НАУЧНЫЙ ИНТЕРНЕТ-РЕСУРС ДЛЯ СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЙ

Юдина Т.Н.

Ведущий научный сотрудник НИВЦ МГУ,

Москва 119899б Воробьевы горы*yudina@mail.cir.ru*;

Компьютерные и Интернет-технологии стимулируют развитие гуманитарных и социальных наук, решая проблему научного информационного обеспечения на качественном ином уровне и обеспечивая равные условия доступа всем исследователям и научным организациям. Формирование и поддержание полноценной информационной базы для профессионального анализа - репрезентативной по охвату, глубокой по ретроспективе, содержащей данные в форматах, удобных для вторичного анализа, регулярно обновляемой - сложная, трудоемкая и дорогостоящая задача, которая не может быть выполнена ни одним университетом или исследовательским центром самостоятельно. Наиболее рациональной формой организации информационной поддержки исследовательских и образовательных задач стали коллективные информационные центры. Подобные центры - архивы социальных данных - действуют в 30 странах мира [1]. Коллективные структуры создавались по инициативе самих научных сообществ на базе крупных университетов, способных обеспечить функционирование таких центров. В задачи коллективных центров входит а)целенаправленное формирование и поддержание информационных массивов за счет получения от государственных учреждений, закупки, соглашений о сотрудничестве с фирмами-владельцами и т.д.; б)академический сервис – содержательная обработка массивов и техническое сопровождение (перевод в форматы, удобные для компьютерного анализа, полная и понятная документация и т.д.), развитие поисковых инструментов; в) обеспечение надежного доступа; г)консультации и техническая помощь; д) координация исследований, информация о ведущихся проектах, е) проведение учебных курсов по компьютерным методам анализа в гуманитарных науках. Начинаясь как корпоративные научные информационные структуры, ориентированные на образовательные и исследовательские потребности, коллективные центры во всех странах стали важным элементом информационной инфраструктуры своей страны.

Проект “Университетская информационная система РОССИЯ” реализует аналогичное решение для России. [2] С февраля 2000 года Университетская информационная система РОССИЯ (УИС РОССИЯ) (**www.cir.ru**) функционирует как коллективная научная информационная база электронных ресурсов для исследований и образования в социально-гуманитарной области и доступна для университетов, вузов, академических институтов РФ и исследователей. Среди пользователей (июль 2001 года) – 130 коллективных членов (университеты, вузы, институты РАН, некоммерческие исследовательские центры) и более 500 индивидуальных. Допуск предоставляется бесплатно после регистрации пользователя.

В рамках проекта удалось решить ряд правовых и организационных вопросов, создать прецедент бесплатного получения на легальных условиях коммерческих ресурсов для некоммерческого использования в рамках коллективной межуниверситетской структуры. Все коллекции получены на основе прямых соглашений с владельцами прав на информационные материалы.

Система включает полнотекстовые документы (300 тысяч документов) и статистические данные в виде таблиц и графиков (15 тысяч документов) и поддерживается как интегрированный ресурс, с развитыми поисковыми возможностями и академическим сервисом, что обеспечено разработанными в рамках проекта технологиями.

Текущая версия включает следующие информационные ресурсы :

- нормативные документы федерального уровня с 1990 года;
- стенограммы пленарных заседаний Государственной Думы Федерального Собрания РФ;
- статистика Госкомстата РФ и Центризбиркома РФ, Межгосударственного статистического комитета СНГ;
- мониторинг Министерства экономического развития и торговли РФ;
- журнал “Эксперт”;
- издания СМИ - газеты “Аргументы и факты”, “Известия”, “Комсомольская правда”, “Независимая газета”, “Сегодня”, “Слово”; дайджесты Агентства “Восточно-европейская пресс-служба”;
- аналитические доклады и статистические массивы, подготовленные государственными организациями и независимыми исследовательскими центрами.

Создан первый предметно-ориентированный ресурс - база данных “Бюджетная система РФ” (www.budgetrf.ru).

Все ресурсы поступают в электронном виде из первоисточников и регулярно обновляются: ежедневно - нормативные акты федерального уровня и издания СМИ (еженедельно газета “Аргументы и факты”, журнал “Эксперт”); ежемесячно - стенограммы заседаний Государственной Думы Федерального Собрания РФ, мониторинг Министерства экономического развития и торговли РФ; одновременно с публикацией - массивы Госкомстата РФ, Центризбиркома РФ, Межгосударственного статистического комитета СНГ, аналитические доклады.

