Археологическая литература стран СНГ в пространстве Сети: подход и реализация

Publications on Archeology of the CIS Countries in the Internet: Approaches and Solutions

Археологічна література країн СНД в Інтернет: підходи та реалізація

Багажков А. К.
Библиотека РАН, Санкт-Петербург, Россия

Вершинин М. И.
Северо-западный технический университет, Санкт-Петербург, Россия

Всевиов Л. М.
Библиотека РАН при Институте истории материальной культуры РАН, Санкт-Петербург, Россия

Alexander K. Bagazhkov
Russian Academy of Sciences Library, St. Petersburg, Russia

Mikhail I. Vershinin
Northwestern Technological University, St. Petersburg, Russia

Leo M. Vseviov
Russian Academy of Sciences Library at the Institute of the History of Material Culture, Russian Academy of Sciences, St. Petersburg, Russia

Багажков О. К.
Бібліотека РАН, Санкт-Петербург, Росія

Вершінін М. І.
Північно-західний технічний університет, Санкт-Петербург, Росія

Всевіов Л. М.
Бібліотека РАН при Інституті історії матеріальної культури РАН, Санкт-Петербург, Росія

Трудности в процессе поиска и получения данных исследователями, трудоемкость создания предметно-ориентированных информационных ресурсов требуют применения новых технологий обработки данных для представления их в Интернет и на оптических носителях. Для представления библиографической базы данных Института истории материальной культуры использованы программно-лингвистические средства VerWEB и VerCON. Предложено решение проблемы учета ошибок различного типа и их автоматического исправления путем модификации данных перед сравнением.

Difficulties with which researchers meet while searching for and retrieving data and labor consuming creation of themed information resources demand that new data processing technologies have to be represented in the Internet and on optical disks. VerWEB and VerCON software and linguistic ware were used to represent bibliographical databases of the Institute of the History of Material Culture. The authors propose to resolve the problem of error check and automatic correction by means of data modification made prior to matching.

Труднощі в процесі пошуку й отримання даних дослідниками, трудомісткість створення предметно-орієнтованих інформаційних ресурсів вимагають застосування нових технологій обробки даних для представлення їх в Інтернет і на оптичних носіях. Для представлення бібліографічної бази даних Інституту історії матеріальної культури використані програмно-лінгвістичні засоби VerWEB і VerCON. Запропоновано вирішення проблеми врахування помилок різного типу і їх автоматичного виправлення шляхом модифікації даних перед порівнянням.

Библиотека Института истории материальной культуры — одна из крупнейших в мире специальных археологических библиотек — основана в 1859 г. как библиотека Имп. Археологической комиссии при Министерстве Имп. Двора. Библиотека в значительной своей части комплектовалась первоначально за счет обмена с изданиями Российской Академии наук, Академии Художеств, духовных академий, научных обществ, даров и покупки книг за границей. Большая роль в реорганизации библиотеки после Октябрьской революции принадлежит академику С. А. Жебелеву и археологу А. А. Спицину. Последним (в 1921 г.) было разработано первое «Положение о библиотеке РАИМК», в котором были определены задачи библиотеки: «... обслуживание научных и просветительских потребностей Академии; удовлетворение запросов всех научных учреждений и лиц, нуждающихся в научных книгах по областям ведения РАИМК».

В начале 20-х годов в библиотеке создан предметный каталог, затем систематический. К 1937 г. фонд библиотеки составлял свыше 150 тыс. ед. Был широко развернут международный книгообмен с 128 научными учреждениями из 28 стран. Традиционно, свыше 12% изданий поступало в дар.

В настоящее время фонд библиотеки насчитывает более 200 тыс. ед. В составе фонда — литература по всемирной археологии, истории древнего мира и средних веков, истории России (в основном до XVIII в.), истории искусства и архитектуры, науки и техники, религии, краеведению, музееведению, охране и реставрации памятников, по применению методов естественных и технических наук в археологии. А так же литература по нумизматике, эпиграфике, сфрагистике, геральдике, исторической географии, палеоантропологии, палеоботанике, палеозоологии, четвертичной геологии. Фонд библиотеки включает коллекцию Н. И. Репникова

В фондах библиотеки представлены российские дореволюционные периодические издания археологических учреждений; полные комплекты иностранных периодических изданий.

