МЕЖУНИВЕРСИТЕТСКИЙ ИНТЕРНЕТ-РЕСУРС ДЛЯ ИССЛЕДОВАНИЙ В ГУМАНИТАРНОЙ ОБЛАСТИ

Татьяна Юдина,

ведущий научный сотрудник
Научно-исследовательского вычислительного центра МГУ им. М.В.Ломоносова, к.и.н.

cir@online.ru

Использование новых информационных технологий в гуманитарных исследованиях требует целенаправленных усилий научного сообщества страны.

Как показал опыт США и стран Западной Европы, компьютерные и Интернет-технологии стимулируют развитие гуманитарных и социальных наук, позволяя организовать исследования и учебные программы на качественно новой информационной базе, превосходящей возможности традиционных способов информационного обеспечения и по охвату и ретроспективе источников, и по оперативности обновления данных и документов. На базе электронных источников и компьютерных технологий развиваются новые методы и направления исследований общественных процессов, основанные на анализе больших массивов данных и документов. Научные результаты исследований, проведенных с использованием реальных данных, имеют важное социальное значение, повышается востребованность. научных исследований обществом.

Формирование и поддержание (регулярное обновление) репрезентативного по охвату, глубокого по ретроспективе ресурсного комплекса - технологически не простая, трудоемкая и дорогостоящая задача, которая не может быть выполнена ни одним университетом или исследовательским центром самостоятельно.

Приступив к изучению компьютерных технологий социальных исследованиях в начале 70-ых годов, и осознав необходимость информационного комплекса, 27 университетов США в 1967 году создали рациональную модель решения проблемы ресурсного обеспечения – коллективную информационную инфраструктуру – Мичиганский межуниверситетский консорциум по социальным и политическим исследованиям. В настоящее время в составе консорциума – около 500 коллективных членов. Опыт США повторен в 80-90-ые годы странами Западной Европы, Австралии, Новой Зеландии, Венгрии. Различаясь по административному устройству, коллективные ресурсные центры выполняют ряд основных функций:

а) целенаправленное формирование и поддержание информационных массивов за счет закупки, соглашений о сотрудничестве с фирмами-владельцами и т.д. ;

б) обработка массивов (перевод в форматы, удобные для компьютерного анализа),

в) информационный сервис (обеспечение доступа к ресурсам);

г) консультации и техническая помощь;

д) информация о ведущихся проектах, координационная деятельность;

е) обучение по специальности "новые методы компьютерного анализа" на базе летних/зимних школ (рациональнее собрать представителей университетов, желающих специализироваться в области сложных методов компьютерного анализа в летней/зимней школе, чем университетам организовывать подобный (дорогостоящий) курс у себя). [1]

Коллективные информационные центры решают проблему информационного обеспечения и обучения перспективным методам анализа для всего университетского/ исследовательского сообщества страны, чем способствуют активизации научных исследований в гуманитарных областях, расширению географии проектов за счет местных университетов, получивших равные возможности доступа к ресурсам. Как результат – университетское сообщество становится более эгалитарным, развивается сотрудничество, создается корпоративная информационная культура.

В 70-80-ые годы машиночитаемые коллекции для конкретных исследований запрашивались университетами и рассылались консорциумами по почте. Сеть Интернет ускорила оперативность информационного процесса, приблизила ресурсы к рабочему столу исследователя и университетским аудиториям, Именно университеты США стояли у истоков, а затем способствовали быстрому развитию Интернет, осознав возможности сети как перспективного канала получения научной информации.

Создание коллективных ресурсов машиночитаемых данных стало важным и динамично развивающимся элементом информационного обеспечения научных исследований в США и Западной Европе при сохранении традиционно высокого уровня обеспеченности университетов информационными материалами в печатном виде.

Сложная, трудоемкая, наукоемкая, дорогостоящая задача организации современной ресурсной базы для учебных программ и научных разработок в гуманитарной области решается путем целенаправленных коллективных усилий самих университетов.

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

Рост числа источников, увеличение объема информации поставил задачу создания инструментов для эффективной обработки и рациональной организации данных для исследовательских целей, прежде всего классификации и интегрирования информационных массивов. Начиная с 1996 года университеты США активно развивают т.н. инициативу Интернет-2, которая включает программу "интеллектуализации" Интернет – развитие технологий содержательной обработки и интеграции больших массивов информации и организации тематических хранилищ данных как элементов "ресурсной базы XXI века" для гуманитарных исследований.[2]

Университеты Западной Европы с середины 90-ых годов развивают проект создания корпоративного ресурса правовых документов на базе единой технологии обработки национальных документов (распределенное индексирование).

Создание технологий, способных интегрировать информационные ресурсы мира в виртуальную библиотеку, - стратегическая цель мировой науки в области информационных систем. Первоочередными задачами признаны: а) разработка международного стандарта представления ресурсов и б) инструменты содержательного анализа информации и формирования тематических хранилищ, в) создание многоязычных поисковых средств. Именно исследования по этим направлениям объявлены приоритетными в правительственных, академических программах научного развития ведущих стран мира и бизнес-планах корпораций. [3, 4]

Информационная система РОССИЯ

При значительном отставании России в общем уровне информационного развития по сравнению с другими странами, университетское и научное сообщество России оказалось одним из наиболее технически оснащенных и подготовленных к восприятию и использованию компьютерных и Интернет-технологий в своей профессиональной деятельности и решению научных задач на современной аппаратной и программной платформе, Интернет-технологий и исследовательских разработках в области информатики, прикладной лингвистики, информационных систем.

