

Северная икона в мультимодальной библиотеке: к функциональной интеграции иконографических и полнотекстовых ресурсов

С.Х. Ляпин¹, А.В. Куковякин²

¹ Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)

lyapins@yandex.ru

² ООО «Константа»

magicmagus@yandex.ru

Аннотация

Описывается приспособление функционала электронной библиотеки, созданной на основе информационной системы T-Libra (разработка ООО «Константа», Архангельск) для интеграции ресурсов различной модальности (текст, графика, аудио, видео) в составе электронной полнотекстовой библиотеки (*внутренняя функциональная интеграция*). Для работы с иконографическими ресурсами использован функционал электронного каталога (адаптированный к стандартному иконографическому описанию), а также *гибридный поиск* по каталожным карточкам и полным текстам. Реализованы также квазисемантические запросы от «иконы к полному тексту» – с использованием предварительной автоматической экспликации предметной области поискового запроса и кластеризации его результатов.

Ключевые слова: сервисы полнотекстового поиска; мультимодальные ресурсы; гибридный запрос; каскадный запрос; квазисемантический поиск; кластеризация результатов запроса; абзацно-ориентированный запрос; частотно-ранжированный запрос; тематические коллекции запросов.

1. Введение. Эволюция музейно-информационного разнообразия

В последнее время в музеях стали появляться электронные библиотеки с возможностями продвинутого полнотекстового поиска, в том числе в распределенной среде. Многофункциональные сервисы этих библиотек могут использоваться для поддержки научно-методической, экскурсионной, лекционной, экспозиционно-выставочной и иной деятельности музея [1].

Вместе с тем достаточно давно во многих музеях функционируют информационные системы, содержащие цифровые изображения объектов музейного фонда с их специализированными описаниями: это прежде всего учетно-фондовые системы, но также и созданные под конкретные проекты различные электронные музейные коллекции.

Эволюция музейно-информационного разнообразия приводит к проблемам взаимодействия различных специализированных систем в рамках интегрированной музейной информационной среды.

Можно говорить при этом о *внутренней интеграции* (в рамках расширения функционала какой-то из специализированных систем), и *внешней интеграции* (в рамках взаимодействия сервисов различных информационных систем, в том числе в распределенной среде). Разумеется, при этом оба типа интеграции могут сосуществовать и дополнять друг друга.

В докладе рассматривается *внутренняя интеграция* в составе многофункциональной и мультимодальной электронной библиотеки, созданной на основе информационной системы T-Libra (разработка ООО «Константа», Архангельск): взаимодействие в ее рамках иконографических описаний икон Русского Севера – так называемых «северных писем» – и полнотекстовых ресурсов гуманитарной тематики.

Целью организации такого взаимодействия является использование *инструментов электронной библиотеки* для изучения иконографического материала в более широком культурно-историческом контексте, чем это позволяет сделать собственно иконографическое описание («музейная этикетка» предмета). Важность рассмотрения иконографического материала в таком контексте подчеркивается многими исследователями и экспертами в этой области, в том числе и касательно «северных писем» [2, 3].

2. Сервисы полнотекстового поиска

В используемой нами текущей версии электронной библиотеки [4] имеются следующие типы полнотекстового поиска: *а) абзацно-ориентированный, б) частотно-ориентированный*. При этом абзацно-ориентированный поиск представлен разновидностями работы как в локальной, так и в распределенной среде. Для целей настоящего доклада используется его версия, включающая кластеризацию результатов запроса в ходе его выполнения.

Абзацно-ориентированный поиск предназначен для поиска и презентации текста с точностью до отдельных авторских абзацев, содержащих заданную пользователем терминологическую структуру (тем самым эксплицируется «горизонтальный» микроконтекст, в котором в составе абзаца находятся смысловые термины). Авторский абзац выбран в качестве естественной единицы смыслового членения текста. Поиск ведется с учетом словоизменительной парадигматики (для русского языка). Обеспечивается поддержка нескольких видов и различных форм презентации результатов этого поиска.