Справочный аппарат состоит из алфавитного и систематического каталогов, а также ряда картотек: работ сотрудников ИИМК, «Rossica», рецензий, годовых картотек «Археологическая литература России и стран ближнего зарубежья», которые аккумулируется в издаваемый с 1959 г. многотомный ретроспективный библиографический указатель «Советская археологическая литература».

С 1965 по 1999 годы вышло в свет девять томов указателя «Советская археологическая литература.1918—1987». В настоящий момент подготовлена рукопись ретроспективного указателя с 1988 по 1991 гг. Вся литература, с 1992 по 1997 годы собрана в ежегодные картотеки в систематизированном виде.

Отечественная литература по археологии с 1998 г. и по настоящий момент вносится в базу данных (БД) «Археологическая литература стран СНГ» в формате ППП CDS ISIS. Структура БД аналогична схеме ранее вышедших указателей. Однако, если в изданиях «Советской археологической литературе» учитывалась всю археологическая литература, вышедшую в СССР, то в БД отражены только поступления в отдел БАН при ИИМК, следовательно, можно утверждать что если в вышедших томах представлено приблизительно 90—95 % всей археологической литературы, то в базе данных не более 70 % литературы вышедшей в странах СНГ.

В БД учитывается литература по археологии, труды по древней истории и истории средних веков, основанные на археологическом материале труды по истории науки и техники, работы по смежным дисциплинам — эпиграфике, сфрагистике, геральдике, исторической географии. Включены основные работы по четвертичной геологии, палеогеографии и палеодемографии, а также по этнографии и языкознанию, содержание которых связан с проблемами археологии. В БД включены работы иностранных авторов, опубликованные в РФ и СНГ. Добавлен раздел «Археология нового и новейшего времени». Но не отражены буклеты и статьи из популярных изданий на языках народов РФ и СНГ. Издания на белорусском и украинском языках, при отсутствии русских резюме или оглавления, описаны на языке подлинника, фамилии авторов приводятся в русской транскрипции. Литература на европейских языках приводится в подлиннике, фамилии авторов даются на языке оригинала и русском. Тезисы докладов различных конференций, не расписываются. Все библиографические описания составлены по ГОСТ 7.1—84

Пользователи БД могут подобрать нужную информацию по указателям археологических культур и памятников, смежным дисциплинам, предметному указателю, указателям по месту издания, издающим организациям и издательств, названиям расписываемых статей, указатель персоналий.

БД «Археологическая литература стран СНГ», поступившая в Отдел БАН при ИИМК 1998—2005гг. « пополняется ежедневно, на 15 марта 2005 года она насчитывает 17894 записей, лучший способ сделать ее доступной и ввести в широкий научный оборот — открыть доступ через Интернет.

www.ban.ru

Для представления БД на Web-страницах используются два основных способа: статическая и динамическая публикация Web-страниц с информацией из БД.

Динамическая публикация используется если необходимо публиковать информацию БД в реальном масштабе времени. Например, в системах электронной коммерции и бизнес-информации. В этом случае Web-страницы создаются после поступления запроса на Web-сервер, который передает запрос на генерацию этих страниц программе, формирующей требуемый документ. Затем готовый документ отсылается обратно браузеру.

При статической публикации БД Web-страницы создаются и хранятся на Web-сервере до поступления запроса на их получение. Этот способ используется при публикации информации, содержащейся в достаточно редко актуализируемой БД. Такая организация публикации БД имеет такие преимущества, как более быстрый доступ к информации и уменьшение нагрузки на сервер при обработке запроса. При их преобразовании в Web-страницы следует решить несколько задач:

- минимизировать ручной труд при преобразовании библиографической БД (ББД);
- выбрать структуру Web-представления ББД;
- сократить время отклика за счет минимизации трафика.

Все эти задачи можно решать с помощью программных средств.

В Библиотеке РАН разработан программный комплекс VerWeb [1], позволяющий автоматически формировать Web-страницы в виде структуры, имеющей от одного до трех ссылочных уровней, плюс нижний, (информационный) уровень, с которого может происходить переход к изображению, полному тексту или другой HTML-странице.

