Методология и технология создания многоцелевой информационной среды T-System на базе электронной библиотеки с гибким полнотекстовым поиском

С.Х.Ляпин, А.В.Куковякин

НП «Центроконцепт», ООО «Константа»
lyapin@atknet.ru, magicmagus@yandex.ru

Аннотация

Описана методология и технология построения многоцелевой информационной среды T-System путем расширения информационной системы T-Libra и предназначенной для интеграции ресурсов и сервисов, характерных для электронной библиотеки с гибким полнотекстовым поиском, виртуального музея, электронного архива, исследовательской лаборатории, образовательного сервера. Методологической основой интеграции является гибридная двухуровневая онтология, основанная на взаимодействии *функциональных систем* (верхний уровень), *библиотеки концептов* и *библиотеки тезаурусов* (нижний уровень). Технологической основой – унифицированная поисковая система, включающая в себя механизм нелинейных каскадных запросов, формирующих соответствующие функциональные системы и соединяющих результаты полнотекстового поиска, релевантные тезаурусы и концепты, текстовые метаданные, а также нетекстовые объекты различной модальности (графика, звук, видео и т.д.). Вся среда проектируется в трехзвенной архитектуре (Веб-браузер / Веб-сервер + Сервер приложений / Сервер баз данных), с использованием специальной системы индексации для повышения эффективности поиска, а также внешней логики, встроенной в сервер приложений и обеспечивающей совместимость с различными СУБД.

1 Введение: от T-Libra к T-System, от электронных библиотек к многоцелевым интегрированным информационным средам

1.1. Электронным библиотекам свойственно развиваться в направлении многофункциональных и мультимодальных информационных сред, с одновременным ростом концептуальности их поисковых и презентационных возможностей. Характерно и поучительно в этом плане, например, развитие американских проектов «Alexandria Digital Earth Prototype» (ADEPT) [1; 2; 3] и «Digital Library for Earth System Education» (DLESE) [4], ориентированных на комплексную поддержку обучения и исследований в сфере наук о Земле, или американо-германского проекта «Archimedes» [5], направленного на создание интерактивной среды по истории механики. Сказанное относится и к другим такого рода специализированным информационным системам (электронным архивам, виртуальным музеям и т.п.) [6; 7].

В этом же направлении развивается информационная система T-Libra, разработанная в ООО «Константа» и НП «Центроконцепт» (г. Архангельск) для создания многофункциональных электронных библиотек.

Она первоначально была спроектирована в двухзвенной Интернет-архитектуре (Web-browser / Web-server + SQL-server) с возможностями гибкого параметризируемого поиска по полнотекстовой базе данных при поддержке пополняемых электронных словарей и под управлением реляционной СУБД Sybase ASA v.7.0. (дополненной специально разработанной объектно-ориентированной инструментальной средой X-Taurus).

Прототип библиотеки был представлен на RCDL’02 в Дубне, действующие версии 5.1. и 5.2. – на конференции «Научный сервис в сети Интернет», Абрау-Дюрсо, сент. 2003 и на RCDL’2003, окт. 2003 [8].

Первый вариант бимодального расширения (текст + графический образ) был показан на EVA’2003, дальнейшее развитие в направлении многофункциональности, мультимодальности и интеллектуальности – на конференциях АДИТ-2004 в Самаре, июнь 2004 (музейная библиотека); Абрау-Дюрсо, сент. 2004 (виртуальная лаборатория) [9]; RCDL’2004, Пущино, сент. 2004 (концепт-ориентированный поиск) [10]; EVA'2004, дек. 2004 и АДИТ-2005 в Казани, июнь 2005 (интерактивные тематические экспозиции в бимодальной среде) [11; 12].

В настоящее время на платформе T-Libra создано несколько специализированных библиотек (для сфер медицины, образования и культуры), а

также ведется разработка регионального образовательного сервера для поддержки дополнительного профессионального образования специалистов культуры.

Архитектура создаваемых информационных систем проектируется с использованием различных вариантов СУБД (Sybase ASA, MS SQL Server, Oracle SQL Server, MySQL) и двух вариантов бизнес-логики приложений (внутренней логики, существенно использующей особенности этих СУБД и реализуемой на языке SQL, и внешней логики, встроенной в сервер приложений и реализуемой на объектно-ориентированном языке C++).

1.2. Ход этих разработок показывает, что методология и технология, реализованные при создании линии T-Libra, могут быть использованы в качестве платформы для функционального расширения и поэтапного создания интегрированной информационной среды, сочетающей в себе функции:

электронной библиотеки с универсальным настраиваемым каталогом, возможностями гибкого поиска по полнотекстовым ресурсам, существующим в формате SQL-базы данных, а также хранения, обработки и презентации нетекстовых единиц информации (собственно линия T-Libra, или T-System/Libra);

электронного архива, дополняющего электронную библиотеку возможностями полнотекстового поиска по описаниям документов (аннотациям, дайджестам, рефератам), а также создания тематических коллекций документов с навигацией по ним (линия T-System/Archive);

интерактивных тематизируемых экспозиций широкого назначения – в составе, например, виртуальных музеев, взаимодействующих в режиме онлайн с музейными библиотеками и системами автоматизированного учета музейных фондов (линия T-System/Media);

исследовательских систем, имеющих в своем составе виртуальные лаборатории как социально-гуманитарной, так и естественнонаучной и технической направленности (линия T-System/Research);

образовательных систем для поддержки автоматизированного дистанционного обучения, в которых система поиска и презентации его результатов адаптирована к учебным планам и рабочим программам (линия T-System/Education).

В целом можно говорить о семействе T-System, основанном на платформе T-Libra, использующем «текст» (в том числе текстовые метаданные) в качестве базисной модальности для функциональной интеграции всей мультимодальной (текстовой + нетекстовой) информации, а некоторую унифицированную настраиваемую поисковую систему по данным и метаданным, – в качестве универсального средства ее обработки.

1.3. В докладе рассматриваются возможности создания интегрированной среды на вышеуказанной платформе с учетом перехода к новой трехзвенной архитектуре T-Libra; обсуждается выбор адекватной онтологии для проектирования такой среды (с учетом ее дальнейшей поэтапной «интеллектуализации») и разработки соответствующей унифицированной поисковой системы; обосновываются преимущества «внешней» бизнес-логики, реализованной в составе сервера приложений, над «внутренней» логикой, реализованной средствами SQL-сервера баз данных; описывается система индексации, применяемая для существенного повышения эффективности поиска; демонстрируются некоторые функциональные блоки создаваемой интегрированной среды (T-Libra в новой архитектуре, T-Media и T-Research).