Первоочередной научной задачей стало информационное обеспечение исследований и образовательных программ в гуманитарной области. С середины 90-ых годов университетские библиотеки практически не получают научной литературы. По данным Министерства науки и технологий РФ, за последние годы обеспеченность отечественными научными изданиями в пересчете на 10,000 исследователей составляет в России – 6,8 единиц, при том, что в Великобритании – 408 единиц, во Франции – 203, ФРГ – 196, США – 162, Японии – 67.

Создание коллективного информационного машиночитаемого ресурса – наиболее рациональное решение в России, учитывая, что сеть Интернет постепенно развивается и университеты получают возможности доступа, т.е. создана техническая основа для решения проблемы ресурсного обеспечения гуманитарных исследований в России.

С начала 90-ых годов в Московском университете предпринимаются целенаправленные усилия по организации Российского межуниверситетского ресурсного центра и разрабатывается проект "Информационная система РОССИЯ". [5, 6,7]. Круг первоочередных источников информации был определен совместно с Кафедрой экономической статистики экономического факультета МГУ и Центром социологических исследований МГУ (проекты "Рабочее место экономиста", "Рабочее место социолога") и включает электронные источники федерального и регионального уровня - правовые документы, статистические и справочные данные, СМИ федерального и регионального уровней, справочную информацию, ведущие научные журналы, издания университетов. Технологические решения позволяют:

а) интегрировать ресурсы и реализовать архитектуру хранилища данных (data warehouse) и

б) обеспечить поисковый сервис с элементами содержательного анализа документов (data mining) и развитыми функциональными возможностями – работа на уровне метаданных (metadata browsing), модификация запросов (query refinement), автообновление запросов по заданной теме. Технология нацелена на решение исследовательских и образовательных задач.

Информационная система РОССИЯ создается как коллективная Интернет-библиотека. Все региональные университеты получат доступ к системе и смогут на равных работать с информационными ресурсами. Наряду с Интернет-доступом предусмотрено предоставление коллекций на электронных носителях.

Прототип

Прототип Информационно-поисковой системы РОССИЯ создан коллективом Центра информационных исследований и действует в НИВЦ МГУ. Прототип реализован на информационных источниках федерального уровня на базе технологии автоматизированной лингвистической обработки текстов. Результат обработки – индексирование, рубрицирование, аннотирование и создание развернутого поискового образа документа – метаинформации. Технология обеспечивает обработку до 15 Мб электронных текстов в день (7-8 тысяч печатных страниц) и интегрирование результатов в ИС РОССИЯ. Технология прошла международную экспертизу в США, результаты – среди лучших достижений в мире. [8]

В ИС РОССИЯ реализованы развитые поисковые средства. Наряду с традиционными приемами, возможен поиск по рубрикатору (200 позиций) и тезаурусу. Тезаурус по политической жизни России – оригинальная разработка коллектива – включает более 30.000 терминов, 70.000 синонимов, 250.000 иерархических связей. Тезаурус позволяет осуществлять навигацию по всему массиву текстов. Пользователь может работать на уровне метаинформации - просматривать не весь текст документа, а лишь его поисковый образ или аннотацию.

В ИС РОССИЯ возможно авто-обновление запросов (пользователь может сформировать набор запросов по интересующей теме и регулярно получать обновление.

В текущей версии ИС РОССИЯ технология работает на массивах правовых документов. В течение 1999 года технология будет реализована на массивах СМИ - около 10 изданий газет и информационных агентств и научных изданиях.

Функциональные возможности ИС РОССИЯ ориентированы на интересы исследователей. До 70-80% времени специалиста уходит на поиск источников и отбор документов, инструменты автоматической обработки информации позволят более рационально организовать работу, расширить круг источников, обеспечив возможности комплексного подхода к решению задач, нацеленных на изучение социальных процессов.

Межуниверситетская корпоративная сеть

Предполагается, что региональные университеты станут не только пользователями, но и создателями информационных

ресурсов: в рамках проекта региональные университеты получат программно-технологические средства, в том числе упомянутый комплекс автоматизированной лингвистической обработки текстов, для организации и поддержки ИС регионального уровня. [9] . Благодаря программе "Университетские Центры Интернет" Института Открытое общество (фонд Сороса) в 33 университетах России создана аппаратно-программная платформа, необходимая для реализации проекта совместными усилиями университетов. Единая технологическая среда, унифицированные способы обработки документов обеспечит интеграцию региональных ИС с ИС РОССИЯ, развитый поисковый механизм с элементами анализа содержания текстов предоставят исследователям возможность системного подхода к процессам и явлениям. ИС РОССИЯ может быть использована для учебных целей и научных исследований в области экономики, управления, социальной сферы, культуры, права, международных отношений. Университетское сообщество России совместными усилиями сформирует современную ресурсную базу, равноправный доступ к которой обеспечит равные исследовательские возможности всем региональным специалистам.

