователем (проверка, создание, удаление);

3) на вызов нужных функций – запрашиваются данные в зависимости от вида функции: данные из форм, из строки при перенаправлении, из констант, из переменных окружения и др.

Запросы к базе данных:

1) отбор экспонатов:

– выбрать все экспонаты для данной таблицы;

– выбрать экспонаты, отвечающие классификационной группировке, для данной таблицы;

– выбрать экспонат или экспонаты, отвечающие введенной строке в заданном поле, для данной таблицы;

– выбрать экспонат или экспонаты, по определенному месту нахождения;

– выбрать экспонат, отвечающий данному уникальному идентификатору, для данной таблицы;

– добавить запись в данную таблицу;

– изменить запись в данной таблице;

– удалить запись в данной таблице.

2) выбор пользователей по выделенному критерию:

– на существование данного пользователя;

– на правильность пароля данного пользователя;

– на добавление нового пользователя;

– на удаление пользователя.

Информация в таблицах:

– жертвенные покрывала: название экспоната, количество квадратов, изображение, материал, цвет, меховая оторочка, размер, номер фотографии, этническая принадлежность, дата изготовления, место – где найден экспонат, музей – где хранится или материалы с упоминанием;

– культовые места – святилища: название культового места, принадлежность; местонахождение; вблизи/вдали от селений; привязанность к природным объектам; жилище духа-покровителя: для амбарчика (количество, размер, материал, опора, стены, фасад, крыша, дверь, лестница, вход, замена амбарчика), для навеса (размер, скат, опора); ритуальная площадка: кострище, стол, деревья с личинами/ножами, жертвенные жерди, изваяния менквов, посуда; фигура духа-покровителя: размер фигуры, внешний вид фигуры; предметы жертвоприношений; оружие; номера фотографий;

– атрибуты медвежьего праздника: название экспоната, описание внешнего вида, номер фотографии, этническая принадлежность, дата изготовления, место – где найден экспонат, музей – где хранится;

– ритуальные колчаны: название экспоната, количество фигур, изображение, материал, цвет, размер, номер фотографии, этническая принадлежность, дата изготовления, место – где найдено, дополнение;

– предметы из бронзы: название экспоната, материал и методика изготовления, размер, номер фотографии, дата изготовления, место – где найдено, описание внешнего вида;

– предметы из серебра: название экспоната, материал, размер, номер фотографии, описание внешнего вида, этническая принадлежность, дата изготовления, место – где найдено.

Выходные данные:

Индексный файл.

Данные, полученные путем запроса:

– таблицы классификацией либо типологией;

– таблицы с отображением экспонатов, соответствующим классификационной группировке;

– таблицы с отображением всех экспонатов, соответствующих определенной группе экспонатов;

– таблицы с отображением всех экспонатов, соответствующих определенной строке запроса: по любому из полей;

– таблицы с отображением всех экспонатов, соответствующих определенной строке запроса: по месту нахождения.

Формы для соответствующих групп экспонатов:

– форма для ввода имени пользователя и пароля;

– форма для создания нового пользователя;

– форма для удаления пользователя;

– формы для создания нового экспоната и изменения существующего;

– формы для выбора одного экспоната;

– формы для удаления экспоната;

– формы для поиска экспоната.

– сообщение об ошибке.

При реализации базы данных использовались следующие возможности языков программирования и других технологий.

1. *Язык XML* (Extensible Markup Language, расширяемый язык разметки) позволяет хранить любую информацию в структурированном виде. Он имеет синтаксис, который позволяет легко писать программы для работы с документами XML, а так же – встроенные средства проверки корректности структур и информации, описанной в документе. Информация в XML заключена в тэгах или их атрибутах. Следовательно, структура документа представляет собой дерево (тэг в тэге). Благодаря этому можно создать гибкую и легко переносимую структуру данных, которую с легкостью можно использовать в различных приложениях: от офисных программ до крупных систем управления базами данных, а также объединять с другими этнографическими системами и базами данных.

2. Для последующей обработки структуры БД и манипуляции с данными (выборка, добавление, удаление, изменение записей) используется *серверный язык сценариев PHP*. PHP является интерпретатором со встроенным блоком трансляции, оптимизирующим ход интерпретации. Таким образом, главной фазой работы PHP является интерпретация внутреннего представления программы и ее исполнение. Вследствие этого, PHP – это язык, который позволяет, с одной стороны, встраивать в код программы "куски" HTML-кода, с другой, встраивать программный код в HTML-страницы. Эти свойства обычно активно используются для продвинутых страниц многих сайтов в Интернет.

3. Для рациональной обработки данных с помощью языка PHP используется стандарт *DOM* – Document Object Model. Он не привязан к какой-то конкретной платформе или языку программирования

и позволяет выполнять все операции по обработке XML-данных (то есть вы можете не только читать их, но и модифицировать содержимое XML-документа, вставляя туда новые тэги, удаляя и изменяя их). DOM предоставляет пользователю простой способ доступа к информации как к дереву объектов. Обход такого дерева позволяет надежно и быстро извлечь всю необходимую информацию.

4. Благодаря использованию возможностей гипертекстового языка разметки HTML вся информация представлена для пользователя в привычной табличной форме с удобной навигацией. Использовались такие элементы языка, которые полностью поддерживаются и одинаково интерпретируются всеми популярными браузерами: Internet Explorer, Netscape Navigator, Mozilla, Opera. В результате вне зависимости от установленного у пользователя программного обеспечения интерфейс системы всегда будет удобен и функционален.

5. Для создания интерфейса были использованы так же и каскадные таблицы стилей (CSS). С их помощью удалось максимально гибко описать представление интерфейса и отделить структуру интерфейса от его представления.

Это позволило:

– минимизировать количество визуальных параметров оформления в HTML-страниц;

– модифицировать структуру программы, не заботясь о том, что может измениться внешнее представление страниц;

– изменить интерфейс системы с минимальными затратами времени, не затрагивая кода программ;

– легко задавать различные формы представления данных: допустим, другое оформление для принтера, отличное от представления в браузере клиента и др.

6. Для безопасного хранения паролей пользователей использован необратимый алгоритм построения цифрового дайджеста MD5.

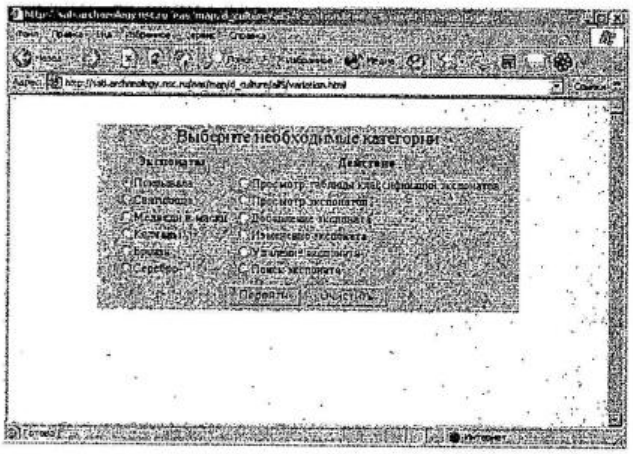


Рис. 19. Главная страница

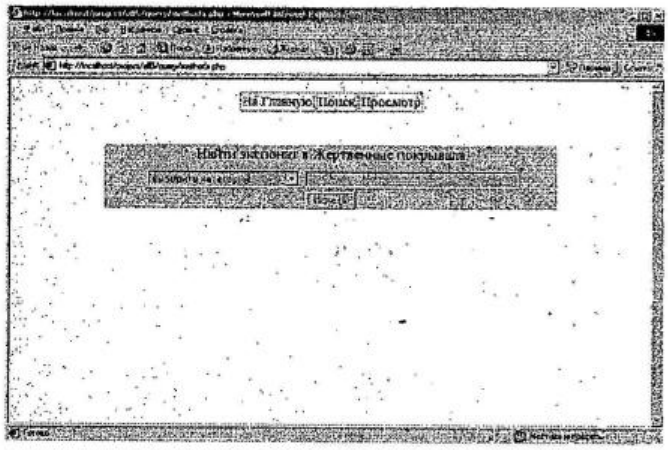


Рис. 20. Поиск экспоната.

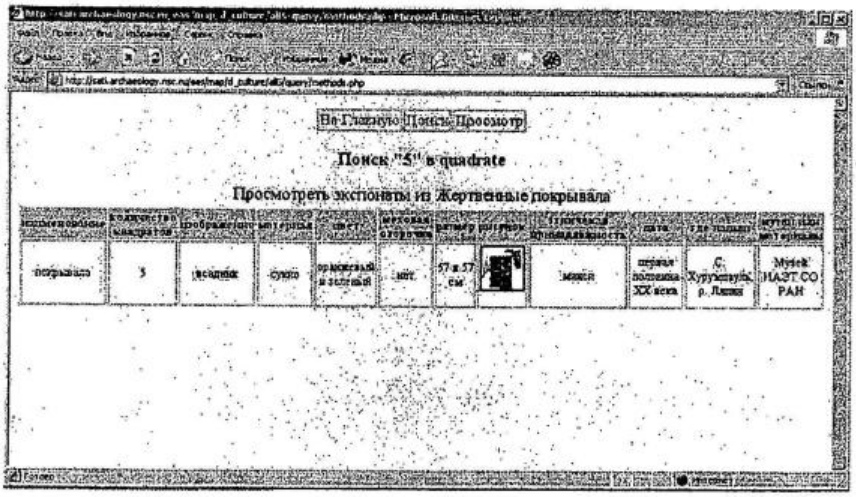


Рис. 21. Результаты поиска.

7. Задействованы возможности протокола HTTP (гипертекстового протокола передачи данных). Сценарии создают заголовки протокола для эффективного взаимодействия с браузером пользователя, например автоматическое перенаправление пользователя на другой сценарий системы.

8. При обработке текстовых данных использовался язык регулярных выражений (RegExp). Благодаря ему реализована гибкая и простая для сопровождения система поиска.

Работа пользователя с системой начинается с главной станицы variation.html, где пользователю выводятся категории экспонатов и возможные действия с ними.

На рис. 19 изображена главная страница системы, она содержит список категорий экспонатов и список возможных действий пользователя. Пользователь может выбрать категорию экспонатов, например, "покрывала" и действие над ними, например, "Просмотр таблицы классификаций". После этого он должен нажать кнопку "Перейти" для продолжения работы с системой или кнопку "Очистить" чтобы изменить своё решение и выбрать другие категории или действия.

Если пользователь не выбрал категорию экспонатов или действие и нажал кнопку "Перейти" выводится сообщение об ошибке, так как система не может начать свою работу.

Пользователь выбрал интересующую его категорию и действие "Поиск экспонатов". Для него открывается страница с формой поиска экспонатов и навигационным меню (рис. 20). Форма поиска состоит из двух полей: определенная характеристика экспоната, по которой необходимо произвести поиск, и поле для ввода пользователем той информации, которую необходимо найти.

Навигационное меню содержит ссылки на "Главную страницу" (см. рис. 19), ссылку на действие "Просмотр экспонатов" (см. рис. 22) в данной категории экспонатов.

На действие "Поиск экспонатов" так же можно попасть из любого другого раздела, нажав кнопку "Поиск" в навигационном меню.

На рис. 21 изображена страница результатов поиска с данными о найденных экспонатах.

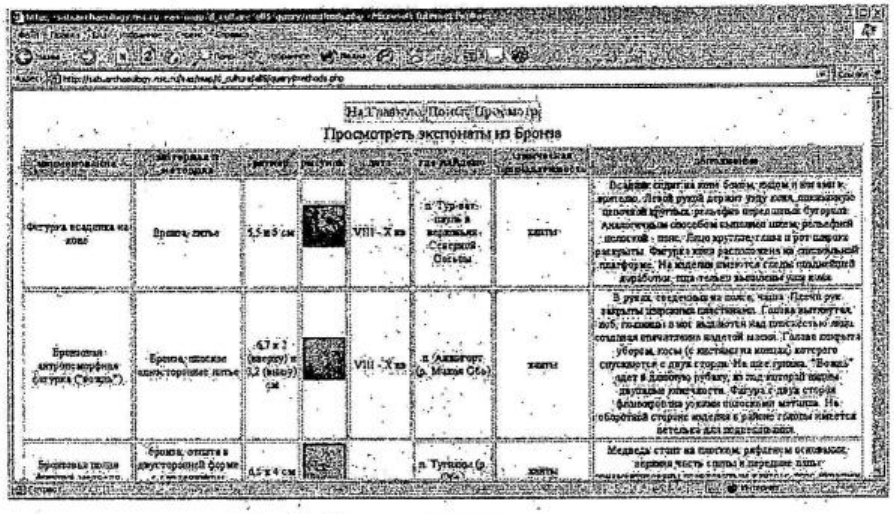


Рис. 22. Просмотр экспонатов.

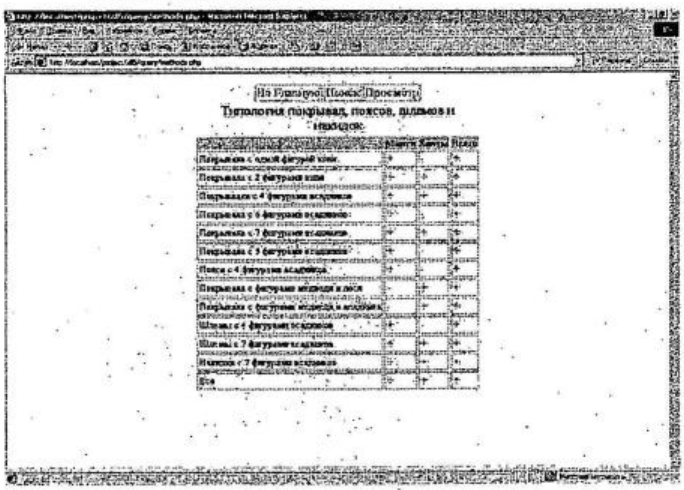


Рис. 23. Просмотр таблицы классификаций.

Пользователь выбрал интересующую его категорию и действие "Просмотр экспонатов". Для него открывается страница, изображённая на рис. 22. Она содержит навигационное меню, заголовок и таблицу с описанием всех экспонатов выбранной категории экспонатов.

Пользователь выбрал интересующую его категорию и действие "Просмотр таблицы классификаций". Для него создается страница (рис. 23) с информацией о типологии или классификации экспонатов данной категории со ссылками на экспонаты, соответствующей классификационной группировки, а так же заголовок и навигационное меню.

Результат обращения по ссылке изображен на рис. 24.

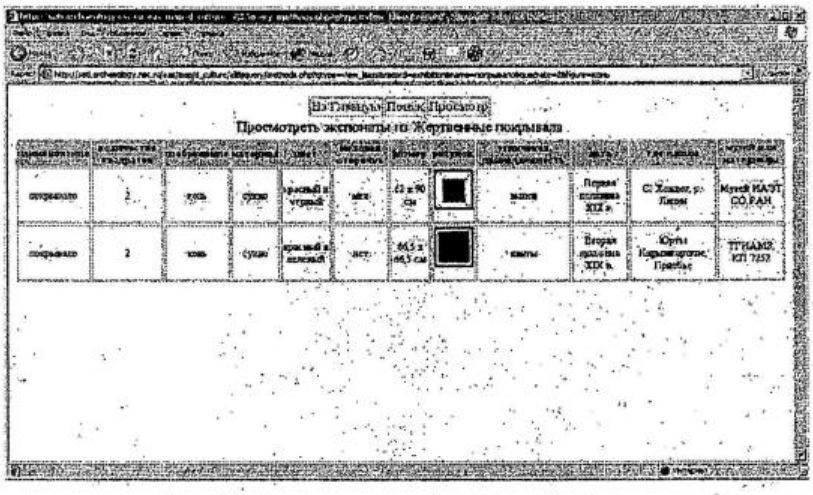


Рис. 24. Просмотр классификационной группировки.

При выборе действий "Добавление", "Изменение", "Удаление" экспоната, выводится страница авторизации пользователя и меню управления пользователями (только для администраторов системы), изображенная на рис. 25.

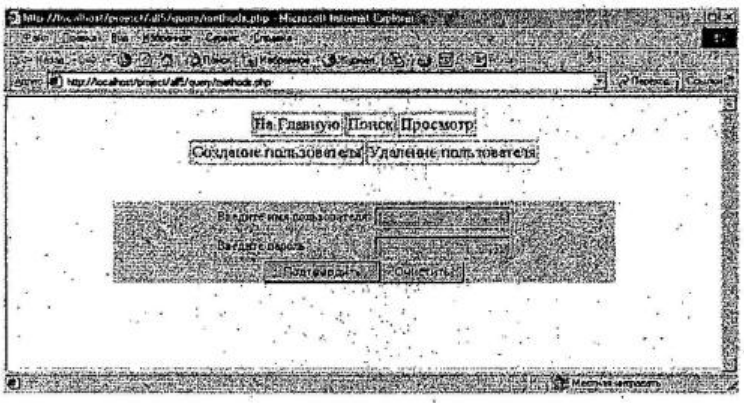


Рис. 25. Авторизация пользователя.

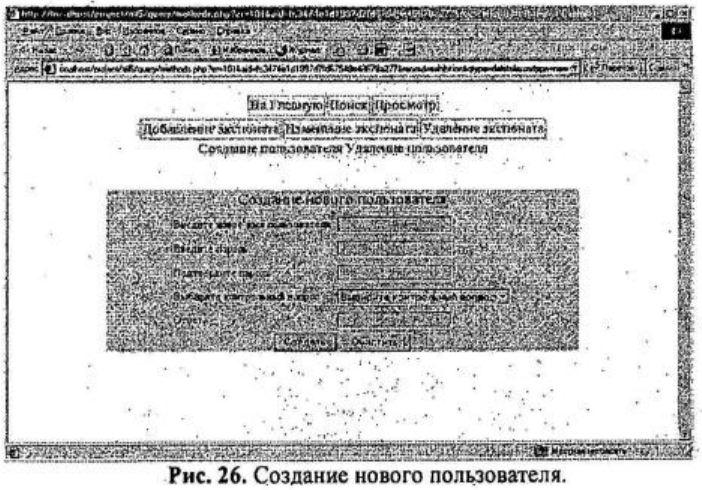


Рис. 26. Создание нового пользователя.

При выборе действия "Создание пользователя" (здесь необходим уровень привилегий администратора) выводится страница с формой, навигационное меню и административная панель управления (рис. 26). Форма включает в себя все поля, которые необходимо заполнить, чтобы создать нового пользователя в системе: поле имени пользователя, поле пароля, поле подтверждения пароля для контроля ошибок, поле контрольного вопроса и поле ответа – для восстановления пароля в случае, если пользователь системы его забыл.

При выборе действия "Удаление пользователя" (здесь тоже требуется уровень привилегий администратора) выводится страница с формой, навигационное меню и административная панель управления (рис. 27). Форма включает в себя поле, в котором необходимо выбрать имя удаляемого пользователя.

Пользователь выбрал интересующую его категорию и действие "Добавление экспоната ". Для него открывается страница, изображенная на рис. 28. Она представляет собой форму для заполнения информации о новых экспонатах, с полями, описывающими все параметры экспоната, навигационное и административное меню.