2 Гибридная двухуровневая онтология для расширения T-Libra и интеграции ресурсов и сервисов: функциональные системы (верхний уровень), библиотеки концептов и тезаурусов (нижний уровень)

2.1. Концептуальное и техническое проектирование многоцелевой интегрированной информационной среды, о которой идет речь, предполагает прежде всего выбор онтологической модели, адекватной назначению этой среды.

Эта модель должна учитывать не только возможности расширения функциональности в вышеупомянутых направлениях (п.1.2.), но и постепенную «интеллектуализацию» информационной среды, причем как с точки зрения организации первичных ресурсов, так и специальных поисковых и презентационных сервисов.

Мы предлагаем для этой цели *гибридную двухуровневую онтологию*, верхний уровень который обеспечивал бы функциональную целостность целенаправленного познавательного акта (осуществляемого «субъектом» – человеком-пользователем и/или программным агентом), а нижний – адекватную для целей продвинутого интеллектуального поиска организацию самих первичных ресурсов, «объектов» поискового запроса.

Поскольку именно верхний уровень модели определяет ее специфику, в целом все это можно было бы назвать *функционально-ориентированной онтологией*, соответственно – функциональной организацией «интеллекта» для многоцелевых информационных систем [13].

2.2. Основной единицей верхнего уровня предлагаемой онтологии является *функциональная система*, архитектоника которой может быть построена путем соответствующей интерпретации и адаптации теории функциональной системы академика П.К.Анохина, разработанной им сначала для описания и моделирования физиологии поведенческого акта, а затем распространенной на

широкий круг естественных и искусственных самоорганизующихся систем [14; 15].

Все ключевые компоненты анохинской функциональной системы – *результат* как доминирующий системообразующий фактор; *афферентный синтез*, предшествующий поведенческому акту, и *обратная афферентация* (обратная связь); *аппарат акцептора результата действия*, а также нелинейное *взаимосодействие* (термин П.К.Анохина) всех этих компонентов функциональной системы в составе целостного поведенческого акта – должны получить интерпретацию *в терминах информационной среды* (для начала в терминах пользовательского запроса к определенным образом организованным первичным ресурсам).

Этот уровень онтологии можно было бы назвать также «эпистемологическим»: конкретные функциональные системы, формируемые поисковыми запросами, фактически репрезентируют ту или иную функциональную организацию «субъекта» (реального пользователя и/или программного агента, формирующих запрос), и, соответственно, функциональную организацию субъект-объектного отношения.

В общеметодологическом плане (и тоже со ссылками на теорию П.К.Анохина) это сформулировано одним из нас еще в начале 1980-х гг. в рамках концептуальной модели «совокупного субъекта», или Σ-субъекта [16].

Для проектирования функциональной организации «субъекта» могут и должны быть использованы также различные варианты *нелинейной эпистемологии*, оперирующей нелинейными моделями субъект-объектного отношения в рамках соответствующих концепций сознания и познания.

Отметим в этой связи, в частности, предметно-редуктивную теорию сознания (теорию «идеального»), восходящую к К.Марксу, методологический потенциал которой далеко не исчерпан; всю экзистенциально-феноменологическую традицию, в особенности Э.Гуссерля и М.Хайдеггера; теорию сознания М.Мамардашвили.

Нелинейные модели субъект-объектного отношения являются ключевыми для моделирования пользователя (User Modeling) и описания взаимодействия «пользователь – информационная среда», в том числе для логического и технического проектирования структуры пользовательских запросов.

2.3. Нижний уровень онтологии представлен *библиотекой концептов* и *библиотекой тезаурусов*. При этом первая ориентирована на смысловую репрезентацию межпредметных и междисциплинарных мультимодальных смысловых единиц информации, – общекультурных «концептов» разного типа, вида и уровня (понимаемых в духе *концептологии* [17; 18; 19]), а вторая – на понятийно-терминологическую репрезентацию предметно-дисциплинарных областей знания в соответствии с традиционным построением тезауруса, ручным [20] или автоматизированным [21; 22; 23].

Отношения между «концептами» и «тезаурусами» (и соответствующими библиотеками) мы рассматриваем как достаточно свободные и взаимообратимые. Тот или иной концепт, представленный, например, терминологическим кластером с включенными нетекстовыми объектами, может оказаться связанным (общей логикой запроса, т.е. соответствующей функциональной системой, и собственной логикой концепта) с развертыванием и использованием некоторого специализированного тезауруса. Наоборот, специализированный локальный тезаурус (или фрагмент глобального тезауруса) может в ходе выполнения запроса включаться в качестве функциональной части в тот или иной концепт с присущей последнему инкапсулированной логикой.

Заметим, что этот нижний уровень (концепты + тезаурусы) нашей онтологической модели сам, в свою очередь, является представителем гибридной онтологии, – в смысле, например, [24].

2.4. В качестве методологических ориентиров для проектирования онтологии нижнего уровня мы предполагаем использовать следующие подходы.

2.4.1. Общепсихологическая модель понятийного концепта Л.М.Веккера [25]. В ней для наших целей особенно важен детальный операциональный и структурный анализ «понятия» как инварианта обратимого межъязыкового перевода (с языка свернутых симультанных образов на язык линейно упорядоченных речевых символов), ведущегося минимум на двух уровнях родо-видовой обобщенности, а также тезис о *понятийном концепте* как «центре кристаллизации» других понятийных концептов и тем самым – как интеллектообразующей единице.

Понятийный концепт (понимаемый по Веккеру) может служить ориентиром или даже образцом для конструирования культурного мультимодального концепта и соответствующих интеллектуальных структур более высокого (надличностного) уровня.

2.4.2. Вышеупомянутая концептология (‘концепты’ как многомерные мультимодальные единицы информации, предрасположенные к культурогенной трансляции смысла из одной предметной области в другую).