Таким образом, ББД преобразуется в систему многоуровневых указателей различного вида: алфавитные, хронологические, смешанные (с переходом к полному тексту и/или изображениям).

Указатели на каждом уровне могут быть различных типов, например:

- текст (любая текстовая информация из полей БД);
- алфавит (точки входа в виде цифр от 0 до 9 и букв латинского и русского алфавитов);
- алфавит-2 (точки входа в виде двухсимвольных сочетаний), который должен присутствовать только после алфавита;
- дата (точки входа в виде цифр и/или диапазонов дат);
- различные комбинации значимых терминов поля записи с 1-го по 4-й.

Под значимыми терминами понимаются термины, не входящие в неинформативную лексику[1].

Еще одна проблема, решение которой важно для автоматизированного формирования Web-представления ББД это наличие ошибок, что также приводит к разрастанию HTML-страниц[2].

С учетом анализа искажений, в том числе в БД, проведенного различными исследователями [2—6], можно предложить следующую типологию ошибок:

- замена одной буквы на другую;
- пропуск букв (преимущественно гласные);
- удвоение букв (преимущественно согласные);
- замена буквы на близкую по звучанию (преимущественно согласные);
- замена буквы на совпадающую по написанию букву из другого алфавита;
- перестановка букв;
- вставка лишних букв (преимущественно не более одной);
- вставка лишних пробелов перед и/или после лексической единицы (ЛЕ);
- неклассифицируемые ошибки;
- сочетание предыдущих ошибок.

Для ББД отмечены специфические ошибки [7—10]:

- ошибочный тег поля;
- ошибочная метка подполя;
- опущенная информация;
- неверная трактовка данных;
- наличие дублетов (не обязательно совпадающих во всех полях);
- ошибки, возникающие при транслитерации текста;
- перестановка терминов (например, в названиях рубрик);
- диахрония терминов.

Изучение статистики ошибок показывает[2—5, 9б 11—13]::в среднем в записях ББД даже при тщательном контроле ввода частота ошибок составляет не менее 0,1%, в том числе

- однобуквенные ошибки составляют 85—95%;
- более вероятно искажение начала лексической единицы (ЛЕ): для слов длиной 3—8 символов наиболее вероятны ошибки в 3—4 позиции;
- примерное распределение ошибок: пропуск буквы составляет 30—40% (в т. ч. до 40% одной из удвоенных букв), вставка — 25—35% (в т. ч. до 45%—удвоение букв), замена — 15—20%, перестановка — 10—15%;

---

[1] Понятие неинформативной лексики (НИЛ) шире, чем распространенный термин “стоп-слова”. Дело в том, что в НИЛ входят не только так называемые стоп-слова, но и значимые термины, не несущие информации в контексте той или иной базы данных. Так, например, термин “археология” в контексте археологической БД не является информативным. С другой стороны, при сегментировании на термины таких полей как заглавие, аннотация и подобных, создается чрезмерное количество точек входа, что приводит к разрастанию объема соответствующих HTML-страниц. Использование файла НИЛ позволяет в несколько раз сократить число точек входа и уменьшить размер HTML-страниц.

[2] . Можно много и долго обсуждать, как избавиться от ошибок в БД, но следует признать их неизбежность. Тем более, что для библиографических БД это не катастрофично так как семантически большинство полей БД устойчиво к ошибкам. Таким образом, следует учитывать наличие в БД ошибок различного рода и учиться работать с ними для получения полезного результата.

- ошибки в гласных буквах (вставка и пропуск) встречаются чаще, чем в согласных;
- более вероятны ошибки в начальных ЛЕ полей ББД.

Кроме того, учет характеристик искажений должен ориентироваться на естественный язык, предметную область ББД и конкретного оператора, то есть на определенное представление знаний.