Пользователь выбрал интересующую его категорию и действие "Удаление экспоната". Для него открывается страница, изображенная на рис. 29. Она содержит форму для выбора нужного экспоната (с записями обо всех экспонатах), навигационное и административное меню.

После того как пользователь выбрал нужный экспонат и подтвердил это нажатием кнопки "Выбрать экспонат", открывается страница с формой, изображенной на рис. 30, с полями изменяемой информации, описывающими все параметры экспоната, навигационное и административное меню.

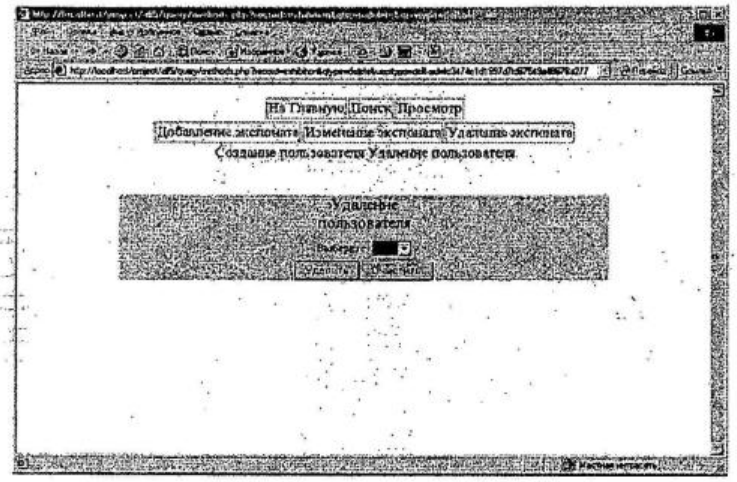


Рис. 27. Удаление пользователя

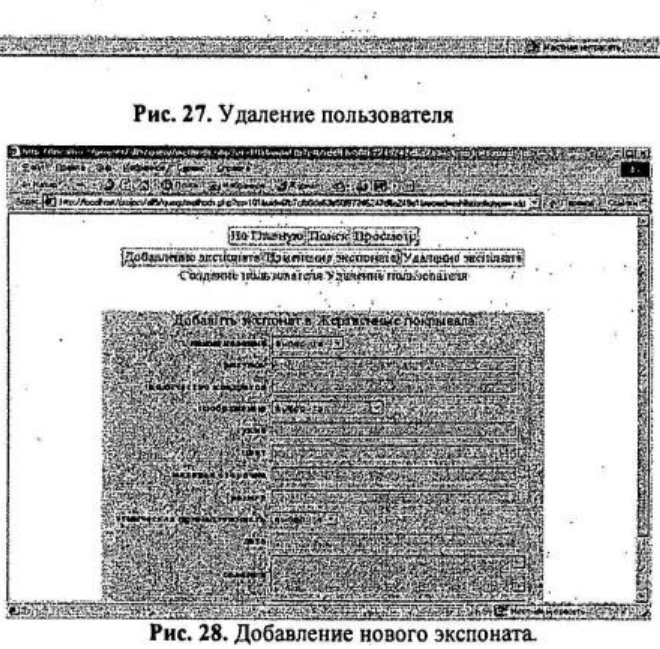


Рис. 28. Добавление нового экспоната.

На основе описанной технологии предусматривается разработка подобных баз данных по материальной и духовной культуре коренных народов Сибири и Дальнего Востока с ориентацией применения в этнографических ГИС.

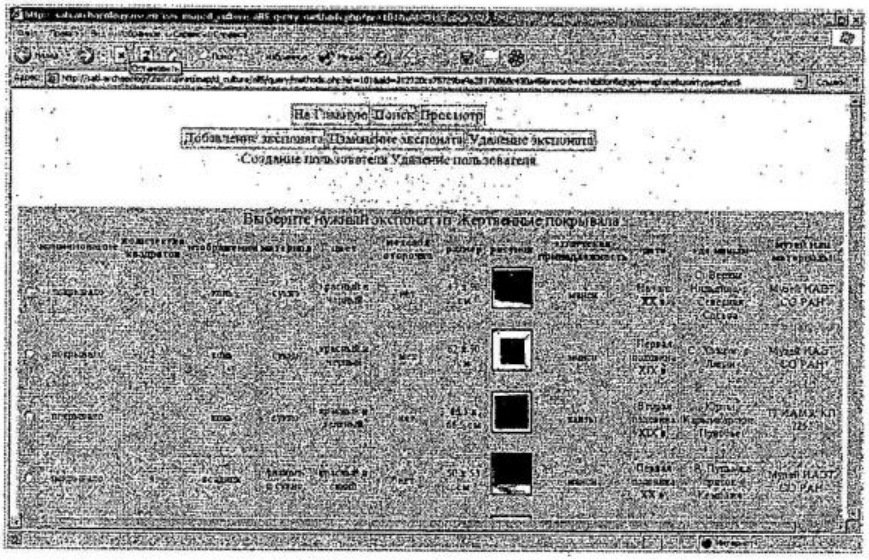


Рис. 29. Выбор экспоната

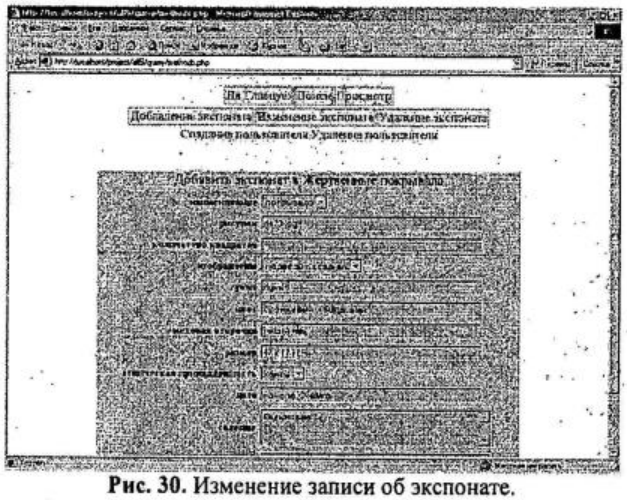


Рис. 30. Изменение записи об экспонате.

5. Разработка Web-интерфейса локальной базы данных электронного каталога библиотеки

В наш век высоких технологий и всеобщей компьютеризации глобальная сеть Интернет становится неотъемлемой частью быта и условий жизни большинства людей. В глобальной сети размещены, постоянно пополняются и расширяются самые разнообразные, необходимые для них (людей) информационные ресурсы: музыка, фильмы, тексты, в том числе разного рода литература, получить которую в виде твердых копий за пределами Интернет весьма затруднительно.

Для этой цели в Интернет имеются или разрабатываются сервисы для поиска текстовых данных в электронных библиотеках, где читатели могут найти нужные книги. Эти сервисы избавляют их от длительных поездок: в другой конец города, если в городе имеются обычные библиотеки, или в другие города (если не в другую страну), иначе.

Существенно облегчают поиски текстовых документов электронные каталоги библиотек, позволяющие осуществлять многоаспектный поиск данных. От того, насколько легок и приятен процесс поиска для пользователя, зависит, найдет ли он интересующую его книгу (или ее какие-либо фрагменты) или нет.

Перед авторским коллективом была поставлена задача по созданию электронного каталога библиотеки Института археологии и этнографии СО РАН.

Для разработок в проекте использовался локальный вариант электронного каталога научной библиотеки института археологии и этнографии СО РАН. Указанный прототип электронного каталога, созданный на базе сервера протокола z39.50 и Web-сервера Apache, ориентирован на двухэтапную схему

размещения базы данных электронного каталога в Интернет. По этой схеме фонды каталога сначала перемещаются из локальной базы в базу данных в Интернет, а затем эти данные конвертируются в формат RUSMARC, обеспечивающий доступ к ним по протоколам z39.50 и HTTP.

В задачу проекта входила разработка новой схемы доступа к ресурсам локального электронного каталога научной библиотеки ИАЭТ СО РАН, обеспечивающей возможность выхода на данные каталога непосредственно из Интернет, минуя промежуточные стадии их перемещения и конвертации.

Важной частью разработки является модернизация структуры прототипа электронного каталога, исходя из того, что разрабатываемый каталог в отличие от прототипа ориентирован на включение в интегрированные информационные ресурсы сектора археологической теории и информатики (САТИ), обособленные от информационных ресурсов корпоративной региональной библиотечной системы (в которую прототип включен непосредственно как неотъемлемое звено). Включение прототипа в корпоративные формы ограничивает номенклатуру и структуру записей в базе данных электронного каталога рамками договорных обязательств организаций – членов корпорации.

Ориентация разработки электронного каталога на автономное функционирование выдвигает другую важную задачу – разработку интерфейса каталога с доступом из Интернет. Предполагается, что сам каталог преимущественно будет наполняться и поддерживаться в рамках локальной сети Института. Поэтому ставится задача, сохраняя локальную форму поддержки каталога, обеспечить удобный удаленный доступ к его ресурсам.

Весь объем перечисленных задач реализован в ходе работ по проекту. Ядром проекта является разработка методов доступа к элементам локализованной копии прототипа.

Целью проекта ставилось обеспечение удобного удаленного доступа к ресурсам локальной актуальной копии электронного каталога научной библиотеки Института археологии и этнографии СО РАН. Эта цель включает задачу обеспечить возможность удаленного доступа с помощью WEB интерфейса, с одной стороны, и с помощью клиента протокола z39.50, с другой.

Интерфейс страницы, задающей параметры поиска, выглядит вполне понятным и стандартным. На этой странице можно задать те данные, которые требуется найти, а так же поля для поиска. Таким же образом можно задать формат вывода результатов. Результаты могут быть представлены как в виде таблицы, так и в виде библиографических карточек. Так же есть возможность представить результаты поиска в виде множества отобранных записей базы данных. Результаты поиска могут сортироваться по одному из полей. С помощью шаблонов есть возможность задавать формат библиографических карточек. Эта возможность позволяет задавать практически любой формат вывода. В Интернет доступ к описанному интерфейсу можно получить из любой точки мира.

На рис. 31 представлена главная страница, на которой задаются критерии поиска (рис. 32), а также формат вывода (в каких полях искать и что нужно найти) – рис. 34-37.

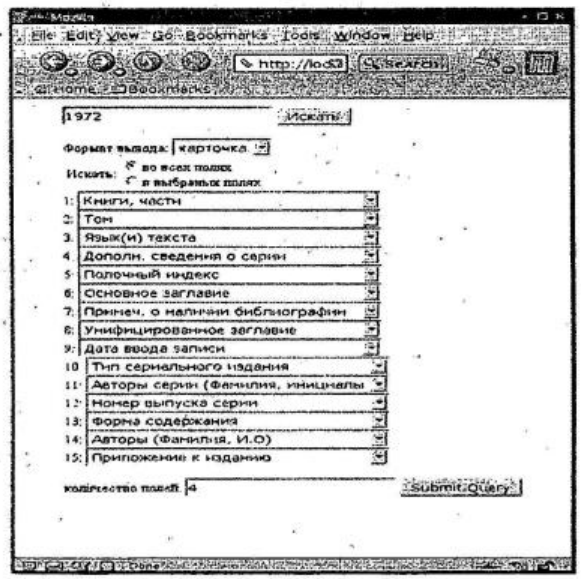


Рис. 31. Задание критериев поиска в базе данных.

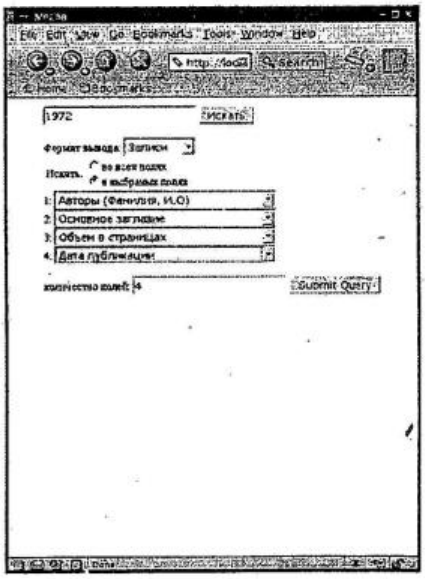


Рис. 32. Вариант задания критериев поиска и полей в которых нужно осуществлять поиск.

Z39.50 определен стандартами ANSI Z39.50-1995, ISO/FDIS 23950. Согласно этим стандартам Z39.50 представляет собой протокол прикладного уровня в рамках семиуровневой эталонной модели взаимодействия открытых систем, разработанной Международной Организацией Стандартов (ISO) и поэтому может быть реализован в различных типах сетей (в сетях TCP/IP, IPX/SPX, OSI и др.), независимо от реализации транспортного уровня. Назначение этого протокола – предоставить компьютеру, работающему в режиме "клиент", возможности поиска и извлечения информации из другого компьютера, работающего как информационный сервер.

Стандарт определяет для компьютеров-клиентов единую процедуру запроса информационных ресурсов – серверов, поддерживающих библиотечные каталоги.

Не вдаваясь пока в детали работы протокола, можно сказать, что стандарт Z39.50 определяет такие правила взаимодействия компьютеров, которые позволяют унифицировать доступ к различным базам данных. Иными словами, пользователь, использующий всего лишь одно приложение на компьютере-клиенте, может производить поиск информации в удаленных распределенных базах данных, имеющих самую разную структуру и форматы представления информации.

Изначально протокол Z39.50 предназначался для обработки библиографической информации. Однако сейчас протокол достаточно развит, чтобы поддерживать различные данные – финансовую, химическую, техническую информацию, полные тексты и изображения.

До появления Z39.50 основным протоколом доступа к распределенным хранилищам информации был протокол HTTP. Однако HTTP – протокол общего назначения и не имеет практически никаких специализированных возможностей, позволяющих, например, унифицировать доступ к разнородной информации, или создавать поисковые запросы к БД.

Разумеется, эти проблемы решаемы и на базе протокола HTTP. Однако для этого приходится применять различные дополнительные средства, языки программирования, библиотеки функций и так далее. Это позволяет разработчику, использующему форматы HTTP, находить свои пути решения подобных проблем, создавая систему доступными ему средствами, используя собственные механизмы, технологии и модели. А в результате каждая поисковая система может иметь свои собственные структуру и форматы хранения данных, не обязательно согласующиеся с используемыми стандартами.

Основная идея представления информации при работе с протоколом Z39.50 лежит в абстрагировании от конкретной структуры какой-либо базы данных. Для этого в стандарте описаны некая абстрактная модель БД. Эта модель включает в себя полный набор элементов, необходимых для доступа и обработки информации, хранимой в БД. Абстрактная модель описывает в виде отдельных элементов не только, например, возможные поисковые поля или форматы выдачи информации, но и все выполняемые сервером операции.

Таким образом, абстрактная модель БД, представленная протоколом, отображается на конкретную модель существующей базы данных. Задача разработчика системы состоит в том, чтобы правильно отобразить абстрактную модель данных протокола на существующую структуру БД и сопоставить соответствующие элементы.

Основная часть проекта реализована на языке программирования Perl. Язык программирования Perl сейчас очень популярен в мире. Язык Perl – это интерпретируемый язык. Эта особенность позволяет использовать программы, написанные на Perl, в большом количестве операционных систем, не изменяя программный код. Большие функциональные возможности языка Perl позволяют легко создавать интерфейсы удаленного доступа к базам данных. Популярность этого языка делает возможным легко изменять проект другими людьми, которые будут впоследствии его дополнять. Так же популярность языка делает возможным использовать в проекте большое количество сторонних модулей, написанных другими людьми в рамках других проектов. Язык Perl в большей степени ориентирован на WEB программирование. На Perl реализована часть, отвечающая за WEB интерфейс к базе данных. В моем проекте программа использует динамическую библиотеку OpenIsis. Эта библиотека позволяет получить доступ к данным базы данных в формате CDS/ISIS и обрабатывать ее. Полученный результат переводится в формат, который будет более понятен пользователю (табличная или иная форма). Программа, написанная на языке Perl, может являться CGI скриптом.

Структура программного комплекса и схема взаимодействия его компонентов приведена на рис. 33.

На сервере в директории "cgi-bin" находятся специальные исполняемые скрипты-файлы, написанные Index Data APS в 1995-2004 годах. Эти скрипты были модифицированы для работы с копией прототипа базы данных электронного каталога. Исходные скрипты размещены на сайте разработчика (http://www.indexdata.dk/). Эти скрипты отвечают за соединение с сервером протокола z39.50, поиск в базе данных по заданному шаблону и формирование HTML-кода.

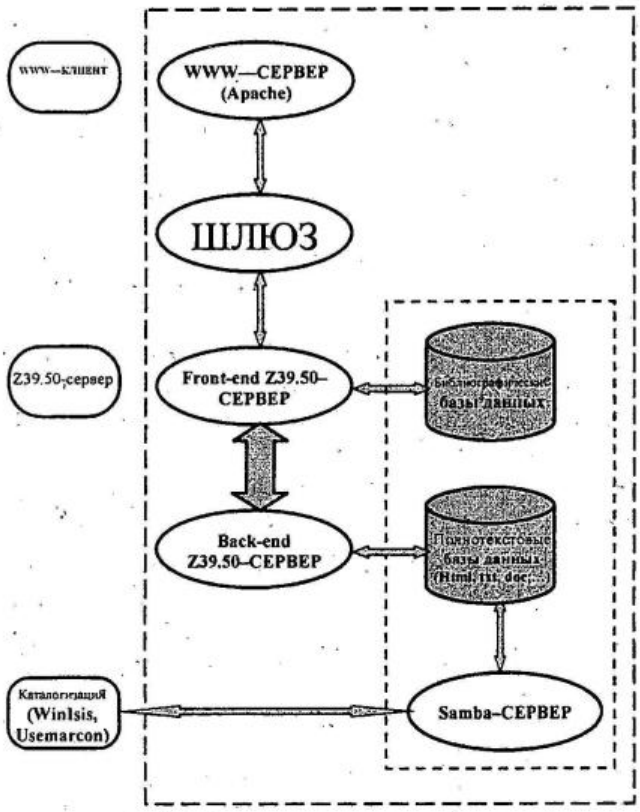


Рис. 33. Структура программного комплекса и схема взаимодействия его компонентов.

Скрипты написаны на языке программирования C.

Когда пользователь подает запрос на поиск, скрипт соединяется с сервером по протоколу z39.50, пересылает ему запрос поиска и выводит результаты поиска в доступной форме.

Так же в этой директории находятся скрипты, написанные, на языке Perl, которые обращаются к базе данных напрямую, осуществляют поиск в базе данных по заданному шаблону и формирование HTML кода. Результаты представляются в нескольких видах (табличный, в виде библиографических карточек, в виде записей и в виде XML).

Рис. 34. Пример запроса и результата вывода записи в формате XML. На рисунке жирным шрифтом выделены поисковые термины.

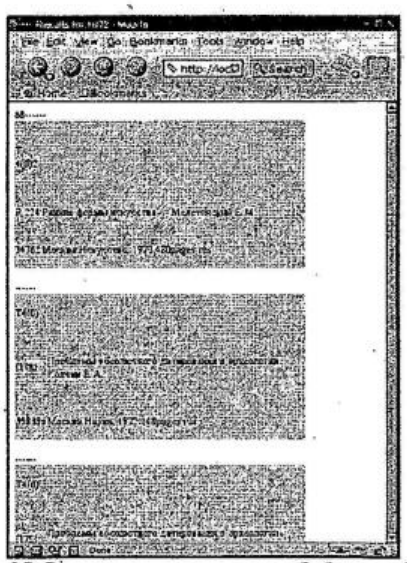


Рис. 35. Результат вывода в виде библиографической карточки с выбранными полями.