Многомерными мультимодальными концептами (ММ-концептами) в рамках развиваемой нами с начала 1990-х годов общей теории концептов, или концептологии мы называем *смысловые единицы информации* (СЕИ), существующие в широком культурно-историческом пространстве, но опирающиеся на понятийный (или псевдо-, или пред-понятийный) базис, закрепленный в значении какого-либо наблюдаемого культурного знака: научного термина, или слова (словосочетания) обыденного языка, или более сложной

семантической структуры (например, нелинейного терминологического кластера), или невербального предметного (квазипредметного) образа, или предметного (квазипредметного) действия и т.д.

Мы употребляем здесь выражение «ММ-концепт», чтобы одновременно подчеркнуть и близость этого термина к общеупотребительному термину «концепт», и его отличие, связанное с принципиальной многомерностью и, соответственно, мультимодальностью обозначаемой этим выражением смысловой единицы информации.

2.4.3. Направление развития теории нечетких множеств, связанное с именем Лотфи Заде (как и вся теория в целом), и ориентированное на вычисления со словами» (Computing with Words) и моделирование человеческих процессов познания и принятия решений на основе «гранулирования информации» (Information Granulation) [26; 27].

Вычисления со словами. Слова рассматриваются как ограничения на *лингвистическую переменную*, и основной компонентой процесса вычисления со словами является распространение ограничений с одних переменных на другие. Решение задачи рассматривается как распространение ограничений от множества исходных данных на множество заключительных данных, задаваемых совокупностью предложений естественного языка. Переход от предложений естественного языка к процессу распространения ограничений и обратный переход к предложениям естественного языка состоит, соответственно, из этапов формализации (разъяснения) ограничений, заданных на естественном языке, и ретрансляции этих ограничений (возможностных, вероятностных, нечетких, истинностных, функциональных и др.) на естественный язык [28].

Этот подход может быть распространен, с нашей точки зрения, на теоретическую и технологическую разработку «вычислений с концептами» *(Computing with Concepts)*, – когда операции производятся не с отдельными словами и предложениями (и соответствующими ограничениями), а с концептами как гетерогенными терминологическими кластерами с инкапсулированной логикой поведения в культурном поле. С соответствующими изменениями можно говорить, видимо, и о «вычислениях с тезаурусами» *(Computing with Thesauri)*.

Гранулирование информации. Слово естественного языка понимается как «метка» гранулы, а «гранула» – как группа объектов, обозначаемых словом и объединяемых неразличимостью, сходством, близостью или функциональностью. Как гранулы (НОС, УХО, ДОМ, ПРОЦЕСС и т.д.), так и их атрибуты (ДЛИНА, ЦВЕТ, ВРЕМЯ) и значения атрибутов (БОЛЬШОЙ, КРАСНЫЙ, БЫСТРЫЙ) в общем случае являются нечеткими. Язык, на котором происходит подобное описание предметной области, называется уточненным естественным языком [28].

Организацию знания в виде «концептов» и «тезаурусов» можно понимать как частный (но для наших целей принципиально важный) случай гранулирования. Соответственно можно использовать язык описания гранулированного знания (например, тот или иной вариант нечеткой логики, дескриптивной логики) для описания концепт-организованного знания и/или предметно-ориентированных тезаурусов.

2.4.4. Нейроинформатика, рассматриваемая как с точки зрения применения нейросетевой парадигмы для организации информационных ресурсов (прежде всего мультимодальных концептов и их совокупностей), а нейросетевых алгоритмов для организации поиска (включая распознавание смысла), так и с точки зрения эвристического моделирования информационной среды в целом. Здесь особенно важен и интересен характерный для нейроинформатики коннекционизм, – то есть рассмотрение *связей* структуры в качестве хранителей и носителей ее основных свойств [29].

2.4. Одним из ключевых моментов технологической реализации вышеописанной онтологии является унифицированная поисковая система, включающая в себя механизм *нелинейных каскадных запросов*, формирующих соответствующие функциональные системы и соединяющих результаты полнотекстового поиска, релевантные тезаурусы и концепты (со встроенной в них логикой развертывания знания), текстовые метаданные, а также нетекстовые объекты различной модальности (графика, звук, видео). (См. подробнее в [10]).

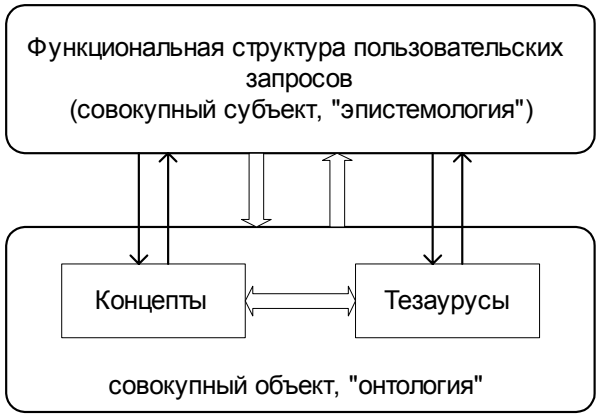


Рис. 1. Гибридная двухуровневая онтологическая модель для T-System

3 Новая T-Libra как база для семейства T-System: трехзвенная архитектура, внешняя логика, система индексации, язык разработки, объектность

3.1. T-Libra проектируется и развивается как информационная система для создания многофункциональных электронных библиотек, ориентированных на поддержку самых разных сфер

деятельности. Вместе с тем она рассматривается нами, – и концептуально, и технологически, – как база для создания на ее основе специализированных функциональных модулей, в совокупности составляющих семейство T-System (см. п.1.2.). Начиная с версии 6.x она обладает рядом существенно новых свойств, кратко обозначенных ниже.

3.2. Для обеспечения многоплатформенности в плане независимости от РСУБД, а также для минимизации затрат пользователя на внедрение T-Libra (и всего семейства T-System), осуществлен переход на новую архитектуру: выделен как отдельно живущая сущность сервер приложений (ранее он существовал "внутри" СУБД в виде набора хранимых процедур). Это позволило перенести в него всю бизнес-логику из базы данных, снимая главное требование к нижнему звену архитектуры: поддержке хранимых процедур. При этом достигается также независимость от конкретного диалекта SQL. В результате система может быть построена и на простых СУБД, типа MySQL. Связь сервера приложений с СУБД осуществляется через ODBC.

3.3. Для полнотекстового поиска применяется новая структура инвертированного индекса, хранящегося в файлах, и другие алгоритмы работы с ним. Вместо применявшегося ранее хранения индекса в таблицах СУБД и обработки его командами SQL используются процедуры, написанные на C++ (сам индекс хранится в файлах). В результате существенно повышена скорость поиска (до двух порядков).