Следующим информационным блоком станут научные издания: в 2001 году – “Социологический журнал”, “Вопросы экономики”, “Федерализм”, гуманитарные серии “Вестника Московского университета” и “Журнала Санкт Петербургского университета”.

В 2001 году с УИС РОССИЯ будут интегрированы данные опросов общественного мнения ВЦИОМ (в перспективе - других аналогичных источников). Ведутся работы по включению изображений: разрабатывается технология поиска изображений на базе Тезауруса [3].

С включением данных опросов/обследований, научных изданий, изображений УИС РОССИЯ будет содержать основные компоненты научной информационной базы для большинства социально-гуманитарных дисциплин.

Система поддерживается как интегрированный ресурс, поэтапно формирующий информационную базу для полноценных системных исследовательских проектов - мониторинга и анализа социально-экономической ситуации в стране и в регионах. В рамках проекта разработан программно-лингвистический комплекс, обеспечивающий содержательную обработку данных и документов и интеграцию ресурсов в систему в автоматическом режиме. В составе комплекса – а)конверторы, каждый из которых настроен на преобразование определенного входящего потока данных и документов (.TXT, MS Word, HTML, . RTF) в единообразный формат хранения – HTML с автоматическим выделением формальных атрибутов документа, и б)автоматизированная лингвистическая обработка текстов (АЛОТ) [4]. АЛОТ последовательно выполняет несколько этапов анализа – графематический, морфологический и терминологический. Терминологический

анализ реализован на базе Информационно-поискового тезауруса по общественно-политической тематике (далее - Тезаурус), специально созданного для автоматического индексирования документов в рамках УИС РОССИЯ. Результатом терминологического анализа является тематическое представление содержания документа – выявление основных тем, которое служит основой для ранжированного рубрицирования. Выявление в документе основных тем позволяет определить предложения или фрагменты, в которых раскрывается тематика. Эти предложения образуют аннотацию документа. Строится также т.н. "структурная тематическая аннотация", где содержание текста представлено в виде совокупностей концептуально связанных терминов, что позволяет оценить содержание текста с первого взгляда. [5]

Ежедневно в систему интегрируется до 2 Мб электронных документов и данных.
Комплекс инструментов автоматической содержательной обработки и интеграции в систему основных типов ресурсов – официальных изданий, мониторингов и отчетов министерств, стенограмм заседаний парламента, опросов общественного мнения, аналитических докладов, научных изданий, СМИ, статистических данных, изображений - позволяет динамично масштабировать систему за счет дополнительных источников, необходимых для конкретных научных проектов

УИС РОССИЯ функционирует под управлением РСУБД ORACLE8i. Технологии реализованы на платформе Windows NT. Технология АЛОТ в 1997-98 годах была представлена на международную экспертизу в рамках программы "Конференции по интеллектуальным поисковым средствам в больших массивах текстов", проводимой на базе Национального института стандартов США и Агентства по развитию передовых (военных) технологий США. Результаты сопоставимы с лучшими достижениями в мире [6]

Академический сервис

УИС РОССИЯ ориентирована на научные потребности исследователей и содержит элементы т.н. "академического" сервиса – функций, выполняемых в научных организациях отделами информации. Сервис включает технологическую и научно-техническую составляющие.

Технология содержательной обработки и информационного анализа входящих ресурсов обеспечивает:

- систематизацию/классификацию всех коллекций по Тезаурусу и рубрикаторам.
- аннотирование полнотекстовых документов;
- индексирование по Тезаурусу оглавлений статистических таблиц и названий показателей;
- рубрицирование оглавлений таблиц статистических коллекций по Сводному оглавлению таблиц,
- рубрицирование научных изданий по Классификатору ГРНТИ, экономических также по рубрикатору JEL (Journal of Economic Literature);

Для массивов статистических данных проводится дополнительный комплекс научно-технических работ:

- перевод данных в формат таблиц MS Excel, удобный для эконометрического анализа - проведения расчетов, сравнений, построения строить графиков, загрузки в пакеты статистического анализа и т.д. для последующей обработки. В текущей версии системы в формате MS Excel доступны данные всех 32 сборников Госкомстата РФ, которые представлены в ретроспективе с 1996 года, - более 15.000 таблиц,

