УДК 021:004.9
DOI: 10.15827/2311-6749.17.3.7

АВТОМАТИЗИРОВАННАЯ СИСТЕМА МОНИТОРИНГА ССЫЛОК НА ВНЕШНИЕ СЕТЕВЫЕ РЕСУРСЫ В АКАДЕМИЧЕСКОЙ БИБЛИОТЕКЕ

Н.Е. Каленов, профессор, д.т.н., директор, nek@benran.ru;
М.М. Якшин, научный сотрудник, greycat.na.kor@gmail.com
(ФГБУН Библиотека по естественным наукам Российской академии наук (БЕН РАН), ул. Знаменка, 11, г. Москва, 119991, Россия)

В современных условиях в задачи академических библиотек как центров, обеспечивающих научной информацией ученых, входит предоставление своим пользователям ссылок на внешние научные ресурсы по тематике их исследований, представленные в Интернете. Это могут быть как коммерческие полнотекстовые электронные издания, приобретаемые библиотеками, так и свободно распространяемые полнотекстовые и реферативные материалы. На своих сайтах библиотеки поддерживают различные системы ссылок на такие ресурсы. Чтобы обеспечить максимальный уровень сервиса пользователей, необходимо периодически контролировать доступность этих ресурсов и при необходимости актуализировать ссылки на них. В статье рассматривается автоматизированная система контроля доступности ссылок на ресурсы, разработанная и действующая в Библиотеке по естественным наукам РАН. Система обеспечивает автоматический контроль доступности ресурсов и поддерживает БД, содержащую историю выявленных и исправленных ошибок в ссылках.

Ключевые слова: *доступность, качество обслуживания, информационное обеспечение науки, коллекции ссылок, автоматическая проверка, актуализация интернет-ссылок.*

В современных условиях бурного развития сетевых технологий и электронных научных публикаций, в значительной мере заменяющих печатные издания, коренным образом меняются задачи академических библиотек как центров, обеспечивающих научной информацией ученых. В дополнение к традиционным информационно-библиотечным процессам обслуживания (библиографический поиск, предоставление реферативной информации, выдача печатных изданий и пр.) академические библиотеки поддерживают на своих сайтах ссылки на внешние научные ресурсы по тематике, представляющей интерес для их пользователей. Это могут быть как коммерческие полнотекстовые электронные издания, приобретаемые библиотеками, так и свободно распространяемые полнотекстовые и реферативные материалы. Так, *Библиотека по естественным наукам* (БЕН) РАН, поддерживающая на своем сайте (http://benran.ru) различные ресурсы и сервисы [1], предоставляет своим пользователям непосредственно из интернет-каталогов ссылки на полные тексты доступных книг и журналов [2, 3], а также со специальных страниц «Естественные науки в Интернет» ссылки на указатели ресурсов по основным разделам естественных и точных наук [4].

Эти три группы внешних для БЕН РАН ресурсов имеют принципиальные различия.

1. Доступ к электронным версиям научных журналов приобретается обычно по ежегодной подписке, причем не отдельными журналами, а коллекциями, включающими десятки и сотни журналов. Как правило, поставщики не предоставляют в явном виде URL каждого журнала, а в заключаемых контрактах на доступ присутствует только адрес коллекции; предполагается, что пользователь входит на сайт поставщика, выбирает по названию нужный журнал и работает с ним. В связи с этим для организации доступа к каждому журналу с соответствующей страницы каталога необходимо определять тем или иным образом URL каждого журнала. Для крупных издательств и поставщиков электронных версий журналов (у которых имеются четкие алгоритмы формирования URL своих журналов) ссылки в БЕН РАН формируются автоматически или полуавтоматически специальными скриптами; для отдельных журналов, не входящих в крупные коллекции, они формируются вручную. Формально контрактами предусматривается поддержка постоянного доступа к журналам, однако стабильность URL журналов никем не гарантирована. В настоящее время в интернет-каталоге журналов БЕН РАН присутствуют ссылки на электронные версии 6 058 наименований отечественных и зарубежных журналов (в целом, в каталоге отражены более 684 000 выпусков 8 054 наименований журналов).
2. Адреса электронных версий монографий при покупке доступа к книжным коллекциям (обычно они приобретаются «навечно»), как правило, присутствуют в явном виде в качестве приложений к контракту. Они вводятся в каталог БЕН РАН либо в пакетном режиме (если с поставщиком удается договориться о передаче библиографических описаний книг в электронном виде), либо в индивидуальном порядке при формировании библиографических описаний изданий, отражаемых в интернет-каталоге. URL книг, представленных в сети в открытом доступе (в первую очередь, это относится к изданиям, подготовленным при поддержке РФФИ и РГНФ), определяются и вводятся в каталог