Прямой доступ к базе данных осуществляется средствами библиотеки OpenISIS, которая позволяет напрямую работать с базами данных в формате CDS/ISIS.

Так же существует возможность доступа к этой базе данных по протоколу smb. Что позволяет получить к ней доступ практически с любого компьютера подключенного к сети интернет, открыть ее в

СУБД WinISIS и без существенных проблем пополнять эту базу данных. После сохранения изменений в базе данных все данные сразу становятся доступными через Web-интерфейс.

В ходе выполнения проекта база данных в формате RUSMARC, была переделана, написанным конвертором, в более доступный формат XML. Этот формат позволяет работать с базой данных в обход сервера протокола z39.50, что в некоторых местах существенно упрощает процесс написания некоторых скриптов. Так же был написан скрипт, который выдает результаты поиска в этой базе данных в требуемом формате.

Выше приведены примеры запросов и результат вывода записи в формате XML (рис. 34), библиографических карточек с выбранными полями (рис. 35), библиографической карточки с полной информацией (рис. 36), в табличном виде (рис. 37).

В результате проделанной работы был создан Web-шлюз для доступа к актуальной копии прототипа электронного каталога Института археологии и этнографии СО РАН с возможностями поиска записей по различным параметрам. Вывод результата осуществляется в различных форматах, понятных для пользователя (библиографическая карточка, таблица, XML).

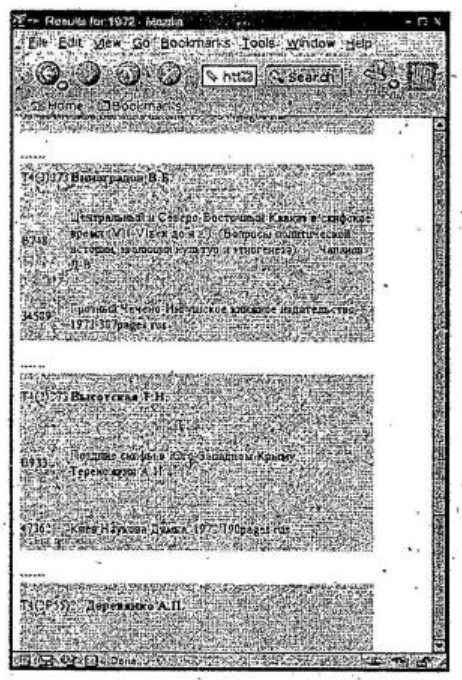


Рис. 36. Результат вывода в виде библиографической карточки с полной информацией.

Ларичев В.Е.	Восточной Азии.	1972	414 с.
Ларичев В.Е.	ПАлеолит Северной, Центральной и Восточной Азии.	1972	414 с.
Ларичев В.Е.	ПАлеолит Северной, Центральной и Восточной Азии.	1972	414 с.
Литвинский Б.А.	Древние кочевники "Крыши мира"	1972	268 с
Матюшин Г.Н.	У колыбели истории	1972	255
Матюшин Г.Н.	У колыбели истории	1972	255
Окладников А.П	Сокровища Томских писаниц. Наскальные рисунки эпохи неолита и бронзы	1972	255 с
Окладников А.П	Петроглифы Средней Лены	1972	270
Окладников А.П.	Петроглифы Средней Лены	1972	270
Окладников А.П.	Петроглифы Средней Лены	1972	270
Стоколос В.С.	Культура населения каменного века Южного Зауралья	1972	
Третьяков В.П.	Культура ямочно-гребенчатой керамики в лесной полосе европейской части СССР	1972	136
Худяков Ю.С.	Археологические открытия 1972 года	1973	520
Худяков Ю.С.	Археологические открытия 1971 года	1972	574
Худяков Ю.С.	Археологические открытия 1971 года	1972	574
Членова Н.Л.	Хронология памятников Карасукской эпохи	1972	247 с

Рис. 37. Результат вывода в табличном виде.

6. Информационная система по подготовке годовых научных отчетов

В секторе археологической теории и информатики (САТИ) ИАЭТ СО РАН с момента его образования в 1996 г. осуществляется комплексная программа исследований по созданию и развитию проблемно-ориентированной среды по гуманитарным наукам и разработке на этой основе гуманитарных информационных ресурсов.

Одним из направлений является разработка ресурсов, освобождающих научных сотрудников от рутинной работы. Создаваемая система в первую очередь предназначена для ученых секретарей научных подразделений Института археологии и этнографии СО РАН. Известно, что ежегодные научные отчеты отнимают много времени. Большая часть времени уходит на рутинную работу по сбору и перепроверке статистической информации (например, список сотрудников, их публикаций, экспедиционных поездок, участий в конференциях и т.д.).

Информация должна храниться из года в год и постоянно пополняться. При этом желательно организовать так, чтобы каждый сотрудник имел возможность добавить или изменить ту часть, которая касается лично его. Конечно эта работа требует некоторой систематизации.

Можно попытаться хранить всю информацию в текстовых файлах (например, в формате Word или Excel, или же в каком-либо ином). Однако этот способ хранения данных может порождать характерный

для подобных таких случаев проблемы. Например::

– проблему синхронизации при попытках внести изменения – один редактирует, остальные вынуждены ждать;

– необходимость соответствующего программного обеспечения для чтения и редактирования общих файлов;

– вольность использования форматирования разными людьми в силу разных представлений о конечном документе.

Оптимальным решением является отделение информации от формата ее представления.

Для этих целей идеально подходит инструментарий, включающий современные базы данных и способы доступа к ним. Современные инструменты манипулирования данными в подобных базах позволяют их пользователям:

– выполнять одновременное и независимое редактирование отдельных записей;

– формировать итоговый отчет в произвольном формате.

Кроме того, в качестве доступа к информации, хранимой в базах данных, целесообразно использовать веб-интерфейс. Эта целесообразность основана на бесспорных преимуществах использования веб-интерфейса. Этот механизм доступа к данным:

– обеспечивает доступность с любого компьютера, подключенного в сеть;

– не требует установки дополнительного программного обеспечения.

Таким образом, наряду с решением перечисленных выше проблем ориентация на использование баз данных в подготовке и ведении научных отчетов дает возможность научным сотрудникам, их руководителям и ученым секретарями получать доступ к их ресурсам почти из любого места.

Исходя из сложившихся традиций оформления и порядка ведения отчетов, в секторе археологической теории и информатики были вычленены основные составляющие статистической информации для отчета, которые мы будем в дальнейшем называть объектами:

– гранты;

– конференции (рис. 38);

– экспедиции;

– публикации сотрудников сектора;

– международное сотрудничество;

– сотрудники сектора.

В силу некоторых ограничений реляционной базы данных в ней нельзя хранить объект произвольной структуры. Это означает, что каждое поле хранимого объекта должно быть простого типа (например текст или число). Поэтому для хранения подобных объектов могут потребоваться дополнительные структуры, которые мы назовем подобъектами.

В подобъектах предполагается хранить те части объекта, которые не укладываются в простой тип. Например, список участников, соавторов, список работ и т.п. Подобъекты по своей структуре аналогичны объектам. Причем, если их структура не укладывается в простые типы, то они могут так же содержать подобъекты более низкого ранга.

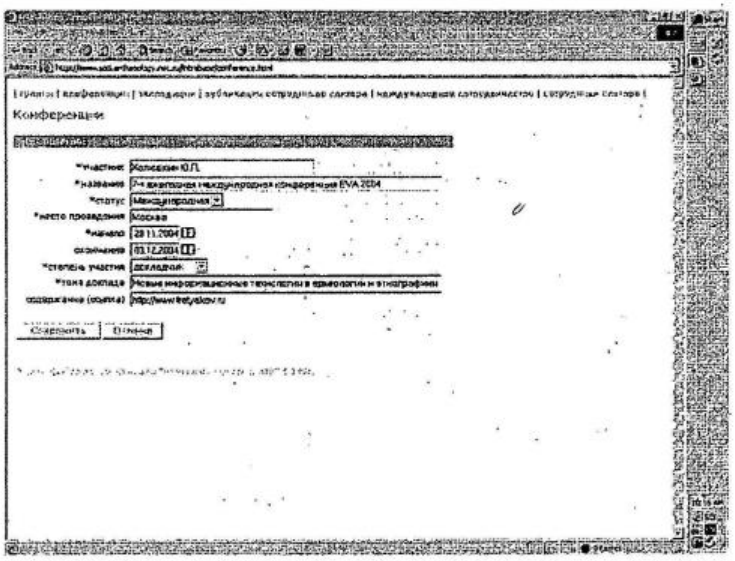


Рис. 38. Пример заполнения формы по отчету об участи в научной конференции.

В указанном выше случае все подобъекты оказались простого типа. В итоге были выявлены следующие объекты и подобъекты:

- ***гранты***
  - ***участники гранта***
- ***конференции***
- ***экспедиции***
  - ***участники экспедиции***
- ***публикации сотрудников сектора***
  - ***соавторы публикации***
- ***международное сотрудничество***
- ***сотрудники сектора***
  - ***научная деятельность сотрудника***

На основе выявленной структуры и ориентировочного содержания отчетов для каждого объекта и подобъекта были созданы таблицы в базе данных и сформированы скрипты для отображения списка уже существующих в базе записей, их редактирования и создания новых.

Полученная система позволяет вводить, редактировать и просматривать всю статистическую информацию. После добавления скрипта, формирующего отчет, на основе введенной информации получаем готовую систему по хранению, редактированию и формированию отчета.

7. Разработка биографической базы данных археологов и этнографов Сибири и Дальнего Востока

Одной из важных задач создания и поддержки информационного центра сектора археологической теории и информатики является разработка биографической базы данных археологов и этнографов Сибири и Дальнего Востока. Этот проект является вторым этапом программы науковедческих исследований, проводимых сотрудниками сектора в сотрудничестве с другими подразделениями Института археологии и этнографии СО РАН.

Предшествующим этапом явились исследования, в результате которых была сформулирована концепция и программа науковедческих исследований, разработаны два варианта анкет исследователей (отдельно для археологов и этнографов), проведено анкетирование археологов и этнографов Новосибирского научного центра и других регионов Сибири и Дальнего Востока. Собранные первичные данные были внесены в базу данных в формате MS Excel для первичного анализа и переноса в более удобные форматы хранения, доступа и редактирования.

Анкетирование проводилось по большому количеству показателей, объединенных в следующие разделы:

сведения о рождении,
среднее образование,
высшее образование,
соискательство и аспирантура,
кандидатская диссертация,
докторская диссертация,
членство в академиях и научных обществах,
трудовой путь,
педагогическая деятельность,
руководство соискателями и аспирантами,
область научных интересов.

Данные анкеты в виде 15 таблиц, сведенные в одну базу данных, стали исходными для разработки биографической базы данных археологов и этнографов.

Исходя из формата и структуры данных для наполнения, доступа и редактирования информации базы данных была СУБД CDS/ISIS for Windows (WinISIS).

Основанием для выбора послужили параллельные разработки по библиографическим базам данных публикаций и коллекций, ведущиеся в секторе, для которых использовался CDS/ISIS. Эта система управления базами данных специально создавалась и используется для ведения баз данных, содержащих

библиографическую информацию. Другой причиной явился текстовый формат переменной длины анкетных данных.

Кроме этого, дополнительными аргументами стали особенности, отличающие CDS/ISIS от других систем управления базами данных, разработанных для общих целей [Бакстон, Хопкинсон, 2002: 15]:

– данные заносятся в поля, поля могут быть использованы с возможным отложенным именованием при описании полей;

– длина полей переменна, благодаря использованию справочника стандарта ISO-2709, в котором каждая запись содержит список полей и указателей позиций данных, относящихся к каждому полю;

– возможны повторяющиеся поля (до 999 повторений);

– возможны подполя, которые указываются специальными дескрипторами, что позволяет обрабатывать разными способами части полей;

– CDS/ISIS использует инвертированные файлы (индексные файлы) для ускорения поиска в базе данных и различные техники индексирования, при этом в индексный файл вносятся различные элементы данных записи и становится возможным индексирование по целому полю, отдельному подполю, каждому слову и другое;

– CDS/ISIS использует поиск по свободному тексту, связанный с последовательным просмотром записей и проверкой их содержания [Бакстон, Хопкинсон, 2002: 55].

Локальная система управления базами данных WinISIS характеризуется следующим:

– выбор базы данных из числа хранящихся в каталогах баз данных;

– для выбранной базы данных предоставляет интерфейс для просмотра, коррекции и поиска данных в обычном и экспертном режиме;

– предоставляет средства индексирования базы данных;

– в интерфейсе есть средства для описания полей и подполей базы данных, формирования индекса, задания формата печати и создания рабочих листов;

– интерфейс имеет импорт в стандарте ISO-2709, экспорт в XML.

Разработка биографической базы данных археологов и этнографов проводилась в два этапа.

Первый этап имел своим результатом базу данных на локальной машине в среде WinISIS и включал следующие решения:

– сведение всех таблиц анкетных данных к одной размером 330:210 в среде Excel;

– приведение полученной таблицы с помощью бейсик-системы автоматизации Excel к формату для использования программного средства Fangorn, преобразующего текстовые данные в коде ASCII в стандарт данных ISO-2709;

– преобразование данных анкеты в стандарт ISO-2709 с помощью программного средства Fangorn;

– преобразование результатов обработки в кодировку Windows с помощью утилиты Tcode;

– проектирование базы данных и импортирование данных анкеты формата ISO-2709 средствами интерфейса системы WinISIS.

Полученная база данных содержит 210 записей в 38 полях и в более чем 90 подполях.

Здесь приводится список полей биографической базы данных археологов и этнографов:

1. Номер анкеты
2. Фамилия имя отчество
3. Пол
4. Дата рождения
5. Место рождения
6. Национальность
7. Среднее образование
8. Служба в армии
9. Высшее образование
10. Соискательство аспирантура
11. Кандидатская диссертация
12. Докторская диссертация
13. Членство в Российских академиях
14. Членство в Российских научных обществах
15. Членство в научных обществах иностранных государств
16. Членство в международных научных организациях
17. Членство академий иностранных государств
18. Почетные научные звания

19. Ученые звания
20. Премии за научные достижения
21. Государственные ордена за заслуги в науке
22. Государственные медали за заслуги в науке
23. Другие государственные ордена
24. Другие государственные медали
25. Медали именные и медали научных обществ
26. Трудовой путь
27. Педагогическая деятельность
28. Руководство аспирантами и соискателями
29. Интересы отрасль науки
30. Интересы область науки
31. Интересы раздел науки
32. Исследуемый период
33. Пространственные интересы
34. Открытия в науке
35. Научные достижения
36. Выбывание из исследовательской тематики
37. Домашний адрес телефон
38. Служебный адрес телефон

На втором этапе разрабатывался Web-интерфейс к созданной базе данных.

Интерфейс был разработан на базе сайта сектора археологической теории и информатики института археологии и этнографии СО РАН. Он позволяет просматривать базу данных, вводить новые данные, искать данные в базе и индексировать их.

Web-интерфейс включает следующие страницы:

– титульная;
– функции интерфейса;
– просмотр записей;
– поиск;
– ввод новая запись;
– ввод добавление полей;
– индексирование;
– авторизация.

Для реализации Web-интерфейса использована библиотека функций ISIS_DLL, разработанная фирмой BIREME, Sao Paulo, август 1997:

Библиотеку можно использовать из языков программирования C, C++, Pascal, Delphi, Visual Basic и др. Имеется также расширение PHP_ISIS support , которое позволяет работать с этой библиотекой на языке PHP. Это расширение было использовано при разработке интерфейса.

Библиотека содержит около 100 функций [BIREME, ISIS_DLL,1997] (в скобках указаны префиксы разделов библиотеки):

– функции приложения (App);
– функции Dll (Dll);
– функции Связи (Lnk);
– функции записи (Rec);
– функции пространства (Spa;)
– функции поиска (Src);
– функции выражения (Trm).

К особенностям этой библиотеки можно отнести то, что для реализации функций по управлению базой данных она предоставляет пользователю-программисту аппарат для использования оперативной памяти компьютера в виде таких понятий [BIREME, ISIS_DLL, 1997: 5] как приложение, пространство, полка. Применение функций библиотеки при этом оказывается эффективным по скорости обработки запросов, потому что большое количество данных в ходе реализации алгоритмов функций размещается и хранится в оперативной памяти компьютера, создавая в ней по существу контекст обработки запросов, что и увеличивает скорость обработки запросов к базе данных.

Таким образом, библиотека функций ISIS_DLL позволяет в приемлемое время обрабатывать запросы на управление базой данных в виде HTML-форм в Web-обмене.

Например, можно обработать без большой видимой задержки запрос на просмотр записей, когда база данных индексирована по <фамилия имя отчество>, выбрав для просмотра сразу 10 строк индекса.

Кроме этого, контекст обработки позволил разработчикам библиотеки реализовать алгоритмы функций ISIS/DLL с учетом расположения данных в оперативной памяти, что сказалось положительным образом на таких свойствах набора функций библиотеки, как простота использования и "интегральность" (мощность) функций. Например, простота изменения индексных полей без переиндексации базы данных при коррекции индексных полей, то есть контекст обработки запросов библиотекой ISIS/DLL в этом случае служит базой для конвергенции свойств функций обработки.

Программные средства, реализующие Web-интерфейс, разрабатывались с использованием процессора гипертекста PHP версии 4.2.3, организованы и введены в Web-систему в виде каталога (около 80 строк).

Из особенностей программного обеспечения нужно отметить следующие:

1. Обработчиком HTML-формы является PHP-модуль содержащий эту форму. Это позволяет организовать некоторую динамику окон Web-интерфейса, которая выражается в том, что определяя какие-либо данные в окне интерфейса, в результате получаешь это окно уже расширенным. Например, в окне ввода данных отображаются разделы вводимых данных. Определяя раздел данных, получаешь окно уже с показателями выбранного раздела, сохраняя при этом в окне перечень разделов для возможных последующих запросов.

2. Определенным образом организована логика обработки HTML-форм, которая позволяет синхронизировать ее с поступлением соответствующей информации в условиях Web-обмена. Например, пока не выбран раздел вводимых данных нельзя отобразить поля для ввода данных этого раздела, так как поля отображаются группами по разделам, ввиду большого их количества.

Определяя перспективы развития Web-интерфейса, нужно отметить следующее:

1. Для выбранной базы данных CDS/ISIS можно создавать несколько индексных наборов для данных и использовать их для обработки [BIREME, ISIS_DLL: глава 4], Такие возможности индексирования базы данных, поддержанные библиотекой функций ISIS_DLL, позволяют более оперативно, чем это реализовано в интерфейсе WinISIS, задавать различные "срезы" базы данных и в сочетании с возможностью ограничения предъявляемых данных, реализованной в Web-интерфейсе улучшить свойства Web-интерфейса, с помощью которых база данных становится более "прозрачной".

2. Наряду с имеющимися в локальной системе Winisis средствами коррекции базы данных, функции ISIS_DLL дают разработчику инструменты для коррекции данных в режиме Web-интерфейса, с учетом опыта разработки в Web-интерфейсе ввода новых данных.