Простой («однослойный») тематический поиск, с одним комплексным полем для ввода терминов и использованием для этих терминов операторов логического *объединения*, *обязательного исключения* или *обязательного*

включения термина в запрос. Результатом поиска является список абзацев, удовлетворяющих заданным условиям.

Каждый из абзацев, входящих в результаты запроса, может быть одним «кликом» мышки раскрыт до своего полного вида. Используя опцию «Контекст» в левом меню, можно последовательно раскрыть абзацы до и после найденного – вплоть до кластера из семи абзацев (три абзаца «до», три абзаца «после», плюс сам абзац – результат запроса).

Имеется возможность посмотреть, с этой же экранной страницы, соответствующий ресурс (статью, книгу и т.д.) в файловом виде; ресурс при этом может быть представлен в различных форматах: текстового документа, графического образа документа (важно для архивных ресурсов), сопровождающего документ аудио- или видеофайла (важно для организации электронных выставок и коллекций).

Расширенный («многослойный») тематический поиск. Этот вид поиска содержит функционал дополнительной тематической фокусировки запроса. Соответствующий инструментарий включает в себя: а) формирование нескольких поисковых полей («слоев») и б) включение в запрос дополнительных количественных параметров его фокусировки.

Поисковое поле "слои" представляет собой технический инструмент для выделения того или иного содержательного "аспекта" интересующей пользователя "темы"; всего может быть сформировано от 2 до 8 слоев. Например, в первом слое вводим термин «человек», во втором – термин «мир», в третьем – термин «жизнь». Тем самым в структуре запроса тематика «человека» специализирована (аспектуализирована) в связи с «миром» и «жизнью».

Еще более точная тематическая фокусировка запроса достигается за счет выполнения дополнительных условий: а) указания минимально необходимого количества поисковых слоев (от 2 до 8); б) указания максимального расстояния между терминами, принадлежащими разным слоям: от 0, когда слова из двух разных слоев запроса в составе абзаца примыкают друг к другу (например, «жизнь человека» и т.д.), до произвольной величины.

Частотно-ориентированный поиск предназначен для построения частотно-ранжированных списков терминов (существительных), и тем самым экспликации различных «вертикальных» макроконтестов, неявно присутствующих в отдельном документе или их выбранной совокупности. Получающиеся таблицы списков терминов, с указанием абсолютного (в обычных числах) и относительного (в %, промилле) количества их встречаемости в тексте, мы называем «терминограммами» (по аналогии с «рентгенограммами»). Поиск может проводиться одновременно по 1, 2 или 3 корзинам ресурсов.

Примечание. Этот тип поиска используется также для предварительной кластеризации результатов абзацно-ориентированных запросов; в этом случае он встроен в механизм их осуществления.

Обеспечивается поддержка двух видов частотно-ориентированного поиска и различных форм презентации его результатов:

абсолютный частотный, результатом которого является частотно-ранжированный список существительных, входящих в ресурсы области поиска

и приведенных к нормальной форме (именительный падеж, единственное число).

относительный частотный, результатом которого является частотно-ранжированный список существительных, входящих только в те абзацы, которые содержат заданный пользователем термин (тем самым список строится «относительно» этого термина).

Эти виды частотного поиска могут использоваться для целей текстологического анализа документа; для выявления и описания предметной области документа; для составления списка ключевых слов; для сравнительного анализа предметных областей различных авторов или различных документов; для кластеризации результатов абзацно-ориентированных запросов и т.д.

3. Использование каталога

Иконографические описания размещены в электронном каталоге ИС T-Libra. Для этого были использованы поля библиографического каталога (стандарт RUSMARC), который имеется в информационной системе. Важную роль в дальнейшем играет поле Аннотация (330 поле стандарта RUSMARC: «Резюме или реферат»), в которое помещается весь текст «музейной этикетки»; обычно это около 1 страницы текста. Термины из аннотации будут использованы в гибридном квазисемантическом запросе.