сотрудниками БЕН РАН. Так же, как и в случае с журналами, поставщики коммерческих версий декларируют наличие доступа к электронным книгам (но не стабильность URL), доступ же к некоммерческим ресурсам (а тем более стабильность их URL) не гарантирован никем. В интернет-каталоге БЕН РАН в настоящее время имеются ссылки на более чем 21 100 электронных версий книг. доступных пользователям библиотеки (общий объем каталога – около 200 000 наименований).

3. Тематические подборки ссылок на указатели интернет-ресурсов по определенным разделам науки формируются специалистами БЕН РАН на основании анализа мирового информационного пространства. Путем просмотра справочников и обработки запросов в Интернете выявляются организации, осуществляющие сбор и систематизацию данных по определенному научному направлению. Описание ресурсов каждой такой организации составляется и загружается вместе с ее URL в соответствующий раздел «Естественные науки в Интернет» сайта БЕН РАН. Очевидно, что «стабильность» URL ни в коей мере не зависит от сотрудников библиотеки, и никакого рычага воздействия на нее не существует. В настоящее время раздел сайта БЕН РАН «Естественные науки в Интернет» содержит более 100 ссылок на «узловые» ресурсы по основным разделам естественных и точных наук.

Всеми перечисленными ресурсами пользуется значительное количество (десятки тысяч) пользователей, и время от времени БЕН РАН получает от них жалобы на то, что они не могут открыть тот или иной ресурс по указанным ссылкам. Таким образом, проблема мониторинга и актуализации ссылок на внешние ресурсы является для БЕН РАН, предоставляющей доступ к ним в режиме 24/7 (семь дней в неделю круглые сутки), достаточно актуальной.

Проблема мониторинга подконтрольных локальных систем в целом хорошо изучена, и для ее решения существует целый ряд готовых решений, таких как Zabbix [5] или Nagios [6]; существуют семейство протоколов мониторинга SNMP [7] и семейство стандартов принятия решений по инцидентам, происходящим внутри организации. Для систем, являющихся внешними по отношению к организации (в данном случае – библиотеке), проблема все еще стоит остро.

Для таких объемов ссылочной массы использовать готовые решения, принятые для мониторинга локальной сети, неэффективно. Подобные решения требуют заводить каждый хост как отдельный объект в специальной БД. Если идти этим путем, потребуется написание процедуры импорта ссылок на ресурсы как объектов, а в дальнейшем – нетривиальное обновление описания таких объектов в БД при каждом очередном изменении набора ссылок. Кроме того, готовые решения эффективно работают с относительно небольшим количеством хостов – порядка сотни, дальнейший рост БД хостов приводит к необходимости построения распределенной системы на нескольких серверах [5, 6], что в данном случае нецелесообразно.

В БЕН РАН было принято решение создать собственную систему скриптов, которая реализовывала бы описанную задачу. Общая логика работы системы показана на рисунке 1.