3. Учитывая конфиденциальный характер некоторых данных, необходимо корректная реализация системы авторизации и разграничения доступа в Web-интерфейсе.

4. Данная база данных в ближайшее время будет интегрирована в информационные ресурсы библиографических баз данных, разработанных также в среде CDS/ISIS для Windows.

II МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В АРХЕОЛОГИИ

В процессе работы над блоком проекта в 2003 г. была опубликована монография: **Деревянко А.П., Холюшкин Ю.П., Ростовцев П.С., Воронин В.Т.** Корреляция среднепалеолитических индустрий Ближнего Востока и Кавказа. Новосибирск: Изд-во СО РАН, 2002. 186 с. и в двух сборниках научных трудов сотрудников сектора информатики (Информационные технологии в гуманитарных исследованиях, Вып. 5-6, Новосибирск, 2003) При этом применялся и был разработан комплекс оригинальных алгоритмов, вычислительных процедур и схем, которые обеспечили многоэтапный анализ археологических данных по разным алгоритмам и правилам. В основе созданных алгоритмов анализа лежит комплекс взаимосвязанных идей: серый анализ для предварительной классификации археологических объектов, выявление структуры таблицы с помощью упорядочения строк и столбцов, нейтрализации неопределенных элементов таблицы, разбиение ее на связные области, построение и применение адекватного критерия качества разбиения, выбор последовательности процедур и критерия остановки разбиения, алгоритмы перемещения границ, анализ матриц сопряженности, понятие и смысл автоматизации типологического группирования, построение критерия качества кластеризации для одномерного и многомерного случаев, для неколичественной переменной, устранение влияния малых групп, логика группирования, интерпретация результатов группирования, устойчивость выявленных структур, устойчивость в анализе структуры таблиц сопряженности, устойчивость при типологическом группировании и т.д. Описанный аналитический базис методологии статистического исследования объектов дополняется схемой применения в анализе данных метода повторной выборки с возвращением (bootstrap). Эти положения, теоретически обоснованные и сформулированные на языке алгоритмов и процедур, легли в основание новой методологии комплексного анализа, пригодной даже для "плохих" данных, какими являются данные археологических исследований.

В развитие этих идей, в ходе реализации проекта в 2003 г. были разработаны следующие алгоритмы статистического анализа археологических исследований.

1. Бета-регрессия как метод восстановления условного распределения случайной величины

2. Построение обобщенной классификации

3. Статистика для сравнения классификаций

1. Бета-регрессия как метод восстановления условного распределения случайной величины

1.1. Постановка задачи.

В развитие идей предшествующего этапа исследований по обработке археологических коллекций участниками проекта была поставлена задача найти такой способ восстановления параметров Бета-распределения, который позволил бы максимально точно восстановить как отдельные значения, так и характер зависимости a и b от z, сведя к минимуму разрушительное действие статистического шума.

Входными данными для рассматриваемого метода восстановления двумерного распределения является выборка из N наблюдений, описываемая случайными величинами x и z. Предполагается, что x изменяется в интервале (0,1) и условная плотность распределения $f(x|z)$ может быть аппроксимирована Бета-распределением с параметрами $a(z)$, $b(z)$:

$$f(x \mid z) = \frac{x^{a(z)-1}(1-x)^{b(z)-1}}{\mathrm{B}(a(z), b(z))} \qquad (1)$$

Зависимость параметров a и b от z может быть достаточно сложной. Поэтому для ее восстановления необходимо выбрать класс функций, позволяющий строить приближение практически любой функции и при этом полностью контролировать точность аппроксимации. С учетом этих требований мы выбрали кубические сглаживающие сплайны, математический аппарат построения которых описан в [Морозов, 1970: 54-58; Носач, 1994: 194-213], а программный код открыто выложен в интернете [IS01R, 2002].

Первым шагом восстановления двумерного распределения является выбор значений z_i, $i=\{0,1, \ldots, m\}$, которые разбивают весь диапазон изменения z на интервалы (z_{i-1}, z_i). Число интервалов выбирается,

исходя из двух противоположных требований. Во-первых, каждый интервал должен быть достаточно широким (содержать много наблюдений), чтобы получить хорошие оценки значений a_i и b_i. Во-вторых, интервалы не должны быть широкими, чтобы не терять информацию о зависимости a и b от z в процессе усреднения. Итак, следует избегать как слишком малого числа интервалов, так и слишком большого.

Кроме количества интервалов, необходимо выбрать и способ разбиения. Здесь существует, по крайней мере, два решения.

Во-первых, сетка с равномерным шагом, где шаг сетки:

$$h=(z_m-z_0)/N \qquad (2)$$

Во-вторых, сетка с равнонаполненными ячейками, где объем i-ой ячейки:

$$N_i = \frac{N - \sum_{k=1}^{i-1} N_k}{m-i+1} \quad i=\{1,\ldots, m\} \qquad (3)$$

Процедура разбиения немного усложняется в том случае, когда наблюдения с одним и тем же z оказываются в разных ячейках.

Если наблюдения распределены равномерно по z, то имеет смысл выбрать равномерную сетку, если же нарушения равномерности существенные, то лучшим вариантом будет разбиение по квантилям.

После разбиения выборки на ячейки по z, можно переходить к вычислению a_i и b_i для каждой ячейки $i=\{1, \ldots, m\}$. Получить оценки параметров a_i и b_i проще всего методом моментов:

$$a_i = \overline{x_i}(\overline{x_i}(1-\overline{x_i})/s_i^2 - 1) \qquad (4)$$

$$b_i = (1-\overline{x_i})(\overline{x_i}(1-\overline{x_i})/s_i^2 - 1) \qquad (5)$$

где $\overline{x_i}$ - выборочное среднее x, а s_i^2 – выборочная дисперсия (смещенная).

Альтернативный способ получить оценки a и b основан на процедуре максимизации функции правдоподобия, имеющей смысл вероятности совместного наблюдения выборочных данных x_j при имеющейся плотности распределения (1):

$$L_i = \prod_{j=1}^{N_i} x_{ij}^{a(z_i)-1}(1-x_{ij})^{b(z_i)-1} / B(a(z_i), b(z_i)) \qquad (6)$$

Учитывая ограничения точности и диапазона представления чисел в компьютере, удобнее оптимизировать не саму функцию правдоподобия, а ее логарифм, вместо произведения накапливая суммы:

$$\ln L_i = \sum_{j}^{N_i}(a(z_i)-1)\ln x_{ij} + \sum_{j}^{N_i}(b(z_i)-1)\ln(1-x_{ij}) - \sum_{j}^{N_i}\ln(B(a(z_i), b(z_i)) \qquad (7)$$

1.2. Генерация исходных данных

Для проверки работоспособности метода необходимо большое количество выборок с точно известным распределением в генеральной совокупности. Отсутствие реального источника данных требуемого объема, разнообразия и качества вынуждает обратиться к альтернативному источнику – методу Монте-Карло, то есть искусственной генерации данных с использованием датчика псевдослучайных чисел. Такой подход хорош тем, что позволяет практически мгновенно получать выборки произвольного размера, в точном соответствии с теоретически заданными параметрами распределения на генеральной совокупности.

Процедура генерации выборки включает следующие шаги:

1. Задание границ изменения z: (z_{min}, z_{max}).
2. Задание функциональной зависимости $a(z)$ и $b(z)$.
3. Задание объема выборки N.

Генерация каждого из N наблюдений:

4. Генерация псевдослучайного числа z в интервале (z_{min}, z_{max}).
5. Определение параметров Бета-распределения по z: $a(z)$ и $b(z)$.
6. Генерация псевдослучайного числа P в интервале $(0, 1)$.
   Здесь P имеет смысл вероятности.
7. Определение x, при котором интегральная функция распределения принимает значение P:

$$F(x,z) = \int_0^x f(t,z)dt = \int_0^x \frac{t^{a(z)-1}(1-t)^{b(z)-1}}{\mathrm{B}(a(z),b(z))}dt = P \qquad (8)$$

8. Добавление полученных значений $\{z, x\}$ к выборке.

Проиллюстрируем выполнение этой процедуры на примере.

1. Границы изменения *z*:
   - $z_{min} = 0,\ z_{max} = 100.$
2. Зависимости *a(z)* и *b(z)*:
   - $a(z) = 0.5 + 0.05 \cdot z$ (9)
   - $b(z) = 5.5 - 0.05 \cdot z$ (10)
3. Объем выборки *N*:
   - $N = 400$

❖ Генерируем первое из 400 наблюдений (см. рис. 39):

4. Генерируем случайное число z в интервале от 0 до 100:
   - $z = 73.009$
5. Определяем параметры Бета-распределения:
   - $a(73.009) = 0.5 + 0.05 \cdot 73.009 = 4.150$
   - $b(73.009) = 5.5 - 0.05 \cdot 73.009 = 1.850$
6. Генерируем случайное число P в интервале от 0 до 1:
   - $P = 0.354463126$
7. Определяем x, при котором интегральная функция распределения принимает значение $P = 0.354463126$:
   - $x = 0.6394$

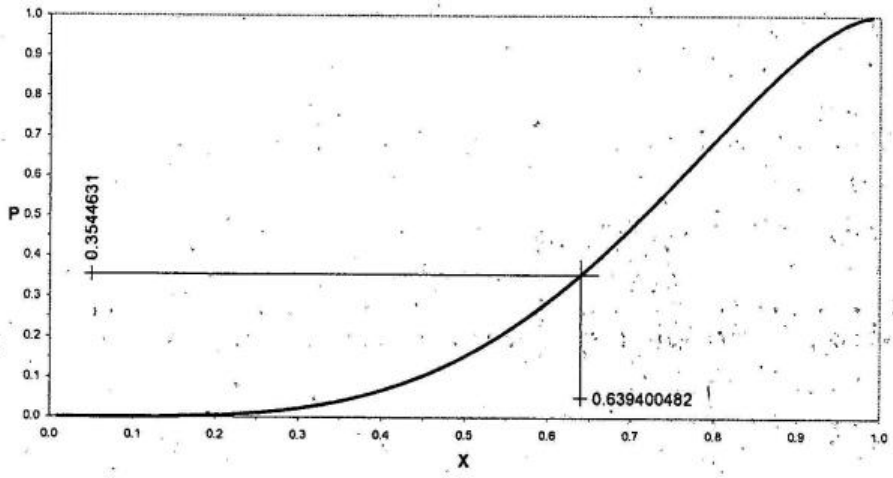


Рис. 39. Преобразование равномерно распределенного случайного числа P в наблюдаемое значение x по Бета-распределению F(x,a(z),b(z)).

8. Добавляем полученные значения *{z=73.009, x=0.6394}* к выборке (рис. 40):

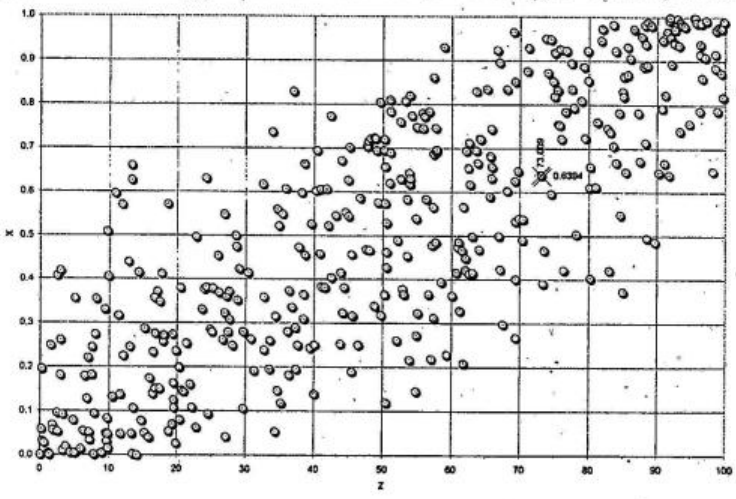


Рис. 40. Выборка из 400 наблюдений по Бета-распределению с параметрами (9, 10). Крестиком обозначено наблюдение *{z=73.009, x=0.6394}*.

1.3. Разбиение на интервалы и оценивание параметров распределения

Как уже было сказано выше, первым шагом восстановления двумерного распределения является задание сетки z_i, $i=\{0,1, \ldots, m\}$. Здесь $z_0 = z_{min}$, а $z_m = z_{max}$. Количество ячеек m должно быть таким, чтобы внутри каждой ячейки было достаточно наблюдений для вычисления a_i и b_i. Если сетка с постоянным шагом не позволяет вычислить параметры во всех ячейках, можно, во-первых, уменьшить m, увеличив наполненность ячеек, а во-вторых, перейти к сетке с равнонаполненными ячейками. В этом случае z_i определяется как значение квантиля i/m.

После того, как интервалы будут получены, можно приступать к расчету a_i и b_i. Для этого достаточно вычислить на каждом интервале среднее значение $\bar{x}_i$ и смещенную оценку дисперсии s_i^2:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j}^{N_i} x_{ij} \qquad (11)$$

$$s_i^2 = \frac{1}{n_i}\sum_{j}^{N_i} (x_{ij} - \bar{x}_i)^2 \qquad (12)$$

После этого можно рассчитать a_i и b_i, применив формулы (4) и (5). Если мы хотим получить оценки методом максимального правдоподобия, необходимо найти максимум логарифма функции правдоподобия (7). А в качестве начального приближения при поиске максимума можно воспользоваться только что полученными (методом моментов) оценками a_i и b_i.

При разбиении приведенной выше выборки на 10 равных интервалов мы получим следующие результаты (см. таблицу и рис. 41 ниже):

Z_{i-1}-Z_i	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
N_i	46	41	35	33	39	52	44	31	37	42
$\bar{z}_i$	5.11	15.56	25.17	35.57	44.86	54.20	64.87	75.51	85.08	94.79
$\bar{x}_i$	0.130	0.239	0.300	0.379	0.510	0.562	0.585	0.767	0.769	0.899
s_i^2	0.019	0.032	0.021	0.035	0.031	0.042	0.033	0.028	0.033	0.012
a_i^z	0.755	1.278	1.759	2.279	2.743	3.210	3.744	4.275	4.754	5.239
a_i^{moment}	0.638	1.099	2.695	2.145	3.581	2.737	3.718	4.201	3.418	6.129
a_i^{maxL}	0.563	0.853	2.442	2.262	3.748	2.900	3.516	5.162	3.396	6.081
b_i^z	5.245	4.722	4.241	3.721	3.257	2.790	2.256	1.725	1.246	0.761
b_i^{momen}	4.265	3.503	6.299	3.508	3.438	2.129	2.637	1.275	1.024	0.692
b_i^{maxL}	3.860	2.839	5.786	3.663	3.634	2.295	2.439	1.599	1.001	0.682

где Z_{i-1}-Z_i – границы ячейки,

N_i – число наблюдений в ячейке,

$\bar{z}_i$ – среднее значение z в ячейке,

$\bar{x}_i$ – среднее значение x в ячейке i, рассчитанное по формуле (11),

s_i^2 – смещенная оценка дисперсии в ячейке (12),

a_i^z – теоретическое значение, рассчитанное по формуле (9),

a_i^{moment} – оценка a_i по методу моментов (4),

a_i^{maxL} – оценка a_i по методу максимального правдоподобия,

b_i^z – теоретическое значение, рассчитанное по формуле (10),

b_i^{moment} – оценка b_i по методу моментов (5),

b_i^{maxL} – оценка b_i по методу максимального правдоподобия.

Как видно из таблицы и рисунка, параметры распределения, восстановленные по случайной выборке, достаточно сильно отклоняются от теоретически заданных. Видно, что не сохраняется монотонность изменения параметров распределения. Иными словами, в восстановленном таким путем двумерном распределении слишком велик уровень статистического шума, который не позволяет увидеть

теоретически заложенной в выборке линейной зависимости a и b от z. Этот шум обусловлен случайным характером формирования выборки и не может быть преодолен иначе, как увеличением ее объема.

В данной работе мы ставим перед собой задачу найти такой способ восстановления параметров Бета-распределения, который позволит максимально точно восстановить как отдельные значения, так и характер зависимости a и b от z, сведя к минимуму разрушительное действие статистического шума.

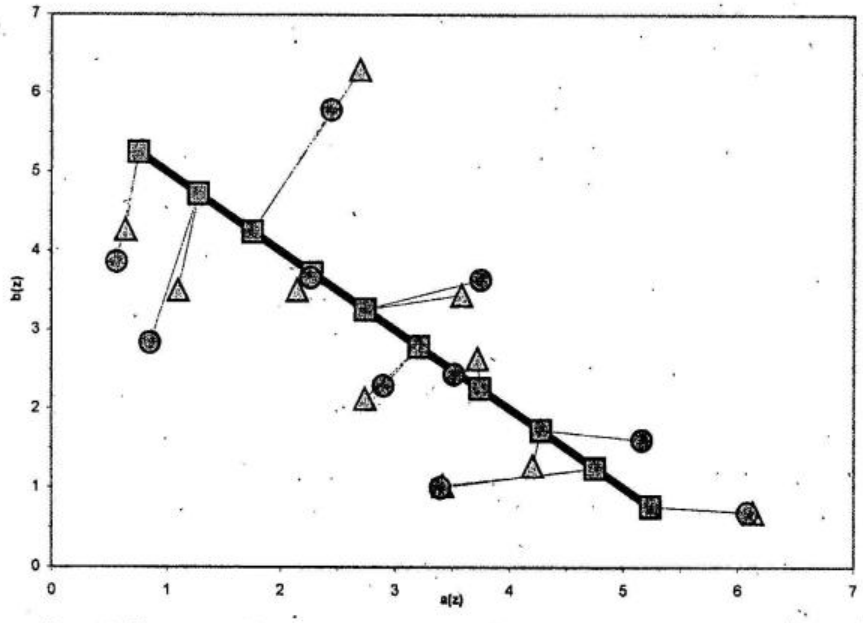


Рис. 41. Параметры Бета-распределения a_i и b_i: квадраты – теория, треугольники – метод моментов, кружки – метод максимального правдоподобия.

1.4. Аппроксимация параметров распределения сглаживающими сплайнами

Чтобы решить поставленную задачу, необходимо найти класс функций, позволяющий аппроксимировать эмпирические данные с контролируемой степенью точности, то есть с произвольно заданной ошибкой аппроксимации. Чем выше точность (меньше допустимая ошибка), тем ближе будет аппроксимирующая функция к эмпирическим точкам. Но, как мы видели на рисунке, точность воспроизведения эмпирических значений далеко не то же самое, что точность воспроизведения теоретически заложенной зависимости, поскольку эмпирические значения в значительной степени поражены статистическим шумом. Должна существовать некоторая оптимальная степень сглаживания, которая бы позволила отсеять шум, сохранив при этом заложенную в данных теоретическую зависимость. Естественно, что восстановленная зависимость будет отличаться от теоретической. Задача состоит в том, чтобы это отклонение минимизировать.

Поиск необходимого класса функций начался с полиномов, коэффициенты которых легко определяются из требования минимизации суммы квадратов отклонений значений полинома от эмпирических точек. Но аппроксимация полиномами обладает весьма существенными недостатками.