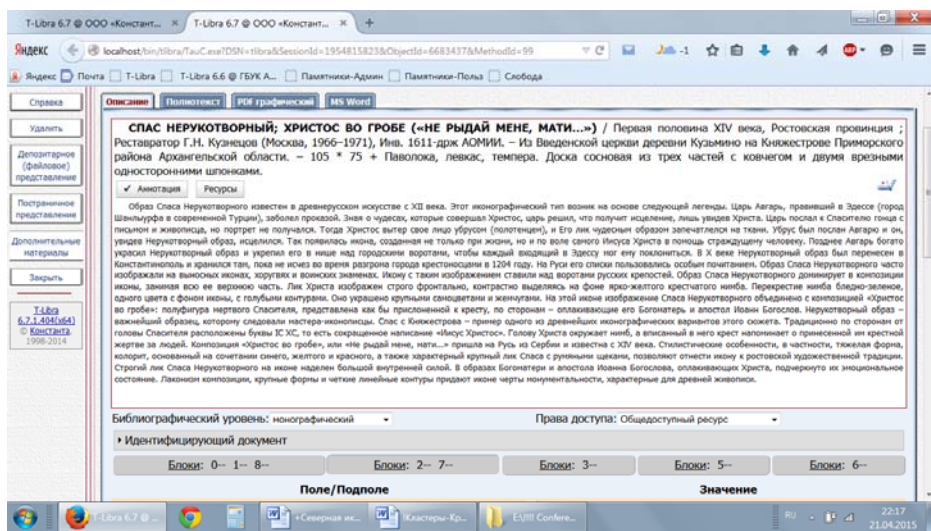


Рис.1. Часть электронной карточки иконы «Спас Нерукотворный» с раскрытой аннотацией

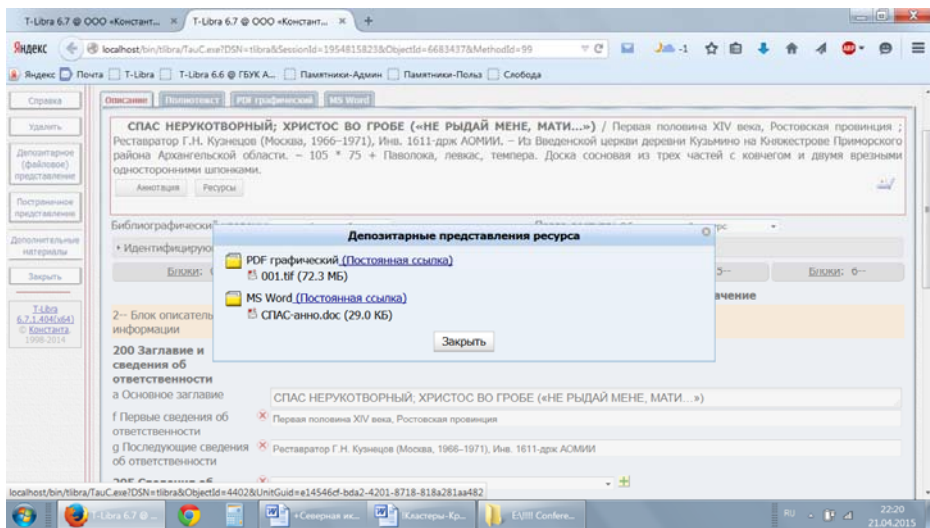


Рис. 2 Активирована кнопка «Ресурсы» электронной карточки. Под этой кнопкой могут находиться различные *представления* ресурса, в различных форматах. Во всплывающем окне видно, что в нашем случае имеются: а) графическое изображение иконы (в формате *.tiff), б) текстовый файл с описанием иконы (в формате *.doc).

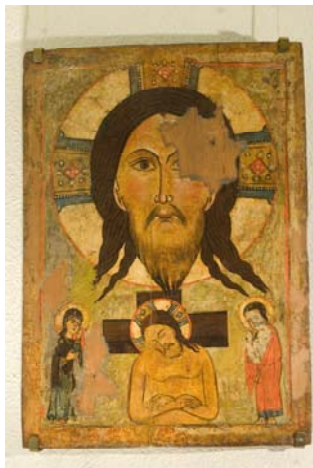


Рис.3. Раскрыт графический файл с изображением иконы «Спас Нерукотворный; ХРИСТОС ВО ГРОБЕ («НЕ РЫДАЙ МЕНЕ, МАТИ...»). Инв. 1611-држ АОМПИ. Из Введенской церкви деревни Кузьмино на Княжестрове Приморского района Архангельской области.