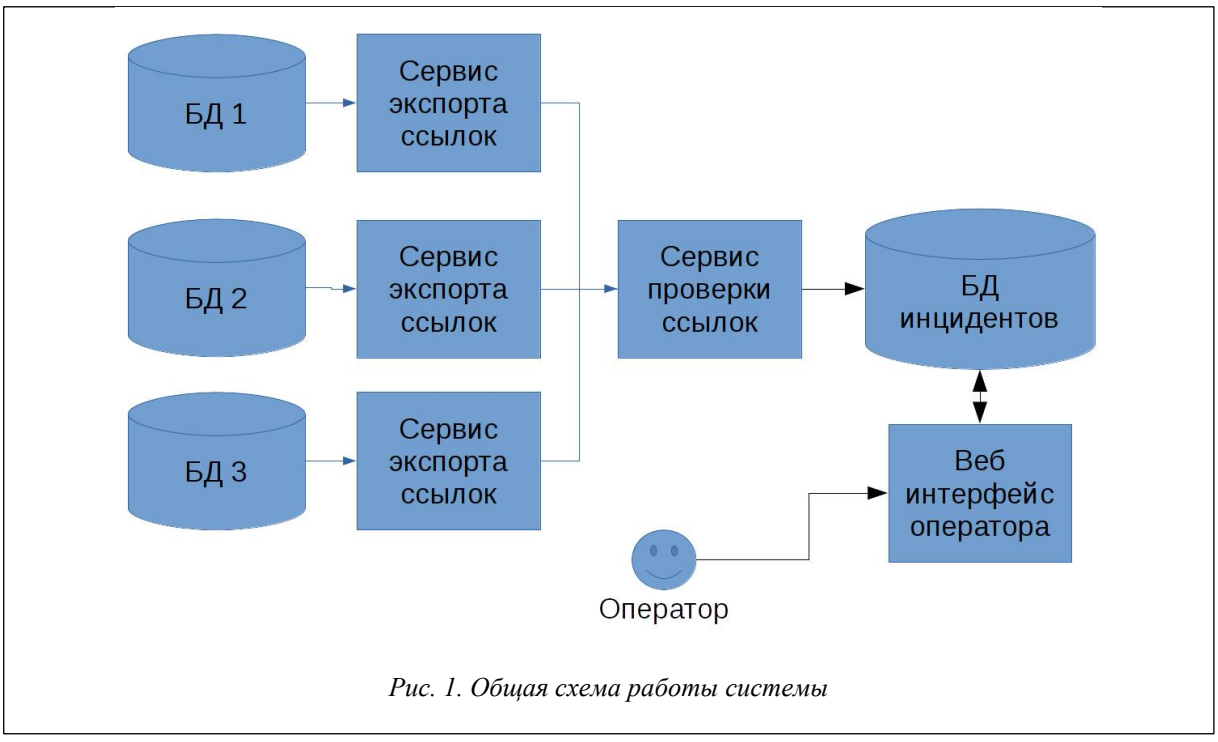

Рис. 1. Общая схема работы системы

Система скриптов имеет следующую архитектуру:

– скрипт, выгружающий данные в системно-зависимом виде (таблицы, XML, JSON, YAML или др. форматы);

– скрипт, преобразующий системно-зависимый вид к общему формату (массиву пар «идентификатор–ссылка»), который будет обрабатывать скрипт проверки;

– собственно единый скрипт проверки ссылок.

Для каждой строки массива пар скрипт проверки проводит ту же самую процедуру, что и браузер пользователя при попытке обращения к ресурсу, а именно:

– из URL выделяется имя сервера;

– имя сервера преобразуется в IP-адрес с помощью запроса к DNS-серверу (ресолвинг имени); на этом этапе можно обнаружить, что DNS-сервер не отвечает или отвечает, но имя не существует;

– производится подключение к серверу по TCP; на этом этапе можно обнаружить, что http-порт закрыт, нет ответа от сервера или есть какие-то другие сетевые проблемы;

– осуществляется GET-запрос по протоколу HTTP к серверу и анализируется ответ на него; на этом этапе должны получить один из стандартизированных http-кодов ответов [8], но можно и не получить ничего или получить какую-то сетевую ошибку; наиболее часто встречающиеся коды HTTP:

200 – OK, все в порядке;

301 – Moved Permanently, ресурс навсегда изменил свой адрес на новый;

302 – Found, ресурс временно изменил свой адрес;

400 – Bad Request, сервер не смог обработать запрос из-за синтаксической некорректности адреса;

404 – Not Found, ресурс не был найден;

500 – Internal Server Error, внутренняя ошибка программного обеспечения сервера;

– если получен ответ 301 или 302 с адресом, куда переместился ресурс, продолжаем запросы новых ресурсов, повторяя алгоритм с самого начала;

– для предотвращения бесконечного цикла вводится счетчик итераций алгоритма, ограничивающийся сверху каким-то разумным числом итераций (например 50); если число редиректов превышает это число, считаем это ошибкой.