Во-первых, для полиномов легко подобрать примеры практически неаппроксимируемых функций. В качестве такого примера можно привести ступенчатую (S-образную) кривую с длинными хвостами. Ни один полином конечной степени не в состоянии дать ее удовлетворительное приближение.

Во-вторых, для сглаживания полиномами нет другой возможности управлять степенью сглаживания, кроме изменения степени полинома. А это очень грубая настройка хотя бы в силу того, что она дискретна.

В то же время в последние десятилетия интенсивно развивается новый раздел современной вычислительной математики - теория сплайнов. Сглаживающие сплайны позволяют не только хорошо интерполировать функции по отдельным точно заданным значениям, но и эффективно строить аппроксимацию эмпирических данных с заданной точностью. При минимальной точности мы получаем чисто линейную зависимость с равной нулю второй производной на всей области определения сплайна. При максимальной точности сплайн становится интерполирующим, то есть проходит строго через все точки. В промежутке от минимальной до максимальной точности параметр сглаживания меняется непрерывно, позволяя найти оптимальную степень сглаживания.

Итак, сплайны действительно являются подходящим классом функций, в точности удовлетворяю-

щим выдвинутым требованиям.

Выбранный нами вид сглаживающих сплайнов [Морозов 1970; IS01R, 2002] минимизирует усредненный квадрат второй производной по всей области определения функции:

$$\int_{z_0}^{z_m} (s''(z))^2 dz \to \min \qquad (13)$$

Кроме того, выполняется ограничение на отклонение сплайна от эмпирически заданных точек:

$$\sum_{i=1}^{m} N_i (s(z_i) - f_i)^2 \le \delta^2 \qquad (14)$$

Здесь квадраты отклонений взвешиваются на количество точек в интервале i, поскольку точность определения значений f_i обратно пропорциональна корню из N_i:

$$\delta(f_i) \sim \frac{1}{\sqrt{N_i}} \qquad (15)$$

Если мы зададим δ^2 в ограничении (14) равной остаточной дисперсии для случая линейной регрессии δ_r^2, то сплайн выродится в линейную функцию от z. Если задать δ^2 равным нулю, то сплайн станет интерполирующим и будет проходить через все эмпирические точки f_i.

Для удобства параметризации степени сглаживания вместо δ^2 будем использовать безразмерный параметр точности λ:

$$\delta^2 = 2(1-\lambda)\delta_r^2 \qquad (16)$$

При уменьшении λ от *1* до ½ δ^2 возрастает от *0* до δ_r^2. Таким образом, λ задает точность аппроксимации, нормированную на единицу. Точность, равная единице, будет означать абсолютную точность. Точность, равная ½, будет означать точность линейного приближения. Точность меньше ½ не будет приводить к дальнейшему огрублению приближения, если пользоваться тем же алгоритмом. Для обобщения мы искусственно продолжим эту зависимость. Что может быть грубее линейной зависимости? Естественно, только отсутствие всякой зависимости, то есть константа. Тогда точность между ½ и *0* будет соответствовать переходу от линейного приближения к среднему значению f:

$$\overline{f} = \frac{1}{N}\sum_{i=1}^{m} N_i f_i \qquad (17)$$

Имеет смысл дополнить область изменения λ еще одним значением: -1. Этому значению можно сопоставить отсутствие не только изменений в распределении, но и фактически отсутствие самого распределения. Для параметров Бета-распределения это означает их равенство единице, при котором распределение вырождается в равномерное.

Изменение кривой, аппроксимирующей зависимость *a(z)*, при изменении параметра точности λ от -1 до 1 можно видеть на рис. 42.

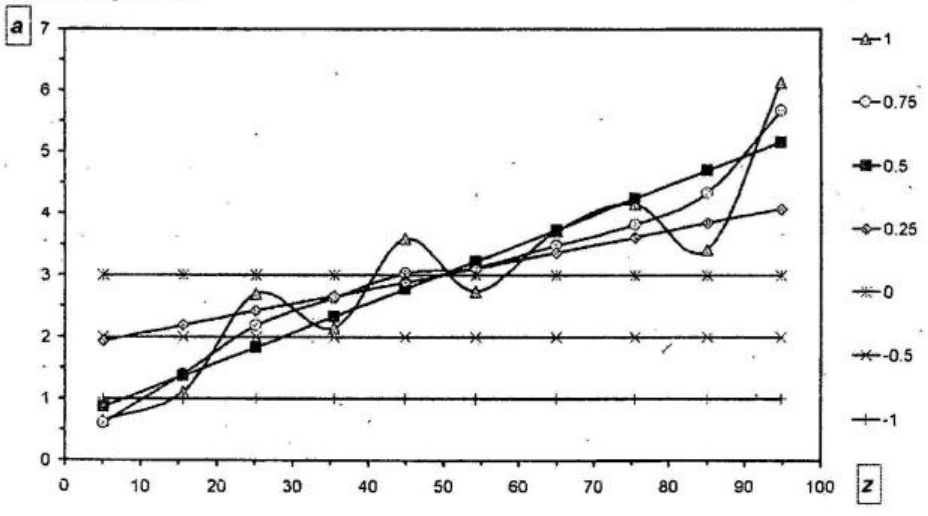


Рис. 42. Восстановление *a(z)* с разной точностью аппроксимации.

В фазовом пространстве {a,b} это же будет выглядеть так:

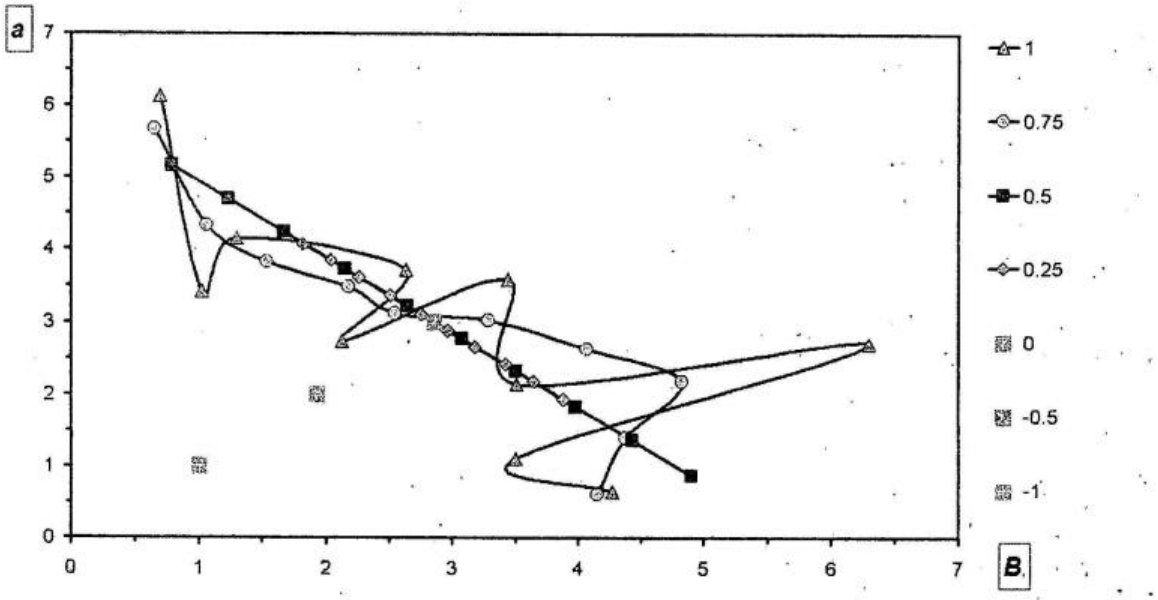


Рис. 43. Восстановление *{a(z), b(z)}* с разной точностью аппроксимации.

Определив вид функции, которая может аппроксимировать параметры распределения с контролируемой точностью, остается найти критерий, который позволит определить оптимальное значение точности.

1.5. Выбор оптимальной степени сглаживания

Отклонения параметров a_i и b_i от теоретических значений $a(z_i)$ и $b(z_i)$ имеют стохастическую природу, то есть вызваны случайным характером формирования выборки. Гипотеза о том, что эти отклонения действительно случайны, а не носят систематический характер, поддается проверке. В математической статистике хорошо известен непараметрический критерий Колмогорова-Смирнова, который позволяет проверить гипотезу о случайности отклонения эмпирического распределения $F_n(x)$ от известного теоретического $F(x)$ на основании статистики D_n:

$$D_n = \sup_x |F_n(x) - F(x)| \qquad (18)$$

Чем больше значение статистики D_n, тем с меньшей вероятностью можно получить его случайно. Чтобы отвергнуть гипотезу о случайности отклонения, эта вероятность должна быть достаточно мала, меньше некоторого порога, например 5% или 0,1%. Но так поступают при сравнении единственного имеющегося у исследователя эмпирического распределения с теоретическим. Мы же имеем дело с целым набором из m эмпирических распределений (по одному на каждый интервал z_{i-1}-z_i, где $i=1..m$), каждому из них соответствует свое теоретическое (восстановленное) распределение. Оказывается, что и в этом случае процедура принятия решения о случайности наблюдаемых отклонений в распределениях ненамного сложнее. Действительно, для этого достаточно отметить тривиальное свойство вероятности, рассчитываемой по статистике D_n. Если гипотеза о случайном отклонении эмпирических распределений имеет место, то рассчитываемая по критерию Колмогорова-Смирнова вероятность $P(D > D_n)$ будет распределена равномерно (рис. 44).

Отсюда сразу вытекает формулировка критерия оптимальности: при наилучшем сглаживании параметров Бета-распределения наблюдается наиболее близкое к равномерному распределение вероятностей $P_i(D > D_n)$, вычисленных для всех интервалов *(i)* по критерию Колмогорова-Смирнова. Для определения близости этого распределения к равномерному можно еще раз использовать тот же критерий (Колмогорова-Смирнова), взяв на этот раз в качестве теоретического распределения равномерное (рис. 45).

Изменение плотности условного распределения на сетке из 10 интервалов в результате сглаживания, можно видеть на рис. 46 и 47.

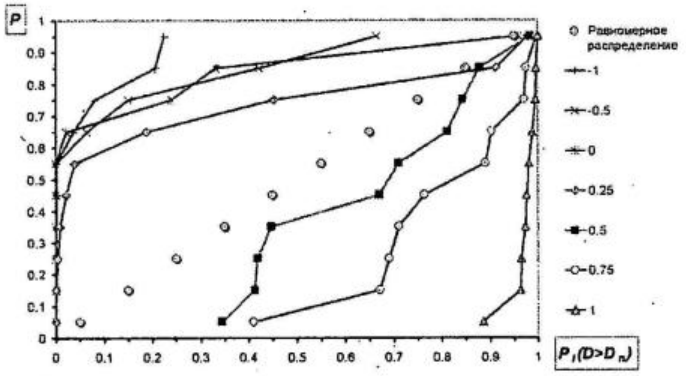


Рис. 44. Зависимость формы распределения $P_i(D > D_n)$ от точности аппроксимации.

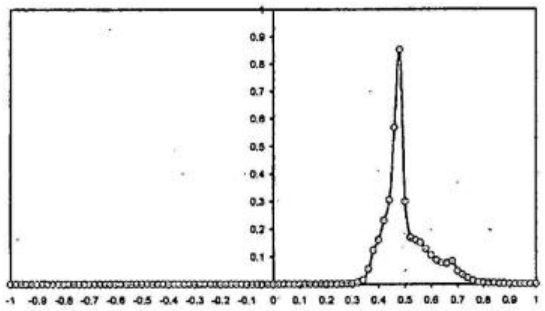


Рис. 45. Поиск оптимальной точности аппроксимации по максимальной близости распределения $P_i(D > D_n)$ к равномерному.

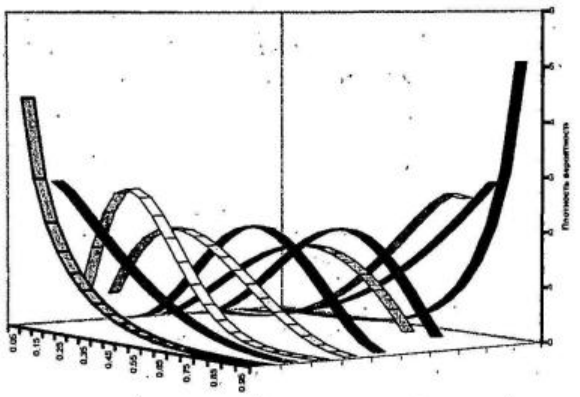


Рис. 46. Плотность условного распределения по интервалам.

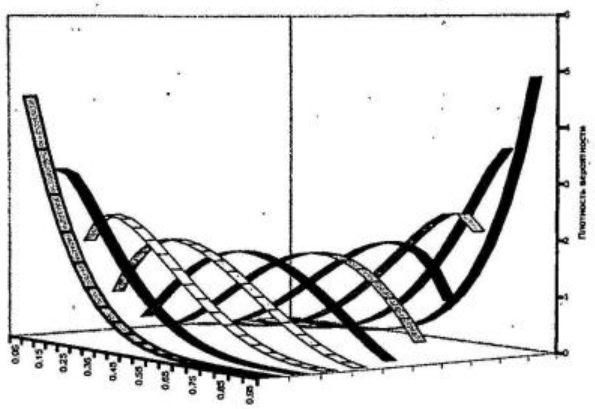


Рис. 47. Плотность условного распределения после сглаживания.

1.6. Проверка точности восстановления Бета-распределения

Расчеты на контрольной выборке показали удовлетворительное восстановление параметров распределения при некоторых разбиениях (рис. 48). Вместе с тем обнаружилось искажение формы кривой *a-b* при числе интервалов меньше 10.

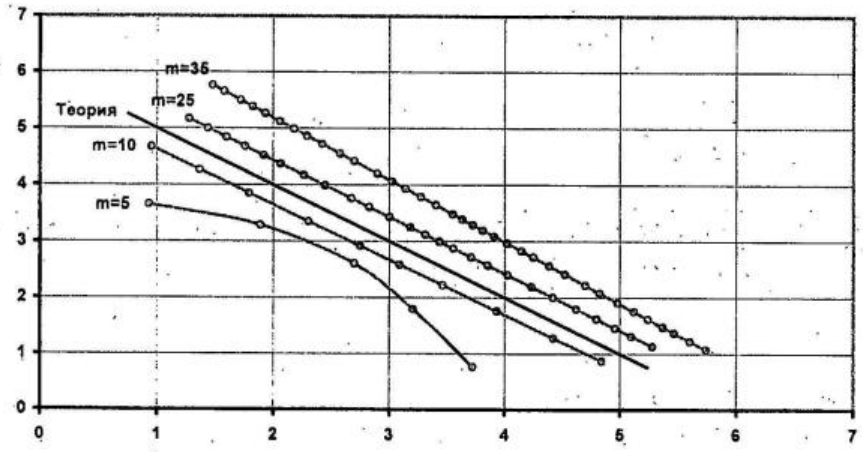


Рис. 48. Восстановление параметров Бета-распределения *a* и *b* на контрольной выборке объемом 400 наблюдений при различных разбиениях (m - число интервалов).

Этот эффект имеет простое объяснение: при малом числе интервалов каждый из них захватывает широкую область значений *z*, что приводит к смешению распределений внутри интервала и искусственному увеличению выборочной дисперсии s_i^2. А поскольку оценки *a* и *b* обратно пропорциональны дисперсии (см. формулы 4,5), то ее увеличение приводит к уменьшению параметров распределения тем более заметному, чем более различаются смешиваемые распределения. Нами были проведены дополнительные вычислительные эксперименты, которые выделили эффект смешивания распределений внутри интервалов в чистом виде. Полученные в этих экспериментах кривые *a-b* дали в точности ту же картину, что на рисунке 10 при **m=5**.

Кроме искажения формы кривой при разбиении на малое число интервалов, наблюдается систематическое увеличение оценок параметров *a* и *b* при увеличении числа интервалов. На рис. 49 показано, как растут средние оценки параметров (теоретическое значение равно 3.0) и разброс этих оценок относительно средних (кривая в нижней части рисунка).

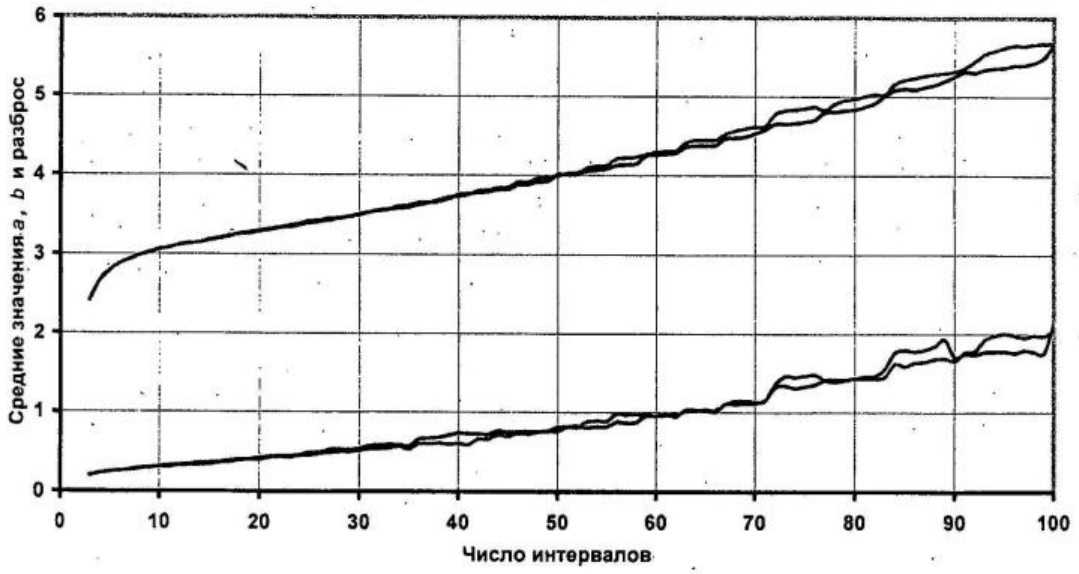


Рис. 49. Зависимость восстановленных значений a и b (сверху) и их среднеквадратичных разбросов (снизу) от числа интервалов m.

Этот эффект обусловлен тем, что при увеличении числа интервалов количество наблюдений внутри каждого из них уменьшается и разброс оценок a_i и b_i возрастает. Когда оценки рассеиваются симметрично относительно теоретического значения, систематического смещения восстановленных значений не происходит. В нашем же случае наблюдается длинный хвост справа, в результате чего усреднение дает завышенные оценки a и b, что мы и наблюдаем.

Для проверки этого предположения мы сгенерировали выборку на основе Бета-распределения с фиксированными параметрами a=0.5 и b=5.5 объемом 400 наблюдений. Имитируя разбиение выборки на 1, 2, 4 и т.д. интервалов скользящим интервалом соответствующего размера, мы получили средние оценки, которые сведены в таблице:

m	1	2	4	5	8	10	16	20	25	40	50	80	100
$\bar{a}$	0.519	0.524	0.524	0.529	0.542	0.554	0.595	0.622	0.658	0.766	0.841	1.254	1.996
$\bar{b}$	5.218	5.240	5.352	5.374	5.529	5.724	6.313	6.672	7.158	8.895	10.234	21.838	38.258
$\bar{x}$	0.0905	0.0910	0.0896	0.0900	0.0898	0.0896	0.0900	0.0906	0.0909	0.0906	0.0905	0.0905	0.0904
$\overline{x^2}$	0.0204	0.0206	0.0201	0.0202	0.0201	0.0200	0.0202	0.0205	0.0206	0.0205	0.0204	0.0204	0.0204

Из таблицы видно, что оценки $\bar{a}$ и $\bar{b}$ существенно отличаются от исходных, начиная уже с m = 10.