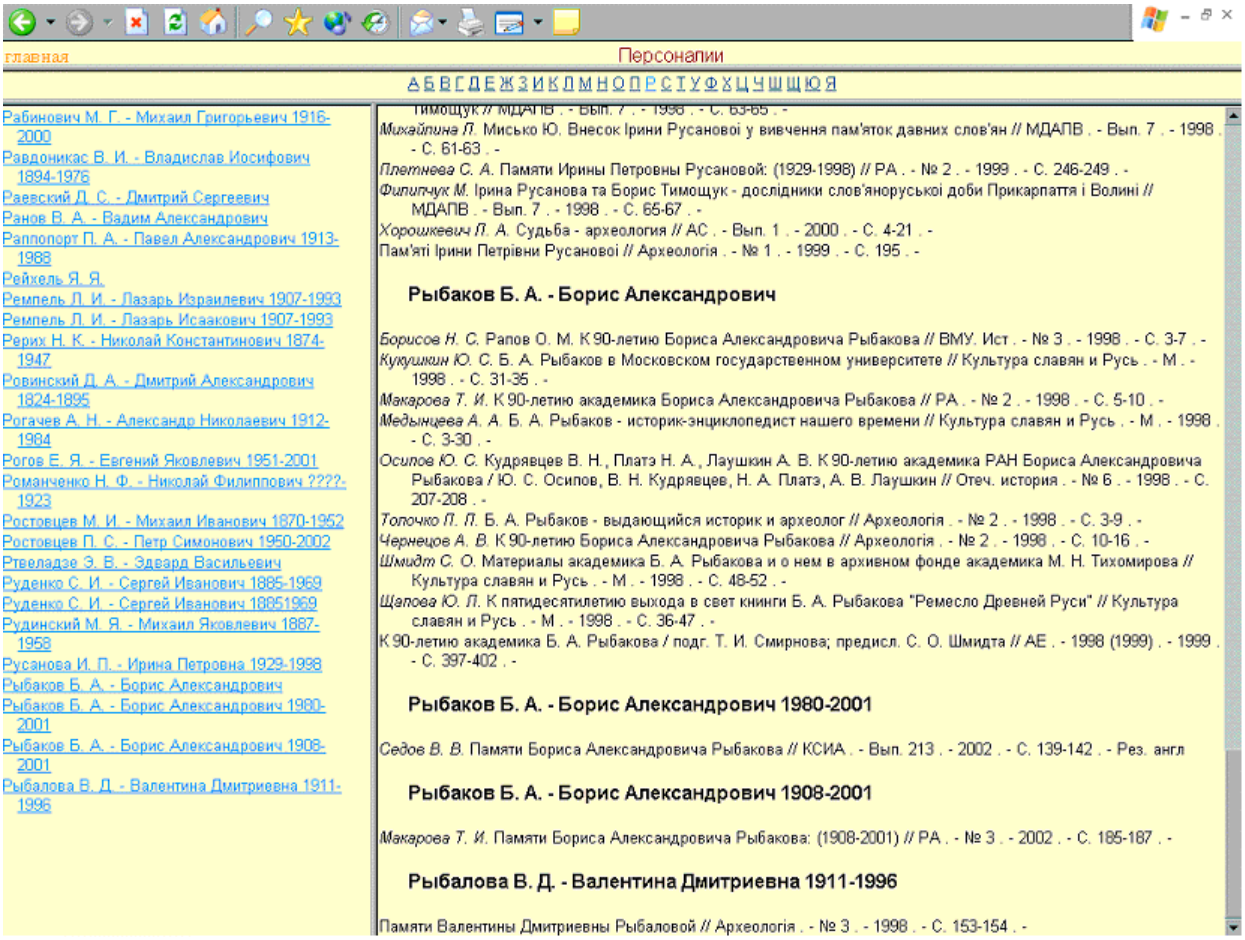
Программное средство VerCON [14], разработанное в развитие программного комплекса VerWEB [1], позволяет учесть наличие ошибок различного рода в ББД путем применения лингвистического конвертора, обеспечивающего устранения влияния ошибок различного рода на процесс построения точек входа на HTML-страницах. Лингвистический конвертор ориентирован на работу с латинским и кириллическим алфавитами.