Таким образом, средствами каталога может быть отображено иконографическое описание, а средствами поиска по каталогу найдена необходимая электронная карточка с прикрепленными к ней текстовыми и нетекстовыми (например, графическими) представлениями ресурса.

Далее пользователь нажимает на специальный ярлычок (**Контекстный поиск**), находящийся в электронной карточке, и запускает *гибридный каскадный запрос* одновременно по каталогу и полнотекстовой базе данных.

Алгоритм этого запроса включает в себя: а) предварительную кластеризацию иконографического описания и, на ее основе, б) многослойный абзацно-ориентированный запрос по полнотекстовым ресурсам. Этот алгоритм учитывает также весовые коэффициенты: с *большим* весом – термины из *названия* иконы и *ключевых слов* иконографического описания, с меньшим – из *аннотации*. Подбор весовых коэффициентов производится экспертами. Мы называем этот запрос *квазисемантическим*, поскольку семантика запроса учитывается косвенно, через указанный алгоритм.

Результатом такого запроса является некоторое количество релевантных тематических абзацев из документов электронной библиотеки (см. далее).

4. Использование многослойного запроса для разработки алгоритма квазисемантического гибридного поиска

Для разработки алгоритма квазисемантического гибридного поиска был использован механизм многослойного тематического запроса. Фактически речь шла о предварительном текстологическом эксперименте, результаты которого могли бы быть использованы для разработки квазисемантического гибридного поиска.

В первый слой запроса были включены термины из названия иконы («Спас Нерукотворный»). Во второй – «убрус» и «лик». В третий – «Константинополь» и «Христос». Все эти термины были взяты из соответствующей электронной карточки (иконографического описания).

The screenshot shows the T-Libra 6.7 search interface. The search query is defined by three layers:

- Слой 1: Спас Нерукотворный
- Слой 2: убрус лик
- Слой 3: Константинополь Христос

Search parameters include:

- Минимально необходимое количество слов: 2
- Максимальное расстояние между словами: 5

 The results section shows:

- Всего найдено: 59 абзацев в 44 документах.
- Страницы результата: [1] 2 3 4 5

 A table of results is displayed:

№	Документ	Кол-во абзацев
1.	Кольцова, Т.М. [составитель]. Иконы XIV – начала XX веков в собрании Государственного музейного объединения «Художественная культура Русского Севера»: каталог-путеводитель по экспозиции и фондам музея / Авт.-сост. Т.М. Кольцова (при участии О.Н. Вешняковой, А.В. Насоновой).	7
2.	СПАС НЕРУКОТВОРНЫЙ; ХРИСТОС ВО ГРОБЕ («НЕ РЫДАЙ МЕНЕ, МАТИ...») / Первая половина XIV века, Ростовская провинция; Реставратор Г.Н. Кунецов (Москва, 1966–1971), Инв. 1611-лрк. АОМИИ – Из. Византийской школы. Деревяны. Юсупово, на. Киевское. Примоского	5

On the right side, there is a section titled "Уточнения" (Refinements) with a list of terms and their counts:

- стих (134)
- христос (80)
- икона (63)
- душ (53)
- образ (51)
- спасом (47)
- бог (43)
- том (32)
- время (31)
- песнь (23)
- уж (23)
- век (22)
- церковь (22)

Рис.4. Результат многослойного запроса с терминами из иконографического описания.

Условия запроса (см. рис. 4): обязательный первый слой – с терминами из названия иконы; учитываются любые 2 слоя из трех; дополнительная фокусировка: расстояние между поисковыми терминами в разных слоях не более 5 слов. При этих параметрах запроса найдено 59 абзацев в 44 документах (по базе в 2572 документа). Документы в поисковой выдаче ранжированы по количеству релевантных абзацев.