Отличительной особенностью используемого индекса является его по-документная организация, – в отличие от построения глобального индекса по всему документному пространству, характерному, например, для поисковых машин Интернет. Это определяется общим назначением информационной системы рассматриваемого типа, и связано с наличием предварительного отбора документов, проводимого пользователем по метаданным (библиографическим, иконографическим и т.д. описаниям), для последующего полнотекстового поиска. В пределах документа индекс разделен на 5 частей (по частям речи), загружаемых в память независимо.

В состав инвертированного индекса в настоящее время входит номер абзаца в документе и позиция слова в абзаце. Операции с инвертированным индексом написаны на C++ с существенным использованием стандартной библиотеки (основным образом, map).

В дальнейшем предполагается переход на более специализированные алгоритмы построения и обработки индекса, в том числе с использованием сигнатурного подхода (см., например, [30; 31]) для повышения эффективности как индексирования, так и собственно поиска.

Мы планируем также использовать различные сочетания алгоритмов инвертированного и прямого поиска (см. в этой связи [32]), используя индексацию для повышения общей скорости поиска при реализации каскадных запросов, а прямой поиск по первичным ресурсам – для экспликации смысловых единиц информации разного типа, вида и уровня.

3.4. В отличие от предыдущих версий, использовавших процедурный подъязык SQL, высокоуровневые предметно-ориентированные объекты приложения разработаны на объектно-ориентированном языке C++.

От ранее существовавшего инструментального слоя X-Taurus, предназначенного для создания объектности над языком разработки SQL, осталась поддержка "точки входа" из необъектного CGI-расширения Веб-сервера в объектную среду приложения.

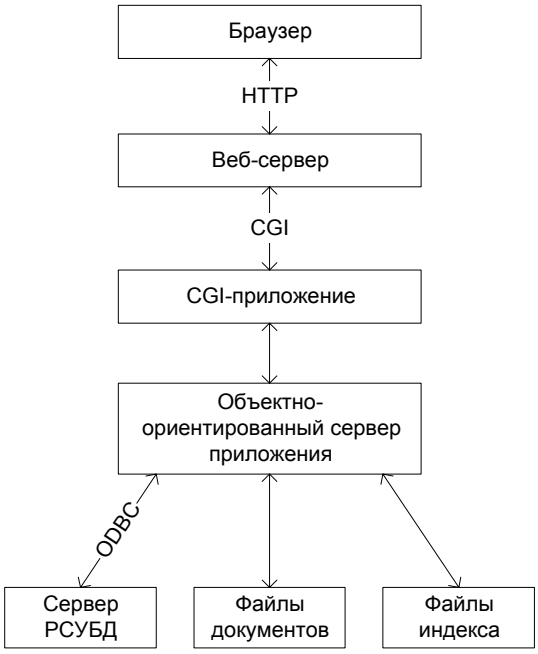


Рис. 2. Новая трехзвенная архитектура T-Libra

4 Основные модификации семейства T-System

4.1 T-Archive (T-System/Archive).

Специфику электронного архива (дополнительно к базовым возможностям T-Libra) определяют:

а) каталог с дополнительными настраиваемыми полями в административном и пользовательском разделах;

б) депозитарий со специализированным настраиваемым рубрикатором;

в) полнотекстовый поиск по аннотациям / рефератам / дайджестам архивных документов (информационные представления которых, как правило, существуют не в символьном, а в графическом формате);

г) модуль для создания пополняемых тематических коллекций архивных документов с возможностями просмотра и навигации по предметам коллекции.

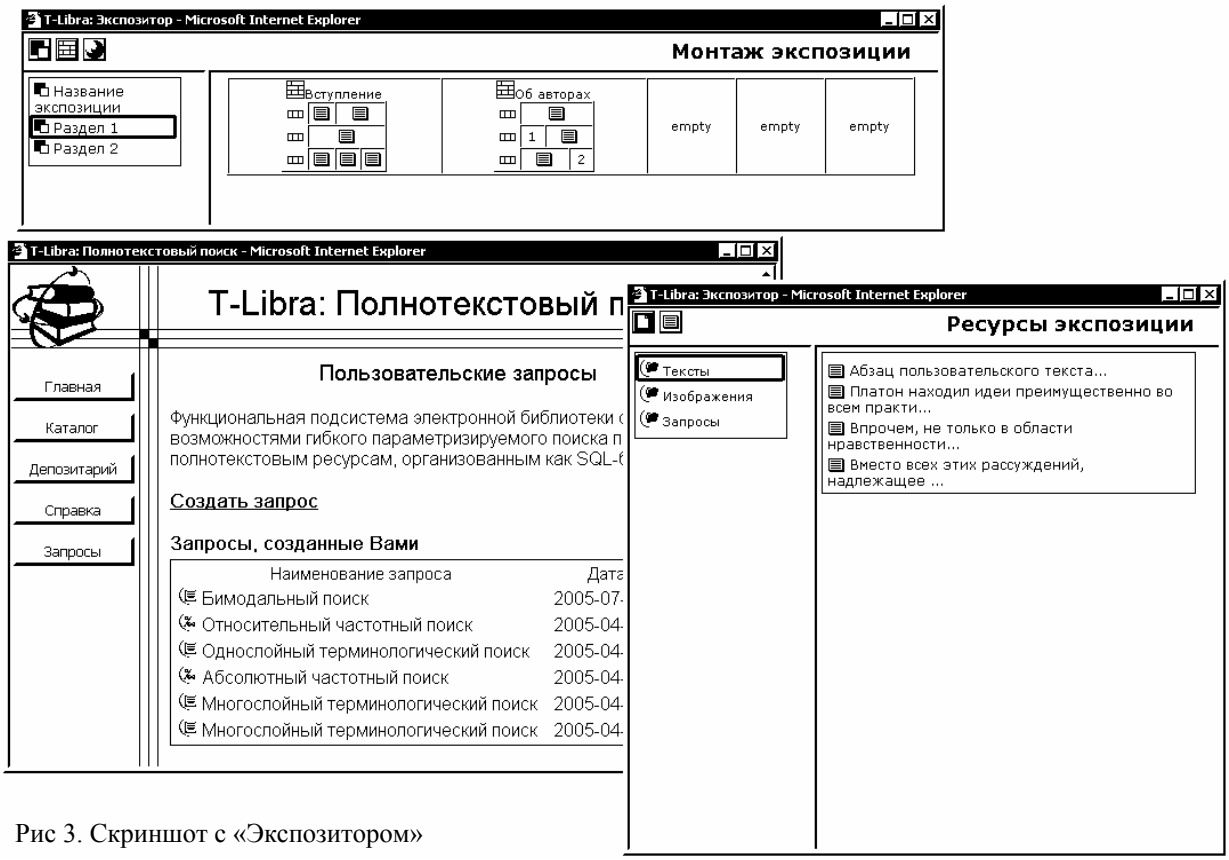


Рис 3. Скриншот с «Экспозитором»

4.2. T-Media (T-System/Media).

Это – информационная система для создания интерактивных тематических экспозиций и многоплановой поддержки виртуальных музеев.