В то же время, оценки $\bar{x}$ и $\overline{x^2}$ не зависят от разбиения и восстановленные по ним a и b также не должны от него зависеть. Тем не менее, алгоритмическая реализация такого подхода является нетривиальной задачей и требует дополнительной проработки, что предполагается осуществить в ходе дальнейшего развития метода.

1.7.Определение минимального объема выборки

Если представить себе последовательный ряд выборок постепенно уменьшающегося объема от бесконечного количества до одного наблюдения, нетрудно увидеть, что к концу этого ряда мы полностью теряем какую бы то ни было информацию о распределении в генеральной совокупности. Но нам хотелось бы более точно определить границу, до которой еще возможно восстановить исходное распределение. Для этого мы должны найти критерий, позволяющий отличить ситуацию, в которой исходное распределение восстановлено, от той, в которой оно не восстановлено или восстановлено неверно. Если такой критерий будет статистическим, то его результат должен носить вероятностный характер, то есть он позволит нам рассчитать вероятность восстановления распределения. Мы можем задать порог, например 95%, ниже которого будем считать, что качество восстановления теоретического распределения нас уже не может устроить. Объем выборки, при котором последний раз достигается этот порог, будем считать минимальным.

Понятно, что минимальный объем выборки не есть фиксированное число, а зависит от самого теоретического распределения – масштаба a и b и характера их изменения. Так что, учитывая многомерность задачи, в данной работе мы сможем определить минимальный объем выборки только для распределений самых простейших видов.

Но, несмотря на невозможность заранее указать, какой минимальный объем выборки необходим в каждом конкретном случае, все же существует способ примерно оценить его по реальным данным. Этот способ состоит в том, что мы принимаем восстановленное распределение за теоретическое и для него запускаем процедуру определения минимального объема выборки. Таким образом нам удастся учесть масштаб и особенности изменения параметров распределения для конкретных данных. Правда, стоит заметить, что восстановленное распределение не должно быть равномерным, иначе не будет работать критерий восстановления распределения.

Указанный способ определения минимального объема выборки легко трансформируется в стратегию точного изучения распределения. Для этого достаточно применять его итерационно: получив на очередном шаге некоторое приближение к функции распределения, по нему определяем минимальный объем выборки, закладывая в критерий все более и более жесткий порог вероятности восстановления распределения. На каждом следующем шаге, увеличивая объем выборки до расчетного, мы получаем все более точные восстановленные распределения, достигая в пределе цель исследования в виде точной теоретической модели условного распределения.

Теперь перейдем к формулировке самого критерия восстановления распределения. Чтобы построить такой критерий, мы должны уметь сравнивать восстановленное распределение с теоретическим. Одна такая мера расстояния между распределениями нам уже известна (18). Значит, мы всегда в состоянии определить, к какому распределению ближе восстановленное – к теоретическому или к равномерному.

Далее, мы можем провести достаточное количество численных экспериментов с генерацией выборок заданного объема и определить процент случаев, в которых восстановленное распределение оказывается ближе к теоретическому, чем к равномерному. Этот процент случаев и будет оценкой искомой вероятности восстановления распределения по выборке заданного объема.

2. Построение обобщенной классификации

При структурном анализе среднепалеолитических индустрий Кавказа и Ближнего Востока [Деревянко, Холюшкин, Ростовцев, Воронин, 2002] возникла задача: по результатам автоматической классификации 64-х археологических памятников, проведенной заранее разными методами (k-средних, иерархического кластерного анализа и типологического анализа) и на разных признаковых пространствах, построить некоторую сводную, обобщенную классификацию.

Исходные данные представляют собой таблицу объект-свойство, где в качестве объектов выступают памятники, а в качестве их свойств – номера кластеров, к которым они были отнесены в результате каждой классификационной процедуры (см. табл. 5).

Вводится статистика для измерения степени близости объектов по результатам нескольких классификаций и предлагается алгоритм использования полученного показателя близости для выделения наиболее устойчиво совместно классифицирующихся объектов в ядра кластеров обобщенной классификации.

Измерение близости пары объектов.

Данные. Рассматриваются две реализации векторной случайной величины $C_i=\{c_i^1, c_i^2, \dots, c_i^m\}$ и $C_j=\{c_j^1, c_j^2, \dots, c_j^m\}$, где i и j – номера объектов в выборке, c_i^k - номер кластера по классификации k для i-го объекта, m - количество рассматриваемых классификаций. Размер всей выборки будем обозначать через N, а размер кластера c_i^k через $N_{c_i^k}$. Принадлежность i-го и j-го объектов к одному и тому же кластеру по классификации k обозначим через $c_{ij}^k = \delta(c_i^k, c_j^k)$, где $\delta(a,b)$ - дельта-функция Дирака, принимающая значение 0 или 1 в зависимости от совпадения или несовпадения аргументов, в нашем случае - номеров кластеров c_i^k и c_j^k:

$$\delta(c_i^k, c_j^k) = \begin{cases} 0, c_i^k \neq c_j^k \\ 1, c_i^k = c_j^k \end{cases}$$

Допущения. Классификация объекта j не зависит от классификации объекта i.

Нулевая гипотеза. Вероятность совпадения номера кластера c_j^k с номером кластера c_i^k равна вероятности для объекта j случайно занять одно из $N_{c_i^k} - 1$ мест соседей по кластеру объекта i:

$$H: P(c_j^k = c_i^k) = \frac{N_{c_i^k} - 1}{N - 1} \qquad (1)$$

Поскольку эта вероятность не зависит от j-го объекта, будем обозначать ее как p_i^k.

Метод. Рассмотрим нормированные отклонения z_{ij}^k от ожидаемого значения:

$$z_{ij}^k = \frac{\delta(c_i^k, c_j^k) - p_i^k}{\sqrt{p_i^k(1-p_i^k)}}$$

При совпадении и несовпадении номеров кластеров z_{ij}^k будет принимать такие значения:

$$z_{ij}^k = \begin{cases} -\sqrt{\dfrac{p_i^k}{1-p_i^k}}, c_i^k \neq c_j^k \\ +\sqrt{\dfrac{1-p_i^k}{p_i^k}}, c_i^k = c_j^k \end{cases}$$

Нетрудно убедиться, что математическое ожидание z_{ij}^k равно нулю, а дисперсия – единице:

$$Mz_{ij}^k = \frac{0 - p_i^k}{\sqrt{p_i^k(1-p_i^k)}} \cdot (1 - p_i^k) + \frac{1 - p_i^k}{\sqrt{p_i^k(1-p_i^k)}} \cdot p_i^k = 0$$

Таблица 5. Результаты классификации памятников Кавказа и Ближнего Востока.

	Памятник	Классификация									
		1	2	3	4	5	6	7	8	9	10
1	Амуд B4	3	1	2	4	1	1	1	1	1	1
2	Амуд B2	3	1	2	4	1	2	2	1	2	1
3	Кеу сл. 1 I	3	3	3	2	1	3	2	2	3	2
4	Кеу сл. II	3	3	3	2	1	3	2	2	3	3
5	Кеу сл III	3	3	3	2	1	3	2	2	3	1
6	Кеу сл V	3	3	3	2	1	3	2	2	3	1
7	Кзар-Акил XXVIA	5	3	2	3	2	4	2	3	4	1
8	Кзар-Акил XXVIB	5	3	3	2	1	3	2	3	3	1
9	Кзар-Акил XXVIIA	4	4	3	2	1	3	3	3	3	1
10	Кзар-Акил XXVIIB	5	3	1	2	3	3	2	3	3	1
11	Кзар-Акил XXVIII	4	4	1	5	1	3	2	2	3	1
12	Кзар-Акил XXVIII	4	4	3	5	1	3	2	2	3	1
13	Кунджи	2	2	2	3	2	4	2	4	4	2
14	Варвази A	2	2	2	3	2	4	2	3	4	2
15	Варвази B	2	2	2	3	2	4	2	4	4	2
16	Варвази C	2	2	2	3	2	4	2	4	4	2
17	Варвази D	2	2	2	3	2	4	3	4	4	2
18	Сефуним A	5	3	3	2	1	3	2	3	3	1
19	Сефуним 12	2	2	2	3	2	4	2	4	4	1
20	Сефуним 13	3	4	3	2	1	3	2	2	3	1
21	Сефуним VI	4	4	3	5	1	3	2	2	3	1
22	Сефуним VII	4	4	1	5	1	3	2	2	3	1
23	Сефуним B	4	4	1	5	1	3	3	2	3	1
24	Сефуним C	3	4	3	5	1	3	3	2	3	1
25	Ябруд 2	5	3	2	3	2	4	2	4	4	2
26	Ябруд 3	5	3	2	4	3	2	2	4	5	2
27	Ябруд 4	2	2	2	3	2	4	1	4	4	2
28	Ябруд 5	1	1	1	5	4	2	3	1	4	2
29	Ябруд 6	1	3	2	3	4	3	3	4	4	2
30	Ябруд 7	1	1	2	4	3	2	3	1	3	2
31	Ябруд 8	5	3	2	3	3	2	1	4	5	2
32	Ябруд 9	1	1	1	1	4	2	3	1	4	2
33	Ябруд 10	5	3	2	3	3	3	1	4	4	2
34	Кударо I 3a	3	4	3	4	1	3	2	2	3	1
35	Кударо I 3б	4	4	3	4	1	3	2	3	3	1
36	Кударо I 3в	3	4	2	4	1	3	2	3	3	1
37	Кударо I 4	4	4	3	2	1	3	2	2	3	1
38	Каркустакау	4	4	3	2	1	3	2	2	3	1
39	Тамарашени	4ы	4ы	1	5	1	3	2	2	3	1
40	Монашеская	4	4	3	2	1	3	2	2	3	1
41	Губский Навес	5	3	1	1	3	3	3	3	4	1
42	Малая Воронцовка	5	3	3	2	3	3	2	3	3	2
43	Таглар 2 сл.	3	4	3	2	1	3	2	2	3	1
44	Таглар 3 сл.	3	4	3	2	1	3	2	2	3	3
45	Таглар 4а	3	4	3	2	1	3	2	2	3	1
46	Таглар 4б	3	4	3	2	1	3	2	2	3	1
47	Таглар 5	3	4	3	2	1	3	2	2	3	3
48	Таглар 6	3	3	3	2	1	3	2	2	3	3
49	Ортвала-Клде I	4	4	1	1	3	3	3	3	4	1
50	Ортвала-Клде II	5	3	3	2	1	3	2	3	3	1
51	Ортвала-Клде III	5	3	3	2	1	3	2	3	3	1
52	Ортвала-Клде IV	5	3	2	3	1	3	2	3	3	2
53	Ортвала-Клде V	5	3	3	2	1	3	2	3	3	1
54	Ортвала-Клде VI	5	3	2	3	1	3	2	3	3	1
55	Ортвала-Клде VII	4	4	1	1	3	3	3	5	5	1
56	Двойной Грот	1	4	1	1	4	3	3	3	4	2
57	Азых 3 сл.	3	3	2	4	1	2	2	3	5	1
58	Среднехаджохская	4	4	3	2	1	3	2	2	3	1
59	Азых 6 сл.	2	2	2	3	2	4	1	4	4	2
60	Медвежье	1	4	1	1	5	3	3	5	5	1
61	Лусакерт D	4	4	1	5	3	3	3	3	4	1
62	Лусакерт A	4	4	1	5	3	3	3	1	4	1
63	Газма	5	3	2	3	1	3	1	4	4	1
64	Баракаевская	1	3	1	1	4	3	3	5	5	2

$$Dz_{ij}^k = \left(\frac{0-p_i^k}{\sqrt{p_i^k(1-p_i^k)}}\right)^2 \left(1-p_i^k\right) + \left(\frac{1-p_i^k}{\sqrt{p_i^k(1-p_i^k)}}\right)^2 p_i^k =$$

$$= \frac{\left(p_i^k\right)^2\left(1-p_i^k\right)+\left(1-p_i^k\right)^2 p_i^k}{p_i^k(1-p_i^k)} = p_i^k + (1-p_i^k) = 1$$

Классификации: метод k-средних по признакам: F1, F2 (1); F1, F3 (2); S_1^2, S_2^2 (3); S_1^3, S_2^3, S_3^3 (4); иерархический кластерный анализ по признакам: F1, F2 (5); F1, F3 (6); S_1^2, S_2^2 (7); S_1^3, S_2^3, S_3^3 (8); типологический анализ: группы (9); типы (10).

Отклонение совпадений от ожиданий по m классификациям для i и j объектов обозначим через Z_{ij}:

$$Z_{ij} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} z_{ij}^k$$

Вычислим матожидание Z_{ij}:

$$MZ_{ij} = M\left(\frac{1}{\sqrt{m}} \sum_{k=1}^{m} z_{ij}^k\right) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} M\left(z_{ij}^k\right) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} 0 = 0$$

Аналогично вычисляется и дисперсия Z_{ij}:

$$DZ_{ij} = D\left(\frac{1}{\sqrt{m}} \sum_{k=1}^{m} z_{ij}^k\right) = \frac{1}{m} \sum_{k=1}^{m} D\left(z_{ij}^k\right) = \frac{1}{m} \sum_{k=1}^{m} 1 = \frac{1}{m} m = 1$$

Но дисперсия суммы равна сумме дисперсий только в случае, если z_{ij}^k для разных k независимы. Будем считать, что это условие выполнено, несмотря на то, что сами c_i^k для разных k не являются независимыми.

Далее заметим, что Z_{ij} изменяется в пределах:

$$Z_i^{\min} \le Z_{ij} \le Z_i^{\max},$$

где $Z_i^{\min}$ определяется при полном несовпадении классификаций для i-го и j-го объектов:

$$Z_i^{\min} = -\frac{1}{\sqrt{m}} \sum_{k=1}^{m} \sqrt{\frac{p_i^k}{1-p_i^k}},$$

а $Z_i^{\max}$ – при полном совпадении:

$$Z_i^{\max} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \sqrt{\frac{1-p_i^k}{p_i^k}}$$

В частности, для всех диагональных элементов матрицы выполняется равенство:

$$Z_{ii} \equiv Z_i^{\max}$$

При выполнении нулевой гипотезы H (1) эмпирическое распределение, построенное на наборе $\{Z_{ij}\}$ для всех $i \neq j$, не должно отличаться от стандартного нормального распределения. Это дает возможность проверить гипотезу о наличии кластерной структуры еще до выполнения процедуры классификации. Принятие нулевой гипотезы H будет означать, что все N объектов в среднем расклассифицированы по m классификациям независимо друг от друга и потому искомая обобщенная классификация будет не более, чем случайным результатом эвристической процедуры.

Если же гипотеза H отвергается, мы можем переходить к построению искомой обобщенной классификации.

Поскольку построенная статистика несимметрична, то есть $Z_{ij} \neq Z_{ji}$, эту меру нельзя использовать для проведения автоматической классификации стандартными методами. Поэтому переходим к описанию процедуры классификации, не требующей симметричности.

Построение классификации.

Данные. Имеется заполненная таблица $\{Z_{ij}\}$:

$$\{Z_{ij}\}=\begin{bmatrix} Z_{11} & \cdots & Z_{1j} & \cdots & Z_{1N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Z_{i1} & \cdots & Z_{ij} & \cdots & Z_{iN} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Z_{N1} & \cdots & Z_{Nj} & \cdots & Z_{NN} \end{bmatrix} \quad (2)$$

Большие положительные значения Z_{ij} говорят о том, что объекты i и j оказываются в одном кластере существенно чаще ожидаемого в условиях нулевой гипотезы H.

Большие отрицательные значения Z_{ij} говорят о том, что объекты i и j оказываются в одном кластере существенно реже ожидаемого в условиях нулевой гипотезы H.

Кластерная структура. Кластер в матрице (2) представляется подмножеством строк (или столбцов с теми же номерами). Пересечение элементов этих строк и столбцов определяет блок матрицы. Удобно рассматривать матрицу после такой совместной перестановки строк и столбцов, которая собирает все клетки блока вместе. Тогда матрица приобретает блочно-диагональный вид (см. рис. 50). Внутри блоков, соответствующих кластерам, значения Z_{ij} должны быть существенно больше, чем за их пределами. В прямоугольных блоках на пересечении разных диагональных блоков, объединены показатели близости объектов, принадлежащих разным блокам, поэтому этот набор Z_{ij} характеризует степень контраста двух кластеров.

Допущения. Некоторое количество объектов может не входить ни в один из выделенных кластеров. Диагональные блоки не должны пересекаться, но могут перемежаться такими объектами, не объединенными в кластеры.

Критерий выделения кластеров. Зададим некоторое пороговое значение α. Будем считать, что объекты образуют кластер только в том случае, когда значимость гипотезы о равенстве нулю среднего значения Z_{ij} внутри блока отвергается с уровнем значимости α. Этот критерий позволяет учесть не только величину среднего значения Z_{ij}, но и объем кластера. Проверку гипотезы можем проводить по Т-критерию Стьюдента.

Метод. Предлагаемый метод относится к агломеративным методам кластерного анализа без обучения, использующих матрицу попарных взаимных близостей между объектами.

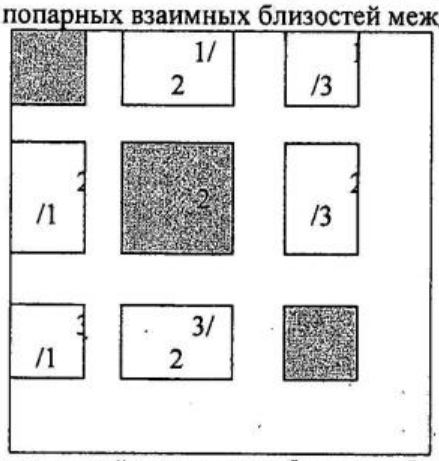


Рис. 50. Блочно-диагональный вид матрицы близости объектов.
1, 2, 3 – номера кластеров обобщенной классификации;
1/2, … 3/2 – блоки межкластерных связей.

Опишем процедуру выделения кластеров (возможно, пересекающихся).

На первом шаге переставляем строки и столбцы матрицы, располагая вместе наиболее близкие объекты. С этой задачей успешно справляется программа поиска структуры в таблицах сопряженности [Ростовцев, Костин, Корнюхин, Смирнова, 1994: 60-61]. В нашем случае получим результат, приведенный в таблице 4.

На втором шаге для всех возможных диагональных блоков вычисляем значимость гипотезы о равенстве нулю среднего значения Z_{ij} внутри блока. Если значимость меньше пороговой (α), присоединяем блок к уже найденным.

Увеличивая пороговое значение, мы можем получить ряд кластерных разбиений, дающих все более и более устойчивые ядра. Таким путем можно построить своеобразную "карту уровней", напоминающую рельеф горной местности на физических картах.