- привязка к статистическим таблицам Методологических пояснений (в развернутом и кратком вариантах) и глоссария;
- выделение в отдельное приложение и перевод в формат MS Excel таблиц и графиков, представленных в правительственных документах, аналитических докладах (в перспективе – в научных изданиях),
- конвертирование в формат MS Excel электоральной статистики,
- представление электоральной статистики в региональном преломлении – на карте,

Технологические решения обеспечивают развитые поисковые инструменты. В дополнение к традиционным видам поиска по библиографическому описанию и пословному возможен:
сквозной тематический поиск по рубрикаторам. В текущей версии системы используются два рубрикатора – рубрикатор УИС РОССИЯ (180 рубрикб 3 уровня иерархии) и рубрикатор Исследовательской службы конгресса США (80 рубрик верхнего уровня тезауруса LIV (Legislative Indexing Vocabulary). На стадии тестирования – технология рубрикации по Классификатору правовых актов РФ, утвержденному Указом президента РФ, 15.03.2000 (около 1200 рубрик, 4 уровня иерархии),
- навигация и уточнение запроса по иерархии Тезауруса. Тезаурус включает более 25,000 понятий, 56,000 терминов, 90,000 прямых и 750,000 наследуемых отношений между понятиями,
- доступ к статистическим коллекциям по Сводному оглавлению таблиц.

В УИС РОССИЯ карточка запроса формируется динамически в зависимости от источника, что позволяет выполнять запросы, специфичные для конкретной коллекции (например, “Номер документа” для нормативных актов, или “Выступающий” для стенограмм ГосДумы). Результаты поиска ранжируются в соответствии с оценкой релевантности содержимого документа запросу пользователя.

Пользовательский сервис включает возможность просмотра аннотации. Аннотирование производится для всех коллекций документов полного текста, за исключение Стенограмм пленарных заседаний Государственной Думы ФС РФ.

Реализован режим автоматического обновления типовых запросов пользователя с устанавливаемой пользователем регулярностью. Эта функция обеспечивает индивидуальное информационное обслуживание, помогая исследователям рационально организовать работу и минимизировать затраты на время нахождения в сети. По некоторым оценкам, до 50-60% времени специалиста уходит на поиск, предварительную оценку, отбор и систематизацию информации..

Академический сервис включает создание тематических ресурсов по наиболее востребованным научным направлениям. С июня 2001 года в рамках УИС РОССИЯ функционирует первый тематический (предметно-ориентированный) ресурс – база данных “Бюджетная система РФ” (www.budgetrf.ru) Ресурс включает ретроспективу документов с 1992 года по 2001, сгруппированных по этапам бюджетного процесса представление, рассмотрение и выполнение бюджета. Содержит бюджетные данные и сопроводительные документы правительственных организаций – Государственной Думы, Совета Федерации ФС РФ, Минфина, Центрального банка, Счетной палаты, Госкомстата и других, а также аналитические материалы, статьи научных журналов “Вопросы экономики”, “Федерализм”, “Эксперт”, статьи СМИ. В текущей версии

представлен федеральный уровень бюджетного процесса и данные о бюджетном процессе нескольких регионов. Ресурс включает развитый справочный блок и глоссарий. Реализованы элементы академического сервиса.

База данных создается в поддержку учебного процесса и включает специальный раздел – учебные материалы, подготовленные преподавателями Экономического факультета и Факультета государственного управления МГУ им.М.В.Ломоносова., в перспективе - преподавателями других университетов и вузов РФ. Ресурс будет развиваться за счет включения данных и документов по бюджетному процессу в регионах, а также путем включения материалов не-государственных исследовательских организаций. База данных "Бюджетная система РФ" представлена в открытом доступе.

В рамках проекта УИС РОССИЯ предпринимаются усилия по обеспечению надежного доступа из удаленных районов страны предполагается определить оптимальную схему зеркалирование ресурса в регионах на базе университетов. Первое зеркало установлено в мае 2001 года Санкт Петербурге на базе Междисциплинароного центра дополнительного профессионального образования Санкт Петербургского университета (www.uisrussia.nw.ru).