В рамках проекта решается комплекс правовых и организационных вопросов. Обсуждаются условия закупки массивов для коллективного доступа университетов, получение скидок для использования в учебных и исследовательских целях.. Госкомстат РФ бесплатно предоставил комплекс статистических ресурсов с разрешением использовать в интересах коллективного межуниверситетского информационного центра. Ведутся переговоры с другими государственными организациями - владельцами информационных массивов.

По предложению Информационного отдела Совета Федерации Федерального Собрания РФ, проект будет представлен на заседании Совета Федерации с целью содействовать региональным университетам в установлении контактов с местными властями в интересах сотрудничества в создании региональных информационных систем.

В рамках проекта предполагается совместная с региональными специалистами разработка учебного курса по использованию новых информационных технологий в социологических исследованиях на базе имеющихся наработок.

С 2000 года предполагается проведение Летних школ по методике социальных исследований с использованием новых технологий. Курс будет организован на базе МГУ с приглашением на конкурсной основе участников из регионов. По возможности, к преподаванию будут привлечены зарубежные специалисты - преподаватели аналогичных школ из Мичиганского межуниверситетского консорциума (США) и Европейского консорциума (Великобритания).

В рамках проекта ИС РОССИЯ реализован двуязычный (русско-английский) комплекс поисковых средств - рубрикатор (200 позиций) и тезаурус (30,000 терминов). Перевод осуществлен с использованием лексики самых известных тезаурусов мира – Исследовательской службы Конгресса США, ООН, службы Легислейт, Евровок (тезаурус ЕС). [10]. Двуязычный комплекс и технология автоматической лингвистической обработки текстов позволят

а) работать с ИС РОССИЯ зарубежным пользователям,

б)включить в ИС РОССИЯ англоязычные массивы за счет регулярной обработки иностранных источников, доступных по Интернет.

В обоих случаях будет доступна метаинформация о каждом документе на двух языках и полный текст на языке оригинала.

Работы по проекту включают разработку программы-шлюза для адаптации международного стандарта (протокола Z 39.50) представления информационных ресурсов.

Развитый двуязычный комплекс и международный стандарт поддержания ресурса обеспечат развитие ИС РОССИЯ как составной части мирового информационного пространства, что значительно расширит исследовательские возможности российских специалистов.

В истории развития новых информационных технологий и Интернет-технологий университетские сообщества сыграли особую роль. В 70-ые годы университеты США, в 80-ые - и университеты стран Западной Европы стали центрами изучения компьютерных технологий и продвижения электронных направлений в общество. Университеты расширили свои образовательные функции до просвещения и обучения общества в целом, содействуя развитию общей информационной культуры страны. В ведущих государствах мира университеты включены в правительственные программы долгосрочного развития государства, разрабатывают стратегию образования, способную обеспечить конкурентоспособность страны в следующем веке. Инициатива Интернет-2, направленная на интеллектуализацию Интернет, усилит научный потенциал и образовательные возможности университетов, а значит возрастет роль университетов в развитии страны. [11].

Университетское сообщество России в настоящее время является одним из наиболее подготовленных социальных групп для восприятия, использования, распространения Интернет-технологий и может сыграть ведущую роль в процессе преодоления информационной отсталости регионов и всей страны.

Список литературы

1. Inter-university Consortium for Political and Social Research. Guide to Resources and Services. 1998, Ann Arbor, Michigan

2. University Corporation for Advanced Internet Development. Mission Statement. 1996, USA

3. Government Support for Computing Research. National Research Counsil. National Academy Press. 1999.

4. The National Information Infrastructure: the Federal Role. CRS Issue Brief. Congressional Research Service. 1999

5. Журавлев С.В., Юдина Т.Н. "Информационная система РОССИЯ". Научно-техническая информация, Сер.2, 1995, N3 .

6. Tatyana Yudina, Paul Dorsey. "IS RUSSIA: an Artificial Intelligence-based Document Retrieval System in Oracle-7 ". Select, 1995, N1б , Chicago.

7. Tatyana Yudina, Sergey Zhuravlev. "IS RUSSIA: an Automated Information Retrieval System". Proceedings of International Association for Computer Information Systems. 1995, Toronto, Canada.

8. B.Dobroff, N.Loukashevich, T.Yudina. IS RUSSIA: Conceptual Indexing Using Semantic Representation of Text. Proceedings of the TREC-6 Conference. 1998, Washington, USA.

9. Tatyana Yudina. "Information System RUSSIA: Russian universities' collective social sciences information center initiative". Proceedings of International Seminar "Data Dissemination and Access in Russia and Eastern Europe". 1998, University of Essex, Great Britain.

10. "Information System RUSSIA: the Bilingual (Russian-English) Search Complex
Proceedings of the Third European Seminar. 1998, Montecatini, Italy.

11. Telematics Applications Programme (1994-1998). European Commission. DGXIII, 1996.