Для реализации данного алгоритма задействуется библиотека HTTP-клиента curl. Коды HTTP возвращаются численно (то есть 200 или 302), а дополнительные ошибки соответствуют внутренней кодировке curl и отображаются как curlXX: например curl56.

Система запускается с помощью cron [8] раз в неделю (в ночь с субботы на воскресенье), результаты агрегируются с помощью стандартных утилит join, sort, uniq [8] и записываются в лог-файлы, которые затем импортируются тремя потоками (соответствующими вышеперечисленным видам ресурсов – журналы, книги, ссылки) в специальную БД Check на платформе SciRus (разработка БЕН РАН [9]).

Структура БД Check включает по одной таблице на каждый из обрабатываемых потоков. Все таблицы одинакового формата, записи в них имеют следующие поля:

– идентификатор проверяемого ресурса во входном потоке (обязательное, формируется для каждого потока отдельно, уникален для каждого ресурса);

– проверяемый URL (обязательное, выбирается из лог-файла);

– коды ошибок (обязательное, выбирается из лог-файла);

– первая дата обнаружения ошибки;

– контрольная дата (вводится оператором, когда требуется ожидание какого-то длительного действия; например, отослан запрос, требуется получить на него ответ) для напоминания о необходимости завершить такое ожидание; используется для формирования соответствующего списка «проблемы, ожидающие решения сегодня»;

– статус записи, принимающий одно из следующих значений: не обработано, повторная проблема, исправлено, запись удалена, направлен запрос, ожидание решения;

– комментарии (для информации о совершаемых в отношении записи действиях).

Алгоритм работы системы

Для каждого проверяемого URL (строки лог-файла) выполняются следующие операции.

1. По итогам анализа HTTP-кодов и других кодов ошибок вычисляется статус «есть проблема»/«нет проблемы».

2. Из БД Check запрашивается запись с идентификатором ресурса, равным идентификатору ресурса в строке лога.

3. Если запись не найдена, то:

– если статус у соответствующей записи лог-файла «есть проблема», в БД Check создается новая запись со статусом «не обработано» и осуществляется переход к следующей записи лог-файла;

– если статус у записи лог-файла «нет проблем», осуществляется переход к следующей записи без корректировки БД Check.

4. Если запись найдена, то:

– если значение ее статуса «исправлено» или «запись удалена», оно изменяется на «повторная проблема» и осуществляется переход к следующей записи лог-файла;

– если значение ее статуса «не обработано», «повторная проблема», «направлен запрос», «ожидание решения», то оно не изменяется и осуществляется переход к следующей записи лог-файла.

В любом случае, если произошло изменение записи, все остальные поля заполняются из строчки лога. После этого обрабатывается следующая запись лог-файла.

Сервисы экспорта ссылок реализованы на языках и с помощью библиотек, соответствующих используемым СУБД в конкретных группах внешних ресурсов и обычно являются частью внешнего ресурса. Сервис экспорта ссылок для тематических коллекций ссылок (которые фактически не имеют СУБД, а представлены просто рядом HTML-файлов) реализован на shell script с помощью библиотеки curl. В качестве системы управления запуском скриптов используется GNU make, вызываемый через cron [8]. В качестве СУБД используется MariaDB v10.0.30. Для реализации пользовательского интерфейса использована платформа SciRus [9], которая, в свою очередь, разработана с использованием языка Ruby v2.4 и фреймворка Ruby on Rails v4.1. Система поддерживается и доступна авторизованным пользователям на сервере БЕН РАН по адресу http://check.labs.benran.ru.