Далее можно раскрыть любой из найденных абзацев (см. рис.5.)

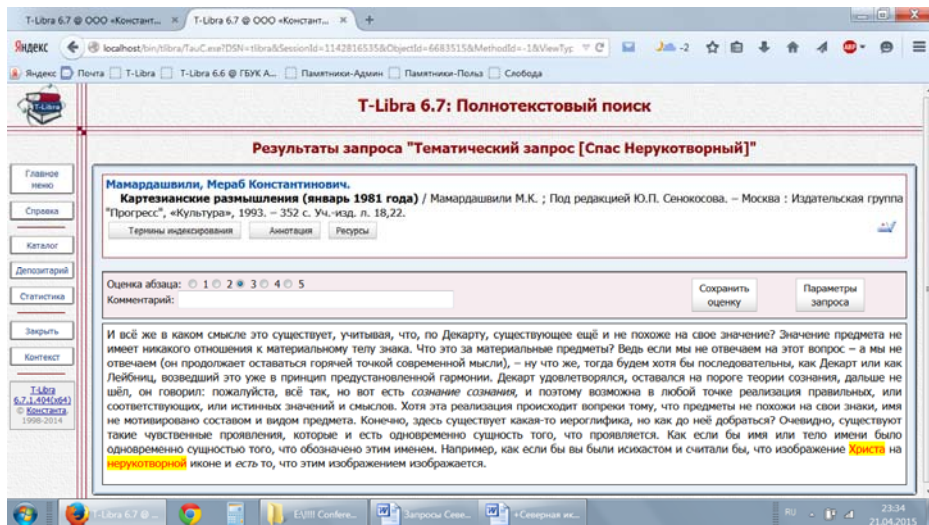


Рис. 5. Раскрыт один из найденных абзацев – из книги М.К.Мамардашвили «Картезианские размышления».

Характерно (для широкого контекстного поиска, о котором и идет речь), что раскрытый на рис. 5 абзац – не из иконографической книги или статьи, а из философской. Такого рода абзацев из общего числа найденных (59) оказалось около 20 (словари, художественная литература, детективы, философия, богословие, история, культурология.). Это обстоятельство позволяет изучать более широкий, чем чисто иконографический, культурно-исторический контекст рассматриваемой иконы.

Аналогичные исследования были проведены на вышеобозначенной ресурсной базе (электронная библиотека объемом около 2500 полнотекстовых документов) еще с 9 иконографическими описаниями (иконы «БОГОМАТЕРЬ ОДИГИТРИЯ»; «ПРЕПОДОБНЫЕ ЗОСИМА И САВВАТИЙ СОЛОВЕЦКИЕ, С ВИДОМ МОНАСТЫРЯ И С ЖИТИЕМ»; «ПРЕПОДОБНЫЙ АНТОНИЙ СИЙСКИЙ, С ЖИТИЕМ» и др. – все из собрания Государственного музейного объединения «Художественная культура Русского Севера», Архангельск). В каждом случае, используя термины из иконографического описания и вариации параметров многослойного запроса, удалось эксплицировать вполне содержательные, с экспертной и пользовательской точек зрения, культурно-исторические контексты иконографических объектов.

Результаты исследований были использованы для разработки алгоритма квазисемантического гибридного поиска, функционально объединяющего

иконографическое описание и полнотекстовые ресурсы. Этот поиск, в свою очередь, будет использован в разрабатываемой в настоящее время (Университет ИТМО, ООО «Константа», ГМО «Художественная культура Русского Севера) интегрированной информационной системе «Северные письма» в музейном информационном пространстве».

5. Коллекции гибридных запросов

Гибридные квазисемантические запросы, эксплицирующие контекстное знание, соотнесенное с иконографическими описаниями, могут быть объединены в *тематические коллекции гибридных запросов*, для которых в настоящее время разрабатывается инструментарий в рамках развития технологии ИС T-Libra. Эти коллекции одновременно могут использоваться и как готовое тематизированное (контекстное) знание, расширяющее состав информационных ресурсов электронной библиотеки, и как пользовательский инструмент для создания и развития аналогичных коллекций [5].