Основываясь на анализе типовых ошибок, сравнение терминов и мультитермов (строк) выполняется следующими способами:

1. Точное совпадение
2. Совпадение без служебных символов: ~@#$%|_.,;!?`-=+:<>«[] (){}'\ /
3. Совпадение без служебных символов и цифр
4. Совпадение без служебных символов, гласных, двойных согласных и цифр
5. Совпадение по частотной карте
6. Совпадение по частотной карте без цифр
7. Совпадение по первому числу

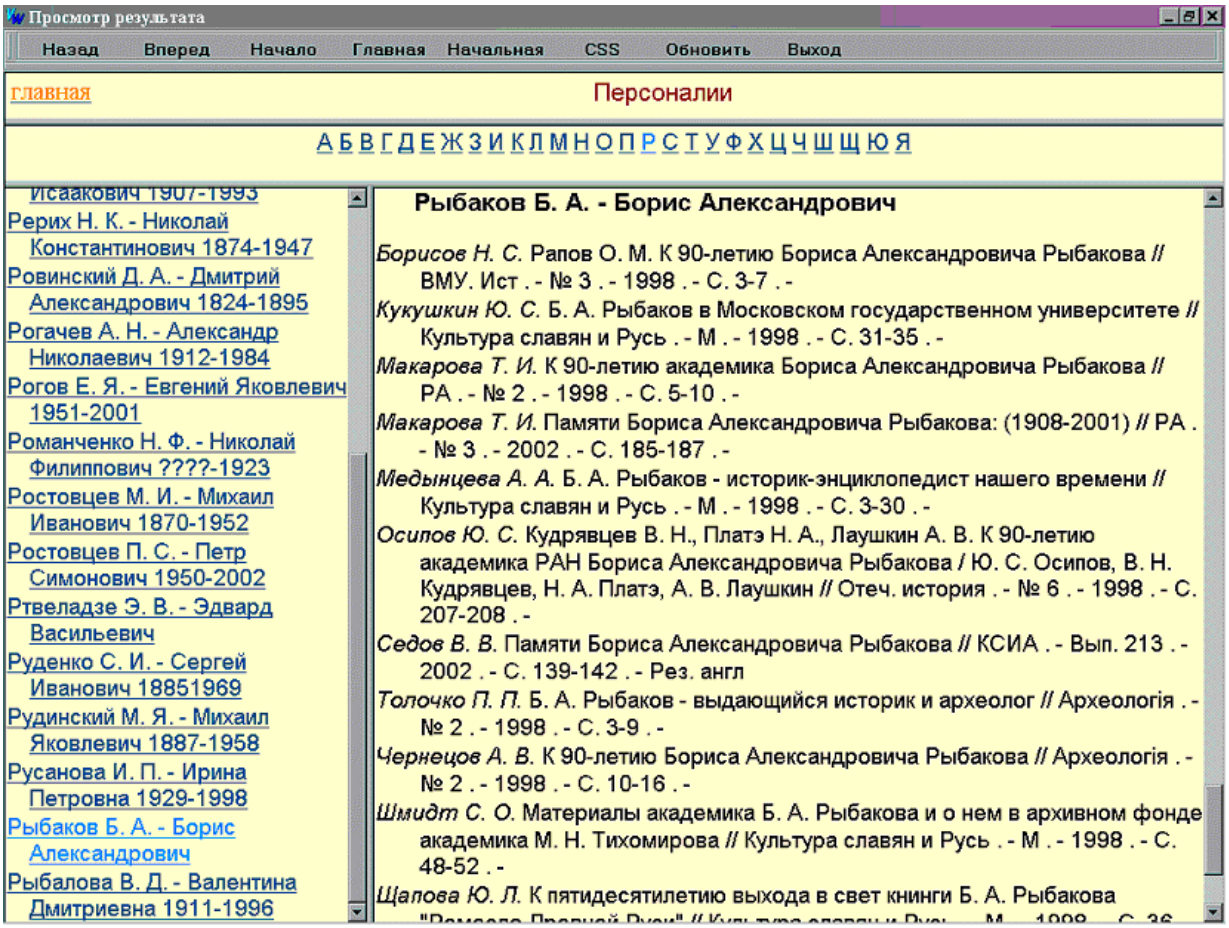
Помимо этого, для проверки может извлекаться часть строки до указанного набора символов или после него. Это полезно когда, например, названия рубрик сопровождаются числовыми индексами рубрик после символа = (например, сравнение в этом случае можно проводить 7-м способом).

Следует отметить, что сравнение по частотной карте — способ чувствительный к длине анализируемой строки и лучше подходит для коротких строк, так как в случае длинных мультитермов частоты некоторых символов начинают стремиться к величинам характерным для данного языка, что приводит к совпадению частотных карт даже весьма различающихся строк символов.