4.2.1. На платформе T-Libra в настоящее время создается информационная система T-Media, позволяющая объединять полнотекстовые и нетекстовые электронные ресурсы (например, ресурсы музейных библиотек и музейных фондов) и формировать виртуальные интерактивные тематические экспозиции: как вручную, так и в автоматизированном режиме.

4.2.2. Осуществить вручную создание такого рода экспозиции можно в специализированном функциональном разделе T-Media – модуле Expositor.

При запуске этого модуля открываются три окна:

окно «Полнотекстовый поиск» (то есть соответствующий функциональный блок пользовательского раздела T-Libra),

и две опции собственно модуля Expositor, окно «Ресурсы экспозиции» и окно «Монтаж экспозиции».

Окно «Ресурсы экспозиции» позволяет собрать корзину ресурсов различного рода из подсистем, имеющихся в T-Media: абзацы (результаты запросов) и *сами запросы* (ссылки на них) из пользовательского раздела «Полнотекстовый поиск», файловые ресурсы из Депозитария, карточки из Каталога и т.д., а также вручную создать новые ресурсы (например, тексты-комментарии автора экспозиции в специальном текстовом окне).

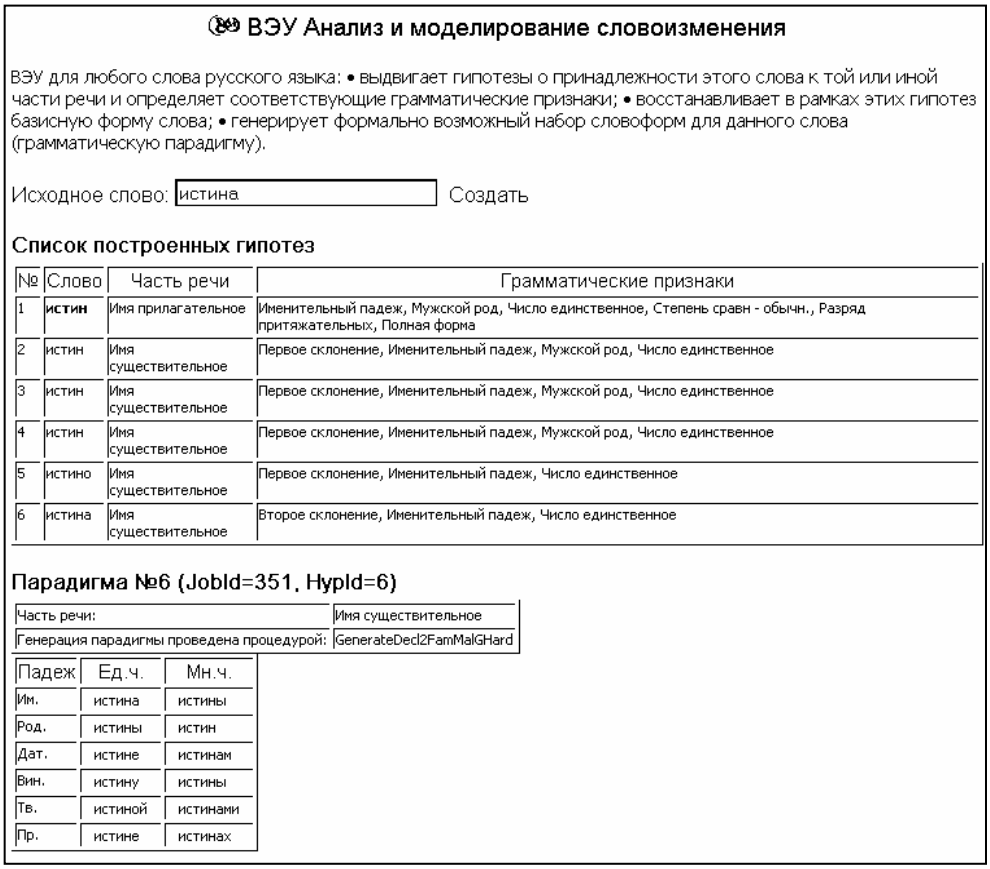
Окно «Монтаж экспозиции» обеспечивает создание оглавления экспозиции, редактирование html-страниц экспозиции путем сборки объектов из окна «Ресурсы экспозиции», дизайнерское и структурное оформление созданных страниц, их отделение от онлайновой связи с сервером и запись на переносимые носители.

Из готовой экспозиции (пока она остается в режиме онлайн) возможен прямой выход в поисковую систему T-Libra.

4.2.3. Тематическая экспозиция может быть составлена не только «ручным» путем, но и в автоматизированном режиме.

Исходной точкой ее построения является либо запуск запроса к полнотекстовой SQL-базе данных, позволяющего эксплицировать смысловой микроконтекст в пределах произвольного авторского абзаца, либо запуск запроса к файловой БД «Депозитарий», позволяющего эксплицировать тот или иной нетекстовый объект («экспонат») и связанное с ним описание (текстовые метаданные).

Результат этого промежуточного запроса – та или иная терминологическая структура (найденная в авторском абзаце или в текстовых метаданных нетекстового объекта), – становится, в свою очередь, началом нелинейного каскадного запроса, выполняющегося в мультимодальной информационной среде с активным использованием других текстовых метаданных, через которые осуществляется выход на релевантные нетекстовые объекты.



ВЭУ Анализ и моделирование словоизменения

ВЭУ для любого слова русского языка: • выдвигает гипотезы о принадлежности этого слова к той или иной части речи и определяет соответствующие грамматические признаки; • восстанавливает в рамках этих гипотез базисную форму слова; • генерирует формально возможный набор словоформ для данного слова (грамматическую парадигму).