УИС РОССИЯ изначально создавалась как ядро корпоративной интегрированной информационной сети. Предполагается, что на базе университетов-зеркал будут создаваться региональные информационные системы с использованием технологии АЛОТ и опыта УИС РОССИЯ. Единая методика и технология обеспечат унификацию работы с документами и данными и сэкономят региональным университетам значительные средства. Коллектив УИС РОССИЯ по соглашению с каждым университетом-партнером передаст весь технологический комплекс региональным коллегам, проведет обучение и будет оказывать необходимую поддержку. Уже ведутся совместные работы с Междисциплинарным центром дополнительного профессионального образования Санкт Петербургского университета. Первым опытом создания региональных информационных ресурсов по технологии УИС РОССИЯ стал журнал "Санкт Петербургский университет". Сотрудничество с Санкт Петербургским университетом рассматривается как пилотный проект, в ходе которого будет отработана методика и практические решения, определена оптимальная схема взаимодействия.. Коллектив УИС РОССИЯ готов сотрудничать с любым региональным университетом, имеющим техническую базу и специалистов и способным взять на себя ответственность за организацию работ по поддержанию информационной системы на базе местных источников.

В истории развития новых информационных технологий и Интернет-технологий университетские сообщества сыграли значительную роль. В 70-ые годы университеты США, в 80-ые годы университеты стран Западной Европы стали ведущей силой в развитии научных исследований на базе компьютерных технологий и продвижении их в общество. Университеты расширили свои образовательные функции до просвещения и обучения общества в целом, содействуя развитию общей информационной культуры страны и повышению уровня компетенции всех уровней власти. В ведущих странах мира университеты включены в правительственные программы стратегического развития информационных технологий .

Университетское сообщество России является одной из наиболее подготовленных для восприятия и использования Интернет-технологий социальных групп и может сыграть ведущую роль в информационном развитии регионов и всей страны.

Работы по проекту УИС РОССИЯ начались в 1993 году и ведутся коллективом Автономной некоммерческой организации Центр информационных исследований, в 1993-1996 годах выполнялись на базе Института США и Канады РАН, с 1996 года работы проводятся совместно и на базе Научно-исследовательского вычислительного центра МГУ им. М.В.Ломоносова. Фундаментальные исследования по проекту поддерживаются фондом РФФИ, проект в целом - зарубежными фондами - Фондом Макартуров, Фондом Форда, с 2001 года - фондом Евразия.

Литература

1. Information Dissemination and Access in Russia and Eastern Europe. Problems and Solutions in East and West. Edited by Rachel Walker and Marcia Freed Taylor. NATO Science Series 4. Vol.26. IOS Press, Amsterdam, 1998, p.34/
2. Журавлев С.В, Юдина Т.Н. Информационная система “Россия”. НТИ. Серия 2, N 3, 1995
3. Добров Б.В., Лукашевич Н.В., Салий А.Д., Сидоров А.В., Штернова О.А., Юдина Т.Н., Постановка эксперимента по индексированию корпуса изображений с помощью специализированного тезауруса // Вторая Всероссийская научная конференция "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Протвино, 26-28 сентября 2000 г., С.274-279
4. Агеев М.С., Журавлев С.В., Ламбурт В.Г., Подготовка Web-версий традиционных изданий –// Открытые системы, 12(56), 2000, С.31-35.
5. Лукашевич Н.В., Добров Б.В., Исследование тематической структуры текста на основе большого лингвистического ресурса // Труды международного семинара Диалог`2000 по компьютерной лингвистике и ее приложениям, Таруса, 2000, С.252-258.
5. Dobroff Boris, Loukachevich Natalia, Yudina Tatyana. UIS RUSSIA: Conceptual Indexing Using Semantic Representation of Text. Proceedings of the TREC-6 Conference. 1998, National Institute of Standards, Washington, USA, p.411.

RESEARCH BASE FOR PROFESSIONAL INVESTIGATIONS AND EDUCATION IN SOCIAL SCIENCES

Tatyana Yudina
Leading researcher of Moscow State University Research Computing Center

*Moscow,119899, Vorobiovy Gory, **yudina@mail.cir.ru***

The University Information System RUSSIA (www.cir.ru) is available since February 2000 as a collective research base for investigations and education in social sciences. The system covers a wide scope of social domain documents and data from first-hand sources under legal agreements. The NLP technology developed under the project provides for up to 2 Mb of electronic documents are classified, conceptually indexed, annotated and integrated on daily bases. Academic services are provided for users. The resource is free for the universities and higher education institutions. The UIS RUSSIA is accomplished as a core for inter-university corporate network with regional universities developing information systems on local sources