Преобразование БД «Археологическая литература стран СНГ», поступившая в Отдел БАН при ИИМК 1998—2005» в систему HTML-страниц занимает не более 1 мин. на ПЭВМ с процессором Celeron-1800 и оперативной памятью 256 Mb.

Оба программных средства хорошо зарекомендовали себя при создании HTML-представления ББД [15,16]:



Литература

1. Вершинин М. И. VerWEB — HTML-генератор для библиографических баз данных/ М.; ВНТИЦ, НГР 50200200489, 2002.
2. Ballard T. Spelling and typographical errors in library databases: One libr. system for rooting out spelling error/ T. Ballard // Computer in libr. — 1992. — Vol.12, № 6. — P.14—19.
3. Bourne C. Frequency and impact of spelling errors in bibliographic data bases / C. Bourne // Inform. processing a. management. — 1977. — № 13. — P.1—12.
4. Pollock J. J. Collection and characterization of spellings error in scientific and scholarly text / J. J. Pollock, A. Zamora // J. of the Amer. soc. for inform. science. — 1983. — Vol.34, № 1. — P.51—58.
5. Randall B. N. Spelling errors in data bases: Shadow or substance?/ B. N. Randall // Libr. resources a. techn. services. — 1999. — Vol.43, № 3. — P.161—169.
6. Szanser A. J. Automatic error correction in natural languages / A. J. Szanser // Inform. storage a. retrieval. — 1970. — Vol.5, № 4. — P.167—174.
7. Aissing A. L.computer-oriented bibliographic control for cyrillic documents with or without script conversion / A. L. Aissing // Inform. technology a. libr. — 1992. — Vol.11, № 4. — P.340—344.
8. Humphrey S. M. Automatic indexing of documents from journal descriptors: A prelim. investigation / S. M. Humphrey // J. of the Amer. soc. for inform. science. — 1999. — Vol.50, № 8. — P.661—674.
9. Nielsen R. Lost articles: Filing problems with initial articles in data bases / R. Nielsen, J. M. Pyle // Libr. resources a. techn. services. — 1995. — Vol.39, № 3. —P.291—293.
10. O'Neil E. T. Characteristics of duplicate records in OCLC's online union catalog / E. T. O'Neil, S. A. Rogers, W. M. Oskins // Libr. resources a. techn. services. — 1993. — Vol.37, № 1. — P.59—72.
11. Бабко-Малая О. Б. Методы и системы автоматизированного обнаружения и коррекции текстовых ошибок / О. Б. Бабко-Малая, В. А. Шемраков. — Л.: БАН СССР, 1987. — 46 с. — (Препр. / Б-ка АН СССР; № 5).

12. Бабко-Малая О. Б. Основные принципы автоматизированной коррекции текстовых ошибок / О. Б. Бабко-Малая, В. А. Шемраков // Распределенные автоматизированные библиотечно-информационные системы и сети. — Новосибирск, 1986. — С.127—131.
13. Белоногов ГГ. Результаты функционирования в ВИНИТИ системы обнаружения орфографических ошибок в режиме опытной эксплуатации / ГГ. Белоногов, Я. П. Штурман, Б. А. Кузнецов // Вопр. информ. теории и практики. — 1984. — № 51. — С.24—44.
14. Вершинин М. И. VerCON — лингвистический HTML конвертор для библиографических баз данных/ М.; ВНТИЦ, НГР 50200401169, 2004.
15. Вершинин М. И., Гроздилова Л. П., Немчинова А. Л. Создание электронного каталога иностранных журналов библиотеки Зоологического института РАН: подходы и реализация / М. И. Вершинин, Л. П. Гроздилова, А. Л. Немчинова // Науч. и техн. б-ки. — 2004. — № 6. — С.17—26.
16. Вершинин М. И., Колпакова Н. В., Золотарев В. М. Разработка предметно-ориентированных информационных баз данных. С.247—252. Научно-Технический Вестник СПбГУ ИТМО, Вып.13. Оптические технологии в фундаментальных и прикладных исследованиях — «Интеграция-2004» /Под ред. В. М. Золотарева. — СПб: СПбГУ ИТМО, 2004, 316с.