Примечание. Для реализации этого функционала проведена модернизация технологии ИС T-Libra. В частности, в разрабатываемой в настоящее время версии T-Libra 7.x реализована технология single page application [6] (одностраничное приложение), переносщая часть выполняемых функций с сервера на клиент/браузер: сервер посылает данные, оформление которых происходит в браузере пользователя. Это позволяет обеспечить вариативную и интерактивную работу пользователя при организации запросов и одновременно разгрузить сервер и ускорить работу системы в целом, что будет заметно в многопользовательском режиме и работе T-Libra в распределенной среде.

6. Заключение

Взаимодействие поиска по каталогу и полнотекстового поиска позволяет осуществить текстологический эксперимент по извлечению контекстного знания и тех параметров релевантного запроса, которые необходимы для разработки гибридного каскадного квазисемантического запроса «от иконы – к тексту».

На этом пути может быть решена и обратная (более трудная) задача – построен аналогичный запрос «от текста – к иконе».

Полученные в результате запроса фрагменты контекстного знания и готовые поисковые структуры соответствующих запросов могут быть организованы в тематические коллекции – новый вид интерактивных и «динамических» информационных ресурсов, дополняющих «статические» ресурсы электронной библиотеки.

Это, в свою очередь, открывает технологические перспективы для создания *внутри и средствами* электронной библиотеки *тематических выставок*, объединяющих ресурсы разной информационной модальности (текст, графика, аудио, видео и др.) и поисково-презентационные сервисы разного функционального назначения (поиск по каталогу и по полным текстам).

Литература

- [1] Костянян С.А., Куковякин А.В., Ляпин С.Х. Музейная библиотека для поддержки музейной деятельности и интеграции ресурсов в распределенной информационной среде (презентация партнерского проекта) // Конференция АДИТ-2014, г. Выборг Ленинградской области, 20–24 мая 2014 г. URL: <http://adit.ru/sites/default/files/mus-bibl.pdf> (дата обращения: 20.04.2015).
- [2] Кольцова Т.М. Иконы Русского Севера (Вступительная статья) // Иконы XIV – начала XX веков в собрании Государственного музейного объединения «Художественная культура Русского Севера»: каталог-путеводитель по экспозиции и фондам музея / Авт.-сост. Т.М. Кольцова (при участии О.Н. Вешняковой, А.В. Насоновой). Архангельск, Изд.-во «Правда Севера», 2013. С. 4–23.
- [3] Реформатская М.А. Северные письма. М., Изд-во «Искусство», 1968.
- [4] Информационная система T-Libra для создания многофункциональных электронных библиотек с возможностями гибкого тематизируемого многоязычного полнотекстового поиска // URL: <http://demo.tlibra.ru> (дата обращения: 20.04.2015).
- [5] Ляпин С.Х., Куковякин А.В. Тематические коллекции полнотекстовых запросов для изучения контекстного знания (проект Humanitarian) // Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23–25 июня 2015 г. (в печати).
- [6] Single Page Application // URL: https://ru.wikipedia.org/wiki/Single_Page_Application (дата обращения: 19.04.2015).

Northern icon in the multimodal library: to functional integration of iconographic and full-text resources

S. Kh. Lyapin¹, A.V. Kukovyakin²
¹ITMO University, ²Constanta, Ltd

Describes the adaptation of the functionality of the digital library, created on the basis of the T-Libra (development "Constanta" Ltd., Arkhangelsk) to integrate resources of various modalities (text, graphics, audio, video) as part of full-text electronic library (*internal functional integration*). To work with iconographic resources used functionality of the electronic catalog (adapted to standard iconographic description), and hybrid search by full text and catalogue cards. Implemented also quasi-semantic queries “from icons to full text” - using advanced automatic explication of the subject area of the search query and clustering of results.

The report was prepared with the support of a grant from the Russian humanitarian scientific Fund No. 14-03-12017

Keywords: services of full-text search; thematic collection of search queries; clustering of query results; paragraph-oriented request; a frequency-ranked query.