Результат построения такой "карты уровней" приведен в таблице 6 путем выделения клеток оттенками серого:

Таблица 6

Пороговое значение α	Цвет клетки
0.05	
0.01	
0.001	
0.0001	

3. Статистика для сравнения классификаций

При проведении автоматической классификации часто возникает вопрос о том, насколько выделенные программой классы отражают реальную структуру данных. Такая задача возникла и при структурном анализе среднепалеолитических индустрий Кавказа и Ближнего Востока [Деревянко, Холюшкин, Ростовцев, Воронин, 2002]. Для сравнительного анализа классификаций были отобраны по критерию полноты данных 64 археологических комплекса (см. табл. 7).

Таблица 7. Исходные данные: результаты разбиения на кластеры 64-х памятников методами k-средних в пространстве S^2 (1) и иерархического кластерного анализа в пространстве S^3 (2)

№	Памятник	Классификация	
		№1	№2
1	Амуд B4	2	1
2	Амуд B2	2	1
3	Кеу сл. 1 I	3	2
4	Кеу сл. II	3	2
5	Кеу сл III	3	2
6	Кеу сл V	3	2
7	Кзар-Акил XXVIA	2	3
8	Кзар-Акил XXVIB	3	3
9	Кзар-Акил XXVIIA	3	3
10	Кзар-Акил XXVIIB	1	3
11	Кзар-Акил XXVIII	1	2
12	Кзар-Акил XXVIII	3	2
13	Кунджи	2	4
14	Варвази A	2	3
15	Варвази B	2	4
16	Варвази C	2	4
17	Варвази D	2	4
18	Сефуним A	3	3
19	Сефуним 12	2	4
20	Сефуним 13	3	2
21	Сефуним VI	3	2
22	Сефуним VII	1	2
23	Сефуним B	1	2
24	Сефуним C	3	2
25	Ябруд 2	2	4
26	Ябруд 3	2	4
27	Ябруд 4	2	4
28	Ябруд 5	1	1
29	Ябруд 6	2	4
30	Ябруд 7	2	1
31	Ябруд 8	2	4
32	Ябруд 9	1	1
33	Ябруд 10	2	4
34	Кударо I 3a	3	2
35	Кударо I 3б	3	3
36	Кударо I 3в	2	3
37	Кударо I 4	3	2
38	Каркустакау	3	2
39	Тамарашени	1	2
40	Монашеская	3	2
41	Губский Навес	1	3
42	Малая Воронцовка	3	3
43	Таглар 2 сл.	3	2
44	Таглар 3 сл.	3	2
45	Таглар 4а	3	2
46	Таглар 4б	3	2
47	Таглар 5	3	2
48	Таглар 6	3	2
49	Ортвала-Клде I	1	3
50	Ортвала-Клде II	3	3
51	Ортвала-Клде III	3	3
52	Ортвала-Клде IV	2	3
53	Ортвала-Клде V	3	3
54	Ортвала-Клде VI	2	3
55	Ортвала-Клде VII	1	5
56	Двойной Грот	1	3
57	Азых 3 sl	2	3
58	Среднехаджохская	3	2
59	Азых 6 sl	2	4
60	Медвежье	1	5
61	Лусакерт D	1	3
62	Лусакерт A	1	1
63	Газма	2	4
64	Баракаевская	1	5

Примечание. Пространство S^2 – двумерное признаковое пространство, полученное процедурой многомерного шкалирования путем проекции точек из многомерного исходного пространства признаков на двумерную плоскость при максимальном сохранении взаимных расстояний между ними. Пространство S^3 – такое же отображение в трехмерное пространство.

Исследователю необходимо получить подтверждение того, что обнаруженная кластерная структура не является случайной флуктуацией.

Для этого требуется выйти за пределы того признакового пространства, в котором была проведена классификация.

Одним из способов убедиться в неслучайности найденной кластерной структуры является сравнение классификаций, построенных на разных признаковых пространствах.

Обратимся к постановке задачи. Для этого рассмотрим, как мы можем сравнить результаты двух классификаций.

В исходных данных, приведенных в таблице 7, первый столбец содержит порядковый номер объекта, второй – его название, третий – номер кластера по первой классификации (разбиение на 3 класса), четвертый – номер кластера по второй классификации (разбиение на 5 классов). Если подсчитать частоты встречаемости всех возможных пар, получим (таблица 8):

Таблица 8. Таблица сопряженности результатов двух классификаций

Классификация 2	Классификация 1			Итого
	1	2	3	
1	3	3		6
2	4		19	23
3	5	6	8	19
4		13		13
5	3			3
Итого	15	22	27	64

Этой таблицы вполне достаточно, чтобы оценить степень согласованности результатов первой и второй классификаций.

Наш план заключается в том, чтобы по приведенной таблице определить:

– степень согласованности классификаций;

– статистическую значимость полученной величины путем построения функции ее распределения в условиях нулевой гипотезы.

Определение степени согласованности классификаций

Сформулируем требования, которым должен удовлетворять искомый показатель степени согласованности классификаций.

Во-первых, он должен быть нечувствителен к порядку нумерации классов. Это требование вытекает из того, что процедура автоматической классификации выделяет классы объектов, не учитывая их содержательной характеристики, а опираясь исключительно на особенности взаимного расположения объектов как точек в многомерном признаковом пространстве. Поэтому номер класса является не более, чем условным идентификатором.

Во-вторых, наш показатель должен измерять степень согласованности даже при несовпадении количества классов в сравниваемых классификациях, поскольку иначе его практическое применение будет неоправданно ограничено.

В-третьих, он должен давать максимальное значение (например, 1) при сравнении классификации с собой.

В качестве ближайшего аналога рассмотрим индикатор æ (каппа), впервые предложенный Дж.Коэном в [Cohen J., 1960; Cohen J., 1968], и затем независимо – Г.Раушенбахом и А.Заславским [Раушенбах Г. В., Заславский А. А.: 126-141], и используемый для сравнения признаков, принимающих сопоставимые значения, например, результатов диагностики больных двумя врачами-экспертами.

Для его вычисления применяется формула:

$$æ = \frac{p_d - p_e}{1 - p_e} \qquad (1)$$

где p_d – сумма долей в диагональных клетках таблицы сопряженности;

p_e – сумма ожидаемых долей в тех же клетках в условиях независимости признаков.

Из приведенной формулы видно, что индикатор æ достигает максимального значения (равного единице), когда все недиагональные элементы равны нулю. Согласованность переменных считается слабой, когда значение æ не превышает 0,4, заметной или хорошей – при значениях 0,4-0,75, и высокой – при значениях более 0,75 [см. Landis, J.R. and Koch, G.G.; Флейс Дж.: 233].

Позаимствуем отсюда идею оценки степени согласованности суммой долей в диагональных клетках. Поскольку в нашем случае однозначное соответствие номеров кластеров разных классификаций заранее не установлено, мы вправе сами установить это соответствие из каких-либо соображений.

Для начала рассмотрим квадратную матрицу, в которой количества строк и столбцов совпадают. Тогда установление соответствия сводится к перестановке строк и столбцов матрицы, после которой

соответствующие друг другу строки и столбцы пересекаются на главной диагонали.

Когда же, как в рассматриваемом примере с трех- и пятикластерной классификациями, количества строк и столбцов различаются, лишние строки или столбцы присоединяем к соседним, то есть сопоставляем одному столбцу сразу несколько строк или одной строке несколько столбцов.

Устанавливать соответствие между номерами кластеров двух классификаций будем так, чтобы сумма частот на главной диагонали принимала максимально возможное для таблицы 9 значение. Для этой таблицы максимальная сумма 46 достигается при указанном в этой таблице порядке строк и столбцов.

Таблица 9. Таблица сопряженности после установления максимального соответствия.

Классификация 2	Классификация 1			Итого
	1	2	3	
2	19		4	23
3	8	6	5	19
4		13		13
5			3	3
1		3	3	6
Итого	27	22	15	64

Делением этой величины на 64 получаем статистику согласованности:

$p_d = 0{,}71875$.

Распределение согласованности в условиях нулевой гипотезы

Сформулируем нулевую гипотезу. Поскольку предлагаемая статистика предназначена для измерения связи, то нулевая гипотеза должна предполагать отсутствие этой связи. То есть, при выполнении нулевой гипотезы результаты каждой классификации остаются теми же, но связь между ними разрушена. Условия нулевой гипотезы гарантированно выполняются в экспериментах с перемешиванием данных, когда значения первого признака остаются на своих местах, а значения второго перемешиваются случайным образом. Алгоритмически перемешивание реализуется случайной выборкой без возвращения. Проведя серию из 100 или 1000 таких вычислительных экспериментов, мы можем получить эмпирическое распределение, близкое к теоретическому. Чем больше будет проведено экспериментов, тем ближе полученное распределение к теоретическому.

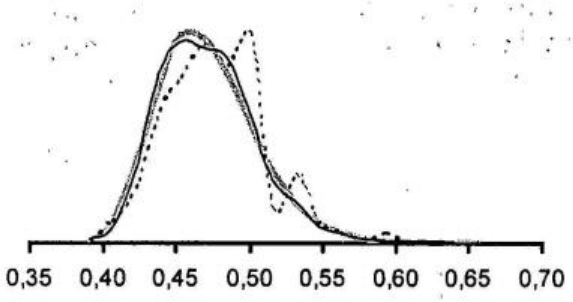


Рис. 51. Эмпирическая плотность распределения степени согласованности классификаций по результатам статистических экспериментов.
Условные обозначения: ····· 10^2 экспериментов
—— 10^3 экспериментов
▒▒▒ 10^5 экспериментов

Однако, такой подход недостаточен для оценки вероятности редких событий. А как раз маловероятные в условиях нулевой гипотезы события и являются практически интересными. Здесь может быть предложено два варианта:

– аппроксимировать полученное эмпирическое распределение каким-либо известным параметрическим распределением, например, нормальным или бета-распределением;

– рассчитать точное теоретическое распределение.

Аппроксимация эмпирического распределения

Для проверки качества аппроксимации нормальным распределением мы провели 10 млн. статистических экспериментов с перемешиванием данных. В результате было получено распределение, приведенное в таблице 10.

Полученное эмпирическое распределение с хорошей точностью совпадает с теоретическим, которое

будет рассчитано далее.

Таблица 10. Эмпирическое распределение степени согласованности классификаций по результатам 10^7 экспериментов

Сумма на диагонали	Процент соответствия	Число экспериментов	Наблюдаемая частота	Значимость
25	39.1	17 325	0.0017325	1.0000000
26	40.6	176 180	0.0176180	0.9982675
27	42.2	709 246	0.0709246	0.9806495
28	43.8	1 415 180	0.1415180	0.9097249
29	45.3	1 874 771	0.1874771	0.7682069
30	46.9	1 867 066	0.1867066	0.5807298
31	48.4	1 543 552	0.1543552	0.3940232
32	50.0	1 072 880	0.1072880	0.2396680
33	51.6	656 910	0.0656910	0.1323800
34	53.1	359 611	0.0359611	0.0666890
35	54.7	176 942	0.0176942	0.0307279
36	56.3	79 562	0.0079562	0.0130337
37	57.8	32 484	0.0032484	0.0050775
38	59.4	12 221	0.0012221	0.0018291
39	60.9	4 172	0.0004172	0.0006070
40	62.5	1 356	0.0001356	0.0001898
41	64.1	390	0.0000390	0.0000542
42	65.6	120	0.0000120	0.0000152
43	67.2	29	0.0000029	0.0000032
44	68.8	3	0.0000003	0.0000003

На рис. 52 приведены результаты аппроксимации, из которых видно, что нормальное и бета-распределение дают близкие друг к другу функции плотности распределения, но оба не воспроизводят особенности наблюдаемого распределения. Следовательно, оценка значимости с использованием нормального или бета-приближения даст значительную ошибку при любом числе экспериментов, но может быть использована в качестве грубого приближения, если расчет теоретического распределения по каким-либо причинам невозможен.

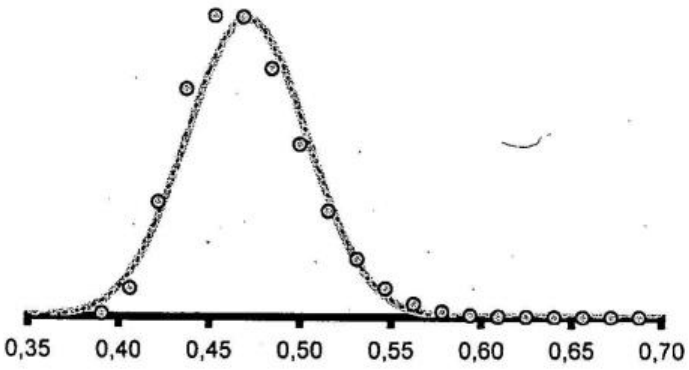


Рис. 52. Аппроксимация эмпирической плотности распределения.

Условные обозначения:

- эмпирическая плотность распределения по результатам 10^7 экспериментов
- аппроксимация Бета-распределением
- аппроксимация нормальным распределением

Для нашего примера получим:

$P_{Norm}(L \geq 46)$ " $3.03 \cdot 10^{-14}$,

$P_{Beta}(L \geq 46)$ " $2.26 \cdot 10^{-15}$.

Как мы увидим в дальнейшем, эти оценки значимости очень далеки от точного значения, полученного из теоретического распределения.

Тот факт, что нормальное распределение не является адекватной аппроксимацией, подтверждается статистически значимым отличием показателей асимметрии и эксцесса от нуля (см. табл. 11).

Таблица 11. Параметры аппроксимации эмпирической плотности распределения нормальным распределением

Число наблюдений	10 000 000
Среднее	0.470654±0.000010
Стандартное отклонение	0.033045
Асимметрия	0.535087±0.000775
Эксцесс	0.324155±0.001549

Построение теоретического распределения

Построение точного теоретического распределения возможно, но требует больших затрат машинного времени на вычисления. Все же для рассматриваемого примера время вычислений оказалось приемлемым (порядка 5 минут) и теоретическое распределение было построено.

Опишем алгоритм расчета. Он распадается на решение следующих задач:

- расчет вероятности (количества элементарных событий) для каждого варианта заполнения;
- полный перебор вариантов заполнения таблицы сопряженности с накоплением суммарной вероятности по значениям статистики согласованности.

Расчет вероятности (количества элементарных событий) для каждого варианта заполнения.

Обратимся теперь к расчету числа элементарных событий, соответствующих каждому варианту заполнения таблицы.

Прежде всего, подсчитаем общее количество элементарных событий при перемешивании данных. Поскольку первый элемент может быть выбран одним из N способов, второй – одним из $(N-1)$ способов и т.д., то общее количество получается равным $N!$, или хорошо известному в комбинаторике числу перестановок для N элементов.

Для наших данных:

$$N! = 64! \approx 1.27 \cdot 10^{89}$$

Далее для примера подсчитаем, сколькими способами мы можем разместить 64 шара 5 цветов по 3 ящикам так, чтобы в каждом ящике было заданное число шаров каждого цвета.

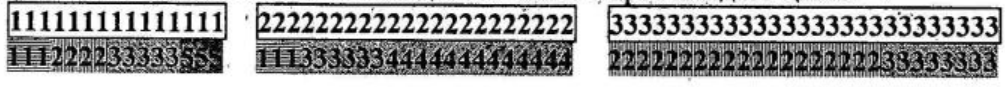


Рис. 53. Размещение шаров по ящикам.

Условные обозначения:

111, 222, 333 – ящики №1, 2, 3 соответственно;
111, 222, 333, 444, ■ – шары пяти различных цветов.

На рисунке 53 количество шаров соответствует наблюдаемым частотам из табл. 10.

Количество вариантов K, которыми можно разложить наши объекты по ящикам, вычисляется следующим образом:

$$K = (15! \cdot 22! \cdot 27!) \cdot \frac{6!}{3! \cdot 3! \cdot 0!} \cdot \frac{23!}{4! \cdot 0! \cdot 19!} \cdot \frac{19!}{5! \cdot 6! \cdot 8!} \cdot \frac{13!}{0! \cdot 13! \cdot 0!} \cdot \frac{3!}{3! \cdot 0! \cdot 0!} \qquad (2)$$

Здесь первая скобка $(15! \cdot 22! \cdot 27!)$ содержит произведение числа перестановок внутри каждого ящика, то есть полное число комбинаций для случая, когда каждый объект жестко привязан к своему ящику. В следующих пяти дробях вычисляется количество вариантов распределения объектов каждого из пяти сортов между тремя ящиками без учета порядка расположения объектов внутри ящиков, который уже учтен в первой скобке. Обобщая это выражение, можем написать:

$$K = \frac{\prod_{i=1}^{k} N_{i\bullet}! \cdot \prod_{j=1}^{m} N_{\bullet j}!}{\prod_{i=1}^{k} \prod_{j=1}^{m} n_{ij}} \qquad (3)$$

Для нашей таблицы:

$$K \approx 9.9 \cdot 10^{73}.$$

Вероятность реализации определенного варианта заполнения таблицы можно получить, разделив K на полное число элементарных событий, равное, как уже упоминалось, $N!$:

$$P_K = \frac{\prod_{i=1}^{k} N_{i\bullet}! \cdot \prod_{j=1}^{m} N_{\bullet j}!}{\prod_{i=1}^{k}\prod_{j=1}^{m} n_{ij} \cdot N!} \qquad (4)$$

Как можно заметить, мы получили гипергеометрическое распределение для таблицы произвольного размера. В нашем случае:

$$P_K \approx 7.8 \cdot 10^{-16}.$$

Полный перебор вариантов заполнения таблицы сопряженности.

Перебор вариантов заполнения может быть реализован в виде рекурсивной процедуры, которая перебирает в цикле все варианты заполнения одной клетки таблицы (см. табл. 12) и вызывает ту же процедуру для следующей клетки. Для определенности примем, что перебор начинается с левого верхнего угла таблицы, двигаясь вправо и вниз. Внутри процедуры требуется найти границы допустимых частот заполнения текущей клетки с учетом того, что частоты всех верхних клеток и левых клеток текущей строки уже заданы.

Рассмотрим клетку таблицы на пересечении строки i со столбцом j и найдем ограничения на частоту n_{ij}, накладываемые условиями постоянства итоговых частот. Для этого сгруппируем все строки и столбцы вне клетки ij.