Исходное слово: истина Создать

Список построенных гипотез

№	Слово	Часть речи	Грамматические признаки
1	**истин**	Имя прилагательное	Именительный падеж, Мужской род, Число единственное, Степень сравн - обычн., Разряд притяжательных, Полная форма
2	истин	Имя существительное	Первое склонение, Именительный падеж, Мужской род, Число единственное
3	истин	Имя существительное	Первое склонение, Именительный падеж, Мужской род, Число единственное
4	истин	Имя существительное	Первое склонение, Именительный падеж, Мужской род, Число единственное
5	истино	Имя существительное	Первое склонение, Именительный падеж, Число единственное
6	истина	Имя существительное	Второе склонение, Именительный падеж, Число единственное

Парадигма №6 (JobId=351, HypId=6)

Часть речи:	Имя существительное
Генерация парадигмы проведена процедурой:	GenerateDecl2FamMalGHard

Падеж	Ед.ч.	Мн.ч.
Им.	истина	истины
Род.	истины	истин
Дат.	истине	истинам
Вин.	истину	истины
Тв.	истиной	истинами
Пр.	истине	истинах

Рис. 4. Скриншот ВЭУ «Анализ и моделирование словоизменения»

Итоговым результатом запроса является совокупность мультимодальных культурных концептов, представленная в виде констелляций релевантных «текстов» и нетекстовых единиц информации, связанных между собой («по смыслу») теми или иными терминологическими кластерами.

4.3. T-Research (T-System/Research).

Информационная система для поддержки исследовательской деятельности и создания виртуальных экспериментальных установок.

4.3.1. Для широкого спектра гуманитарных исследований (лингвистика, история, философия, культурология, юриспруденция, социология, экономика и т.д.), в том числе междисциплинарных, принципиально важной является возможность продвинутой компьютерной обработки «текста», – различных по содержанию полнотекстовых ресурсов, существующих в виде SQL-базы данных.

Эта возможность реализуется в информационной системе T-Research с помощью *виртуальных экспериментальных установок* (ВЭУ), представляющих собой программные комплексы, каждый из которых предназначен для решения исследовательских задач определенного типа, снабжен интерфейсом для организации взаимодействия с полнотекстовыми (и мультимодальными) базами данных, входящими в состав T-Research, с результатами исследований, полученными на других установках, имеет специализированную подсистему презентации результатов поиска.

В настоящее время в составе T-Research имеются четыре таких ВЭУ.

4.3.2. Пример (ВЭУ «Анализ и моделирование словоизменения»).

Программный агент для любого слова русского языка, введенного исследователем:

а) выдвигает гипотезу о принадлежности этого слова к части речи и определяет соответствующие грамматические признаки;

б) восстанавливает в рамках этой гипотезы базисную форму слова сообразно его грамматической сущности;

в) генерирует формально возможный набор словоформ для данного слова, т.е. его грамматическую парадигму;

г) выбирает из всех сгенерированных таким образом парадигм правильную;

д) показывает «правильную» парадигму в контексте всех «неправильных» (см. Рис.4).

Эта ВЭУ может использоваться для автоматизированного пополнения электронных словарей словоформ (именно в этом качестве она используется в ИС T-Libra, входя в состав программного обеспечения для оператора полнотекстовой базы данных); для разработки экспериментальных, в том числе количественных методов анализа процесса словообразования; для освещения некоторых вопросов истории

естественного языка с точки зрения потенциально возможного и актуально реализуемого словаря; для анализа выразительного потенциала языковых форм, в том числе в связи с задачами поэтики; для новой постановки задач компаративной лингвистики

4.3.3. Другие ВЭУ в составе T-Research:

«Анализ и моделирование микроконтекста» (позволяет выявлять и анализировать *терминологические поля* в пределах предложения или авторского абзаца);

«Анализ и моделирование макроконтекста» (позволяет выявлять и анализировать на основе частотно-ранжированного поиска специфические терминологические поля – «терминограммы», или «вертикальные контексты» – в пределах произведения или их совокупности);

«Экспликация культурных концептов» (на основе нелинейного каскадного запроса с активным использованием метаданных позволяет эксплицировать и репрезентировать мультимодальный .культурный концепт, в итоге представленный в виде констелляции релевантных «текстов» и нетекстовых единиц информации, связанных между собой («по смыслу») тем или иным терминологическим кластером).

4.3.4. Развитие этого функционального модуля связано с разработкой ВЭУ не только для социально-гуманитарных, но и для естественных и технических наук.

При этом мы полагаем, что виртуальный эксперимент (как учебный, так и научный) должен иметь дело с *информационными моделями*, которые репрезентирует ту или иную реальность, а не с фрагментами самой реальности и экспериментальными установками (приборами) на другом конце Интернета. Этот подход обоснован в ряде публикаций по созданию различных виртуальных лабораторных практикумов (см., например, [32]).

Мы планируем также разработку специализированной инструментальной среды для конструирования различных ВЭУ администратором T-Research.

4.4. T-Education (T-System/Education).

Этот функциональный блок рассматривается нами как содержательная и технологическая основа образовательного сервера, используемого в рамках как традиционных, так и дистанционных форм обучения.

Он обладает всеми основными возможностями рассмотренных выше блоков (библиотеки, архива, музейно-экспозиционной и исследовательской среды), плюс некоторыми дополнительными.

В частности, проектируется и планируется:

а) специализированный интерактивный интерфейс, позволяющий администратору / пользователю собрать учебную тему вручную с использованием запросов по полнотекстовым и мультимодальным ресурсам, а также по базам данных со вспомогательным материалом (тестами, справочниками, иллюстрациями);

б) адаптация модуля Expositor к задачам формирования учебных тематических экспозиций (содержательную структуру лекции или семинара можно репрезентировать и с помощью набора соответствующих тематических экспозиций);

в) возможность выбора темы из списка готовых хорошо структурированных тем, который (список) может пополняться пользователем;

г) возможность использовать специальный инструментарий (программные агенты, экспертные системы) для формирования учебной, учебно-методической, научно-методической темы;

д) возможность использовать различные исследовательские ВЭУ, адаптированные к содержательной структуре учебных планов и рабочих программ.

5 Заключение

Гибридная двухуровневая онтология и унифицированная поисковая система будут использоваться в функциональных блоках T-System (обозначенных в п. 1.2.) в различных редуцированных формах, с различной степенью функциональной полноты и «интеллектуальности».

Эти функциональные блоки находятся сейчас в разных стадиях готовности:

T-Libra существует в коммерческом варианте, причем в нескольких версиях;

T-Media, T-Research, T-Archive – в виде отдельных работающих функциональных фрагментов;

T-Education – на стадии концептуального и технического проектирования.

Внедрение и апробация этих функциональных блоков и в целом многоцелевой информационной среды T-System будет происходить в рамках различных двух- и многосторонних проектов, а также на базе Архангельского областного центра повышения квалификации специалистов культуры (АОЦПК) в ходе создания региональной информационной инфраструктуры сферы культуры в течение 2005-2009 гг.

Литература

[1] *Terry Smith, Alex Ushakov, Bill Heller*. Some Aspects of Developing and Using the Digital Learning Environment in Alexandria Digital Earth Prototype // Proc. of the 5th National Russian Research Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections” RCDL’2003, St.-Petersburg, Russia, 2003, p.18-25, http://rcdl2003.spbu.ru/proceedings/C1.pdf.

[2] *O.Agapova, R.Mayer, T.Smith, A.S.Ushakov, A.A.Ushakov, Stefan Decker*. Developing Digital Library Visual Services for Building a Lesson-Design Environment Prototype // Proc. of the 5th National Russian Research Conference

RCDL'2003, St.-Petersburg, Russia, 2003, p.130-139, http://rcdl2003.spbu.ru/proceedings/C3.pdf..

[3] *Ushakov A.S., Agapova O., Smith T., Gerber M., Ushakov A.A*. Exploring the Semantic Types of Relationships for Visual Query Development // Proc. of the 6th National Russian Research Conference RCDL'2004, Pushchino, Russia, Sept.19 – Oct. 1, 2004, p.141-149, http://www.impb.ru/~rcdl2004

[4] *Mary Marlino, Tamara Sumner*. A Model and Research Agenda for Educational Community-based Digital Libraries: The Digital Library for Earth System Education // Proc. of the 5th National Russian Research Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" RCDL'2003, St.-Petersburg, Russia, 2003, p.26-34, http://rcdl2003.spbu.ru/ proceedings/C2.pdf..

[5] *U.Schoepflin*. The Archimedes Project: Realizing the Vision of an Open Digital Research Library for the Study of Long-Term Developments in the History of Mechanics // Proc. of the 5th National Russian Research Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" RCDL'2003, St.-Petersburg, Russia, 2003, p.124-129, http://rcdl2003.spbu.ru/proceedings/G2.pdf..

[6] *Марчук А.Г*. Электронные архивы, музеи и экспозиции // Труды 5-ой Всеросс. науч. конф. RCDL'2003, Санкт-Петербург, Россия, 2003. - Изд-во СпбГУ, 2003, с. 106-111, http://rcdl2003.spbu.ru/proceedings/E3.pdf.

[7] *Ю.И.Шокин, В.А.Ламин, А.М.Федотов, В.Б.Барахнин, О.Л.Жижимов, Н.А.Мазов, Б.Н.Пищик, Н.Н.Покровский*. Распределенная информационная система «Виртуальный музей науки и техники СО РАН» // Труды 5-ой Всеросс. науч. конф. RCDL'2003, Санкт-Петербург, Россия, 2003. - Изд-во СпбГУ, 2003, с. 112-116, http://rcdl2003.spbu.ru/proceedings/E4.pdf.

[8] *С.Х.Ляпин, А.В.Куковякин*. Многофункциональная электронная библиотека T-Libra: WWS-архитектура, интегрированный каталог, настраиваемый мультирубрикатор, гибкий параметризируемый полнотекстовый поиск // Труды 5-ой Всеросс. науч. конф. RCDL'2003, Санкт-Петербург, Россия, 2003. - Изд-во СпбГУ, 2003, с. 292-299, http://rcdl2003.spbu.ru/proceedings/J4.pdf

[9] *С.Х.Ляпин, А.В.Куковякин*. Виртуальная лаборатория для гуманитарных исследований на основе электронной библиотеки с гибким полнотекстовым поиском // Труды Всеросс. науч. конф. «Научный сервис в сети ИНТЕРНЕТ», г. Новороссийск (пос. Дюрсо), 20-25 сент. 2004 г. - М.: Изд. Московского гос. университета, 2004. - С. 45-47.

[10] *С.Х.Ляпин, А.В.Куковякин*. Концепт-ориентированный поиск в электронной полнотекстовой библиотеке с мультимодальным расширением // Труды 6-й Всеросс. науч. конф. RCDL'2004, Пущино, 29 сент. - 1 окт. 2004 г. - С. 127-134, http://www.impb.ru/~rcdl2004.

[11] *С.Х.Ляпин, А.В.Куковякин*. T-Media: от музейной библиотеки к информационной среде для интеграции музейных ресурсов и сервисов // Материалы 7-й ежегод. межд. конф. EVA 2004 Москва. - Москва, ГТГ, 29 нояб. - 3 дек. 2004 года. - М.: Изд. Центр ПИК Минкультуры России; Гос. Третьяков. Галерея, 2004, http://conf.cpic.ru/upload/eva2004/reports/doklad_57.doc.

[12] *Ляпин С.Х., Куковякин А.В*. T-Media: среда для создания виртуальных музейных экспозиций на основе расширения электронной полнотекстовой библиотеки // Культурное многообразие в едином информационном пространстве. Тезисы докладов IX ежегодной конференции АДИТ-2005, Казань, 30 мая - 3 июня 2005 г. - Казань, Национальный музей РТ, 2005. - 156 с. (с.74-75). http://www.adit.ru/rus/conference/adit2005/papers/paper.asp?nomer=4.

[13] *С.Х.Ляпин*. Функциональная организация интеллекта в многоцелевой информационной среде // Труды Международных научно-технических конференций «Интеллектуальные системы (IEEE AIS'05)» и «Интеллектуальные САПР» (CAD-2005). - Краснодарский край, пос. Дивноморское, сент. 2005 (в печати).

[14] *П.К.Анохин*. Избранные труды. Философские аспекты теории функциональной системы. М., 1978. / Философские аспекты теории функциональной системы, с. 27-48; Принципиальные вопросы общей теории функциональных систем, с. 49-106; Философский смысл проблемы естественного и искусственного интеллекта, с. 107-124.