Таблица 12. Параметры аппроксимации эмпирической плотности распределения нормальным распределением

Строки	Столбцы			Итого
	{1,j-1}	j	{j+1,m}	
{1,i-1}	$\sum_{p=1}^{i-1}\sum_{q=1}^{j-1} n_{pq}$	$\sum_{p=1}^{i-1} n_{pj}$	$\sum_{p=1}^{i-1}\sum_{q=j+1}^{m} n_{pq}$	$\sum_{p=1}^{i-1} N_{p\bullet}$
i	$\sum_{q=1}^{j-1} n_{iq}$	n_{ij}	$\sum_{q=j+1}^{m} n_{iq}$	$N_{i\bullet}$
{i+1,k}	$\sum_{p=i+1}^{k}\sum_{q=1}^{j-1} n_{pq}$	$\sum_{p=i+1}^{k} n_{pj}$	$\sum_{p=i+1}^{k}\sum_{q=j+1}^{m} n_{pq}$	$\sum_{p=i+1}^{k} N_{p\bullet}$
Итого	$\sum_{q=1}^{j-1} N_{\bullet q}$	$N_{\bullet j}$	$\sum_{q=j+1}^{m} N_{\bullet q}$	N

Чтобы найти ограничения, накладываемые на n_{ij}, запишем условия неотрицательности для каждой из четырех клеток, выделенных оттенком серого:

$$\begin{cases} n_{ij} \geq 0 \\ \sum_{p=i+1}^{k}\sum_{q=j+1}^{m} n_{pq} \geq 0 \\ \sum_{q=j+1}^{m} n_{iq} \geq 0 \\ \sum_{p=i+1}^{k} n_{pj} \geq 0 \end{cases} \qquad (5)$$

Теперь выразим неизвестные частоты через известные и n_{ij}:

$$\begin{cases} n_{ij} \geq 0 \\ (N - \sum_{p=1}^{i-1} N_{p\bullet} - \sum_{q=1}^{j-1} N_{\bullet q} + \sum_{p=1}^{i-1}\sum_{q=1}^{j-1} n_{pq}) - (N_{i\cdot} - \sum_{q=1}^{j-1} n_{iq}) - (N_{\cdot j} - \sum_{p=1}^{i-1} n_{pj}) + n_{ij} \geq 0 \\ N_{i\cdot} - \sum_{q=1}^{j-1} n_{iq} - n_{ij} \geq 0 \\ N_{\cdot j} - \sum_{p=1}^{i-1} n_{pj} - n_{ij} \geq 0 \end{cases} \quad (6)$$

Отсюда получаем систему неравенств, определяющих границы изменения n_{ij}:

$$\begin{cases} n_{ij} \geq 0 \\ n_{ij} \geq (N_{i\cdot} - \sum_{q=1}^{j-1} n_{iq}) + (N_{\cdot j} - \sum_{p=1}^{i-1} n_{pj}) - (N - \sum_{p=1}^{i-1} N_{p\bullet} - \sum_{q=1}^{j-1} N_{\bullet q} + \sum_{p=1}^{i-1}\sum_{q=1}^{j-1} n_{pq}) \\ n_{ij} \leq N_{\cdot j} - \sum_{p=1}^{i-1} n_{pj} \\ n_{ij} \leq N_{i\cdot} - \sum_{q=1}^{j-1} n_{iq} \end{cases} \quad (7)$$

Таблица 13. Распределение степени согласованности классификаций, полученное полным перебором вариантов заполнения таблицы сопряженности с использованием вероятностей (4) и ограничений (7)

Сумма на диагонали	Процент соответствия	Вероятность	Значимость	Вариантов заполнения
25	39.1	0.001749	1.000000	22
26	40.6	0.017560	0.998251	355
27	42.2	0.071013	0.980690	2258
28	43.8	0.141812	0.909677	8303
29	45.3	0.187544	0.767865	20943
30	46.9	0.186256	0.580321	41835
31	48.4	0.154406	0.394065	71618
32	50.0	0.107233	0.239659	107056
33	51.6	0.065733	0.132426	144617
34	53.1	0.035935	0.066693	182606
35	54.7	0.017707	0.030759	212644
36	56.3	0.007954	0.013052	237641
37	57.8	0.003262	0.005097	253386
38	59.4	0.001225	0.001835	258359
39	60.9	0.000423	0.000610	255557
40	62.5	0.000134	0.000187	242486
41	64.1	3.91E-05	5.31E-05	223009
42	65.6	1.05E-05	1.39E-05	198010
43	67.2	2.63E-06	3.39E-06	171144
44	68.8	6.06E-07	7.67E-07	143010
45	70.3	1.30E-07	1.61E-07	116589
46	71.9	2.57E-08	3.14E-08	91949
47	73.4	4.73E-09	5.69E-09	70683
48	75.0	8.10E-10	9.60E-10	52722
49	76.6	1.28E-10	1.50E-10	38563
50	78.1	1.88E-11	2.17E-11	27468
51	79.7	2.54E-12	2.89E-12	19250
52	81.3	3.14E-13	3.53E-13	13115
53	82.8	3.53E-14	3.93E-14	8734
54	84.4	3.60E-15	3.95E-15	5620
55	85.9	3.29E-16	3.57E-16	3541
56	87.5	2.66E-17	2.86E-17	2098
57	89.1	1.88E-18	2.00E-18	1219
58	90.6	1.14E-19	1.20E-19	623
59	92.2	5.72E-21	5.97E-21	293
60	93.8	2.39E-22	2.46E-22	108
61	95.3	6.77E-24	6.95E-24	32
62	96.9	1.79E-25	1.79E-25	8

Приведенных ограничений достаточно для реализации процедуры полного перебора частот заполнения в таблице сопряженности произвольного размера.

В результате расчета для нашей таблицы было получено распределение, приведенное в табл. 13.

Значимость нулевой гипотезы для нашего случая:

$$P(L \geq 46) = 3.14\text{E-}08$$

Исходя из "принципа практической невозможности маловероятных событий", гипотезу о независимости классификаций в нашем случае можно уверенно отвергнуть.

В ближайшем будущем предполагается применить предложенный в данной статье математический аппарат для анализа структур на многослойных памятниках Верхнего Енисея [Васильев, 2003].

Одним из возможных применений метода может быть получение обобщенной классификации по результатам сопоставления более, чем двух исходных.

4. Графическое представление и интерпретация статистических результатов.

Для наиболее адекватной интерпретации результатов статистических исследований требуется их наглядное графическое представление. С этой целью в проекте использовался VRML как средство визуализации результатов исследований в археологии (рис. 54).

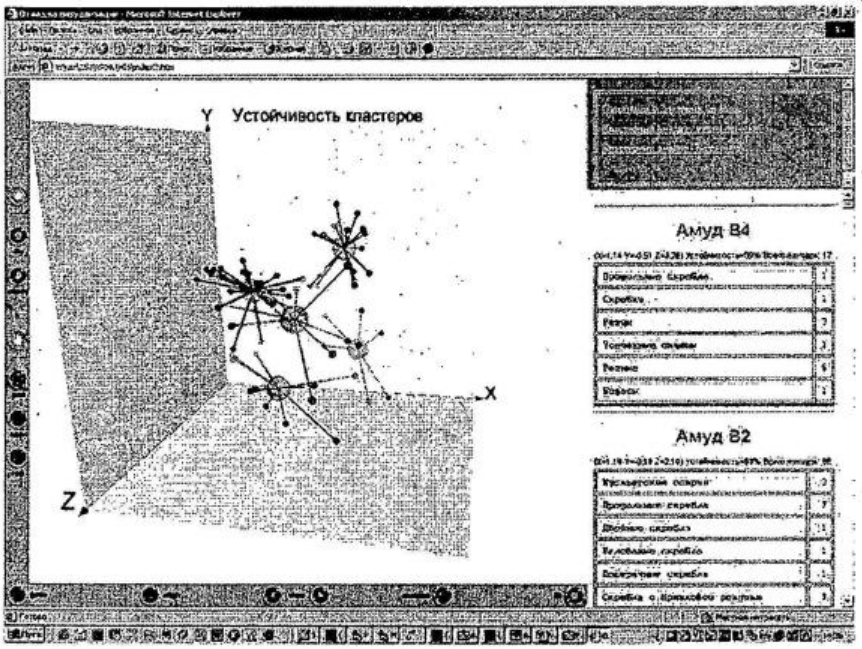


Рис. 53. Пример наглядного представления устойчивости кластеров в ходе экспериментов.

Среди средств представления результатов научных исследований большое распространение получили табличная форма, упорядочивающая числовые данные, графы (в том числе деревья) для демонстрации структуры взаимосвязей и различные виды диаграмм деловой графики для представления массивов данных – преимущественно в прямоугольной системе координат на плоскости. Язык моделирования виртуальной реальности (VRML) предоставляет новые возможности для развития подобных средств. Он позволяет реально вывести деловую графику в трехмерное пространство, давая читателю возможность зримо представить результаты математико-статистического анализа археологических данных в виде плавающих в 3-х мерном координатном кубе простых геометрических тел различного цвета и прозрачности, обозначающих объекты и их совокупности, связи между ними, границы областей, что дает более наглядную картину явлений и закономерностей. Возможность варьировать прозрачность тел открывает перспективы значительного увеличения информационной емкости такого представления.

Поскольку VRML изначально создавался для встраивания в HTML-документы, он идеально подходит для Web-публикаций, где пользователь также сможет исследовать модель в различных масштабах и с разных точек зрения, произвольно перемещаясь между ними. Удобно использовать

заложенные в VRML динамику (анимационные клипы) и интерактивность (гиперссылки или триггеры процессов, срабатывающие от нажатия на кнопку мыши или захвата и перемещения объектов). Дальнейшее развитие этого направления может привести к созданию научных публикаций нового качества, иллюстративный материал которых компактно размещается в виртуальном трехмерном пространстве на HTML-страницах.

Список изданий по теме интеграционного проекта СО РАН № 149 "Новые информационные технологии и математические методы в археологии, культурной и социальной антропологии", изданных на средства проекта

Монографии

Деревянко А.П., Холюшкин Ю.П., Ростовцев П.С., Воронин В.Т. Корреляция среднепалеолитических индустрий Ближнего Востока и Кавказа. Новосибирск: Изд-во СО РАН, 2002. 186 с.

Derevianko A.P., Kholuchkin Yu. P., Rostovtsev P.S., VoroninV.T. Corrélation des industries Paléolithique Moyen du Proche-Orient du Caucase. – Новосибирск, 2004. 116 с.

Сборники научных статей:

Информационные технологии в гуманитарных исследованиях. Вып. 5. – Новосибирск: РИЦ НГУ, 2003.

Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск: РИЦ НГУ, 2003.

Информационные технологии в гуманитарных исследованиях. Вып. 7. – Новосибирск: РИЦ НГУ, 2004.

Литература

Бакстон Эндрю, Хопкинсон Алан. Руководство по CDS/ISIS для Windows. Москва 2002.

Боровикова О.И., Загорулько Ю. А. Организация порталов знаний на основе онтологий. // Труды международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии" – Т. 2. – Протвино, 2002: 76-82.

Булгаков С.В. Подход к построению мультиагентной системы для проведения содержательного поиска во множестве информационных источников. // Труды VIII Междунар. конф. по электронным публикациям. – Новосибирск, 2003. 3 с. (Электронное издание, гос. регистр. 3521)., http://www.ict.nsc.ru/ws/show_abstract.dhtml?ru+76+5988

Булгаков С.В., Загорулько Ю.А., Костов Ю.В. Проект интеллектуального интернет-портала информационных ресурсов о научном и производственном потенциале региона // Труды V-й международной конференции "Проблемы управления и моделирования в сложных системах" – Самара: Самарский Научный Центр РАН, 2003: 255-260.

Воронин В.Т., Холюшкин Ю.П., Бердников Е. В., Федоров С.А., Жилицкая Г.Ю. Электронный каталог научной библиотеки института археологии и этнографии СО РАН //Информационные технологии в гуманитарных исследованиях. Вып. 6. Новосибирск: Редакционно-издательский отдел НГУ, 2003: 81-85.

Деревянко А. П., Холюшкин Ю.П., Бердников Е. В.Воронин В.Т. ГИС "Палеолит Северной Азии" // информационные технологии в гуманитарных исследованиях. Вып. 6. Новосибирск: Редакционно-издательский отдел НГУ, 2003: 21-29.

Деревянко А. П., Холюшкин Ю.П., Воронин В.Т., Костин В.С. Предварительные данные по структурному анализу технологических индексов мустьерских комплексов Средней Азии и Казахстана // Информационные технологии в гуманитарных исследованиях. Вып. 3. Новосибирск: Редакционно-издательский отдел НГУ, 2003: 46-56.

Жигалов В.А., Загорулько Ю. А., Нариньяни А.С., Россеева О.И. Предел однородности поиска в интернете // Системная информатика: Сборник научных трудов – Новосибирск: Наука, 2002. – Вып. 8: Теория и методология программирования: 29–71.

Загорулько Г.Б., Нариньяни А.С. Интеллектуальные таблицы: новые возможности в решении сложных задач // Материалы Международной научно-практической конференции "Информационные технологии, информационные измерительные системы и приборы в исследовании сельскохозяйственных процессов" Ч. 1. – РАСХН Сиб. отд-ние. – Новосибирск, 2003: 240-242.

Загорулько Ю.А., Боровикова О.И. Подход к разработке настраиваемого web-портала знаний // Материалы Международной научно-практической конференции "Информационные технологии, информационные измерительные системы и приборы в исследовании сельскохозяйственных процессов" Ч.1. – РАСХН Сиб. отд-ние. – Новосибирск, 2003: 235-240.

Загорулько Ю.А., Боровикова О.И. Подход к разработке настраиваемых порталов знаний.// Материалы III-й Международной научно-практической конференции "Математическое моделирование в образовании, науке и производстве". – Тирасполь: РИО ПГУ, 2003: 319-320.

Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Проблемы организации электронного архива с семантическим индексированием документов // Труды международной конференции Диалог'2003 "Компьютерная лингвистика и интеллектуальные технологии". – Протвино, 2003: 724–731.

Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Система InDoc: интеллектуальная обработка, распределение и поиск документов в электронном архиве // Труды V-й международной конференции "Проблемы управления и моделирования в сложных системах" – Самара: Самарский Научный Центр РАН, 2003: 248-254.

IS01R. Сплайн-сглаживание. БЧА НИВЦ МГУ. (Текст подпрограммы построения одномерного сглаживающего кубического сплайна на языках Фортран и С++).// http://www.srcc.msu.su/num_anal/lib_na/cat/i/is01r.htm

Кирсанов Д. Веб-дизайн: книга Дмитрия Кирсанова. – СПб.: Символ-Плюс, 2001. – 376 с.

Коржинский С. Н. Настольная книга Web-мастера: эффективное применение HTML, CSS и JavaScript. Издание второе, исправленное и дополненное. – М.: Издательский дом "КноРусь", 2000. – 320 с.

Костин В.С. Статистика для сравнения классификаций // Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск, Изд. НГУ, 2003: 57-65.

Костин В.С., Корнюхин Ю.Г. Построение обобщенной классификации // Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск, Изд. НГУ, 2003: 65-74.

Костин В.С., Нуртдинов А.Н., ЖдановА.С., Корнюхин Ю.Г. Бета регрессия как метод восстановления условного распределения случайной величины // Информационные технологии в гуманитарных исследованиях. Вып. 5. – Новосибирск, Изд. НГУ, 2003: 16-27.

Котеров Д. В. Самоучитель PHP 4. – СПб.: БХВ-Петербург, 2001. – 576 с.

Краткий обзор технологии j2ee и особенностей Веб-служб ("An overview of J2EE and Web Services Features") // htpp://www.oracle.com.

Марчук А.Г., Холюшкин Ю.П., Загорулько Ю.А., Воронин В.Т. Разработка новых методов и информационных технологий представления и обработки археологических и этнографических данных // Информационные технологии в гуманитарных исследованиях. Вып. 5. – Новосибирск, Изд. НГУ, 2003: 58-63.

Морозов В.А. О задаче дифференцирования и некоторых алгоритмах приближения ксперимента льной информации // Вычислительные методы и программирование. 1970. Вып.XIV. .:Изд-во МГУ: 46-62.

Нариньяни А.С. ТЕОН2: От тезауруса к онтологии и обратно // Труды международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии". – Т. 1. – Протвино, 2002: 307-313.

Носач В.В. Решение задач аппроксимации с помощью персональных компьютеров. М: МИКАП, 1994: 382 с.

Основные принципы JSP("JavaServer Pages Fundamentals") // htpp:// www.sun.com.

Ростовцев П.С. Бета-анализ иерархии социально-экономических объектов // Регион: экономика и социология. 2001. N4: 121-138.

Справочник CSS // htpp://www.manual.css.ru.

Справочник HTML // htpp://www.manual.html.ru.

Старыгин А. А. XML: разработка Web-приложений. – СПб.: БХВ-Петербург, 2003. – 592 с.

Холюшкин Ю. П. Место археологической историографии в системной классификации археологической науки // Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск, Изд. НГУ, 2003: 8-14.

Холюшкин Ю. П. О месте типологической археологии в системной классификации археологии // Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск, Изд. НГУ, 2003: 14-20.

Холюшкин Ю. П., Гемуев И.Н., Бауло А.В., Воронин В.Т., Нуртдинов А.Н. Религиозно-мифологические представления народов Западной Сибири // Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск, Изд. НГУ, 2003: 73-77.

Холюшкин Ю. П., Гемуев И.Н., Воронин В.Т. , Фурсова Е.Ф., Бауло А.В., Воробьев В.В., Грищенко А.А. Межэтнические взаимодействия и межконфессиональное согласие в Сибири (электронная библиотека как инструмент решения проблемы) // Информационные технологии в гуманитарных исследованиях. Вып. 6. – Новосибирск, Изд. НГУ, 2003: 77-81.

Холюшкин Ю.П., Гражданников Е.Д. Системная классификация археологической науки (элементарное введение в археологическое науковедение). Новосибирск: Изд-во ИДМИ Минобразования, Новосибирск, 2000. 58 с.

Benjamins V. R., Fensel D., et. all, 1998, "Community is Knowledge! in KA2", Proceedings of the KAW'98, Banff, Canada, 1998.

BIREME, ISIS_DLL Руководство пользователя. Sao Paulo, август 1997.

FAQ "Using XSQL servlet" // htpp://, источник www.oracle.com.

Genesereth, M.R. and Nilsson, N.J. Logical Foundation of Artificial Intelligence. Morgan Kaufmann, Los Altos, California, 1987.

Gruber Thomas R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing // International Workshop on Formal Ontology, March, Padova, Italy, 1993.

Guarino, N. 1997. Understanding, Building, and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga. International Journal of Human and Computer Studies(46): 293– 310.

Plog F.T. Laws, systems of law and the explanation of observed variation // The explanation of culture change: Models in prehistory – L., 1973: 649-661.

Takeda H., Takaai M., & Nishida T. Collaborative development and Use of Ontologies for Design // Proceedings of the Tenth International IFIP WG 5.2/5.3 Conference PROLAMAT 98, September 9 – 10 – 11, 12, Trento, Italy, 1998.

Ushold Mike, Gruninger Michael. Ontologies: Principles, Methods and Applications // Knowledge Engineering Review, Volume 11,Number 2., 1996.

Ushold Mike, King Martin. Towards a Methodology for Building Ontologies // IJCAI-95, Workshop on Basic Ontologica Issues in Knowledge Sharing, 1995.

Using Dublin Core. http://dublincore.org/documents/usageguide/

http://mapserver.gis.umn.edu/index.html

http://mapserver.gis.umn.edu/doc.html

http://www.easytrace.com/work/russian/news.html

http://www.xml.org/

Научное издание

Информационные технологии в гуманитарных исследованиях

Выпуск 8

Ответственный редактор:
академик. РАЕН, д.и.н. *Ю.П.Холюшкин*

Макет – *В.Т.Воронин,* дизайн – *А.Г.Микшин*
Компьютерная вёрстка – *В.Т.Ворони,*
Обложка – *Е.В.Бердников*

Подписано в печать 24.09.2004.
Заказ № 443

Формат 60x84 1/8. Офсетная печать.
Тираж 200 экз.
Учетно-издательских листов 8,5

Лицензия ЛР № 021285 от 6 мая 1998 г.

Редакционно-издательский центр НГУ 630090, Новосибирск 90, Пирогова, 2