[15] *П.К.Анохин*. Функциональная система как основа физиологической архитектуры поведенческого акта / П.К.Анохин. Избранные труды. Системные механизмы высшей нервной деятельности. – М., «Наука», 1979, с. 13-99.

[16] *С.Х.Ляпин*. Функциональная организация субъекта и ценностные детерминации в научном познании // Ценностные детерминации в научном познании. Межвуз. сб. науч. статей. Отв. ред. д.ф.н., проф. Л.А.Микешина. – Вологда, 1984. – С.24-44.

[17] *С.Х.Ляпин.* О концептах и концептологии (в поисках нового подхода к моделированию деятельности) // XIX World Congress of Philosophy. – Moscow 22-28 August 1993. – Book of abstracts. Сборник резюме. Vol. I. Секция 13 (Философия деятельности). – с.322.

[18] *С.Х.Ляпин.* Концептология: учение о концептах, методология культурогенных трансляций, технология эвристического развертывания смысла // Вестник СЗО РАО. – №3, 1998, СПб.-Архангельск: Поморский гос. университет им. М.В.Ломоносова, 1998. – с.28-41.\

[19] *С.Х.Ляпин.* Концептологическая формула факта // Концепты. Научные труды Центроконцепта. – Отв. ред. С.Х.Ляпин. – Вып. 2(2) 1997. – Архангельск: Изд-во Поморского гос. университета, 1997. – с.5-71.
[20] *Аджиев Алим Сапарович, Нгуен Хунг Мань*. Подходы к описанию и использованию тезаурусов в информационных системах // Труды 5-ой Всеросс. науч. конф. RCDL'2003, Санкт-Петербург, Россия, 2003, с.191-200, http://rcdl2003.spbu.ru/proceedings/F1.pdf..
[21] *Н.В.Лукашевич*. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер. 2. - 1995. - №3. - С.21-24.
[22] *Б.В.Добров, Н.В.Лукашевич, О.А.Невзорова*. Автоматизированное построение прикладной онтологии: технологические аспекты // Международная IEEE конференция «Искусственные интеллектуальные системы» (IEEE AIS'02), Геленджик-Дивноморское. В сб.: Обработка текста и когнитивные технологии (вып. 7). Под ред. В.Д.Соловьева. - Казань: Отечество, 2002. - С.103-109.
[23] *Б.В.Добров, Н.В.Лукашевич, С.В.Сыромятников*. Формирование базы терминологических сочетаний по текстам предметной области // Труды 5-ой Всеросс. науч. конф. RCDL'2003, Санкт-Петербург, Россия, 2003. - Изд-во СпбГУ, 2003. - С. 201-210, http://rcdl2003.spbu.ru/proceedings/F2.pdf..
[24] *Leonid Kalinichenko, Michele Missikoff, Federica Schiappelli, Nikolay Skvortsov*. Ontological Modeling // Proc. of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003, p.7-13, http://rcdl2003.spbu.ru/proceedings/D2.pdf.
[25] *Л.М.Веккер*. Психические процессы. Т.2. Мышление и интеллект. – Л., Изд-во Ленинградского университета, 1976. (Глава VI. Мышление как интегратор интеллекта, с.276-338, §2. Понятийная мысль как вид мышления и как форма работы интеллекта, с. 280-320).
[26] *Л.А.Заде***.** Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976. – 165с.
[27] *Л.А.Заде.* Роль мягких вычислений и нечеткой логики в понимании, конструировании и развитии информационных / интеллектуальных систем (пер. с англ. И.З.Батыршина) // Новости искусственного интеллекта, №2-3, 2001, с. 7 – 11, http://www.raai.org/resurs/resurs.shtml?publ?ainews
[28] *И.З.Батыршин*. Общий взгляд на основные черты и направления развития нечеткой логики Л.Заде // Новости искусственного интеллекта, №2-3, 2001, http://www.raai.org/resurs/resurs.shtml?publ?ainews
[29] *А.Н.Горбань, В.Л.Дунин-Барковский, А.Н.Кирдин, Е.М.Миркес, А.Ю.Новоходько, Д.А.Россиев, С.А.Терехов, М.Ю.Сенашова, В.Г.Царегородцев*. Нейроинформатика. – Новосибирск: Наука. Сибирское предприятие РАН, 1998. - 296с.
[30] *О.С.Бартунов, Т.Г.Сигаев*. Научная Сеть: алгоритмы и структуры данных // Труды Всеросс. науч. конф. «Научный сервис в сети ИНТЕРНЕТ», г. Новороссийск (пос. Дюрсо), сент. 2002 г. - М.: Изд. Московского гос. университета, 2002. - С. 211-213.
[31] *C. Faloutsos, S. Christodoulakis*. Description and performance analysis of signature file methods. ACM TOIS 1987
[32] Glimpse, Webglimpse, Unix-based search software… http://webglimpse.org
[33] *А.Н.Давыдов, Б.С.Ишханов, Э.И.Кэбин*. Виртуальный практикум по физике ядра и частиц // Труды Всероссийской научной конференции «Научный сервис в сети ИНТЕРНЕТ», г. Новороссийск (п. Дюрсо), 22-27 сентября 2003 г. - М.: Изд-во Московского университета, 2003, с.59-60.

Methodology and technology for creating of the multi-purposed information environment T-System based on the digital library with flexible full-text search

Sergey Lyapin, Alexey Kukovyakin

We describe hereby the methodology and technology for creating of the multi-purposed information environment 'T System' based on the extension of the digital library T-Libra. The environment is destined for the integration of resources and services, which a typical for digital library with flexible full-text search, virtual museum, digital archive, research laboratory and educational server. The methodological basis this sort of extension is a hybrid two-level ontology based on interaction of functional systems (top level), concepts library and thesauri library (lower level). The technological basis of extension is an administrative division, which has tools for flexible set-up of T-System's, as well as an unified search system includes the mechanisms of nonlinear cascade inquiries, which of that generate a relevant functional systems and combine the results of full-text search, related thesauri and concepts, text metadata and non-text objects of different modalities (graphics, sound, video etc.). The above-mentioned environment is designed in three-tier architecture (Web-browser / Web-server + Application server / DB server) with using of special indexing system for increase of search effectiveness, as well as of external logic which is built-up in Application Server.