Процесс работы с системой

Общая блок-схема действий операторов (сотрудников БЕН РАН, обеспечивающих мониторинг ссылок) представлена на рисунке 2.

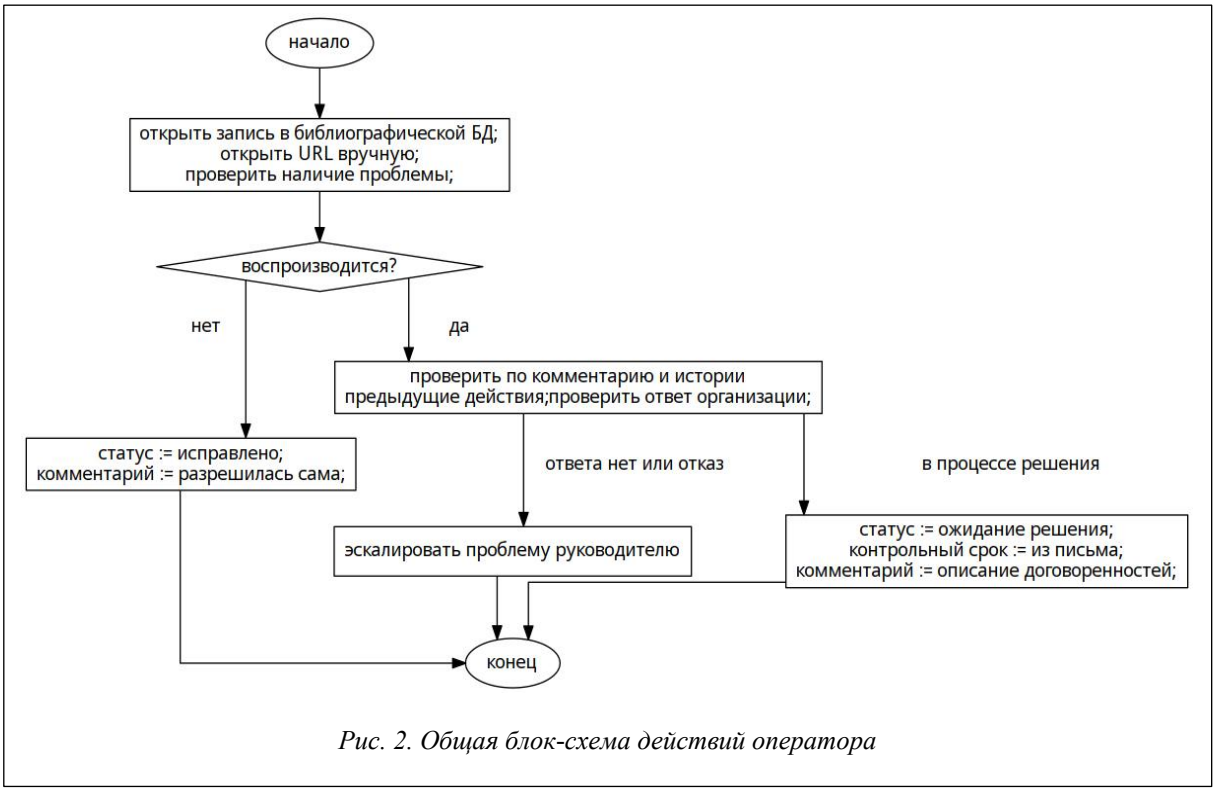

Рис. 2. Общая блок-схема действий оператора

Сформированная в ночь с субботы на воскресенье информация доступна для обработки в течение следующей недели; историю работы с информацией за прошедшие недели можно посмотреть в системе по каждой записи во вкладке «История».

Войдя по своему паролю в систему, оператор попадает на начальную страницу, выбирает вкладку, соответствующую потоку, с которым он собирается работать (журналы, книги, ссылки на сайте), и начинает анализировать строки открывшегося списка «Новые проблемы» (рис. 3). Этот список включает все записи со статусом «не обработано» или «повторная проблема».

Каждая строка списка является активной ссылкой, при переходе по которой на экран выводится информация о конкретной «проблемной» записи (рис. 4).

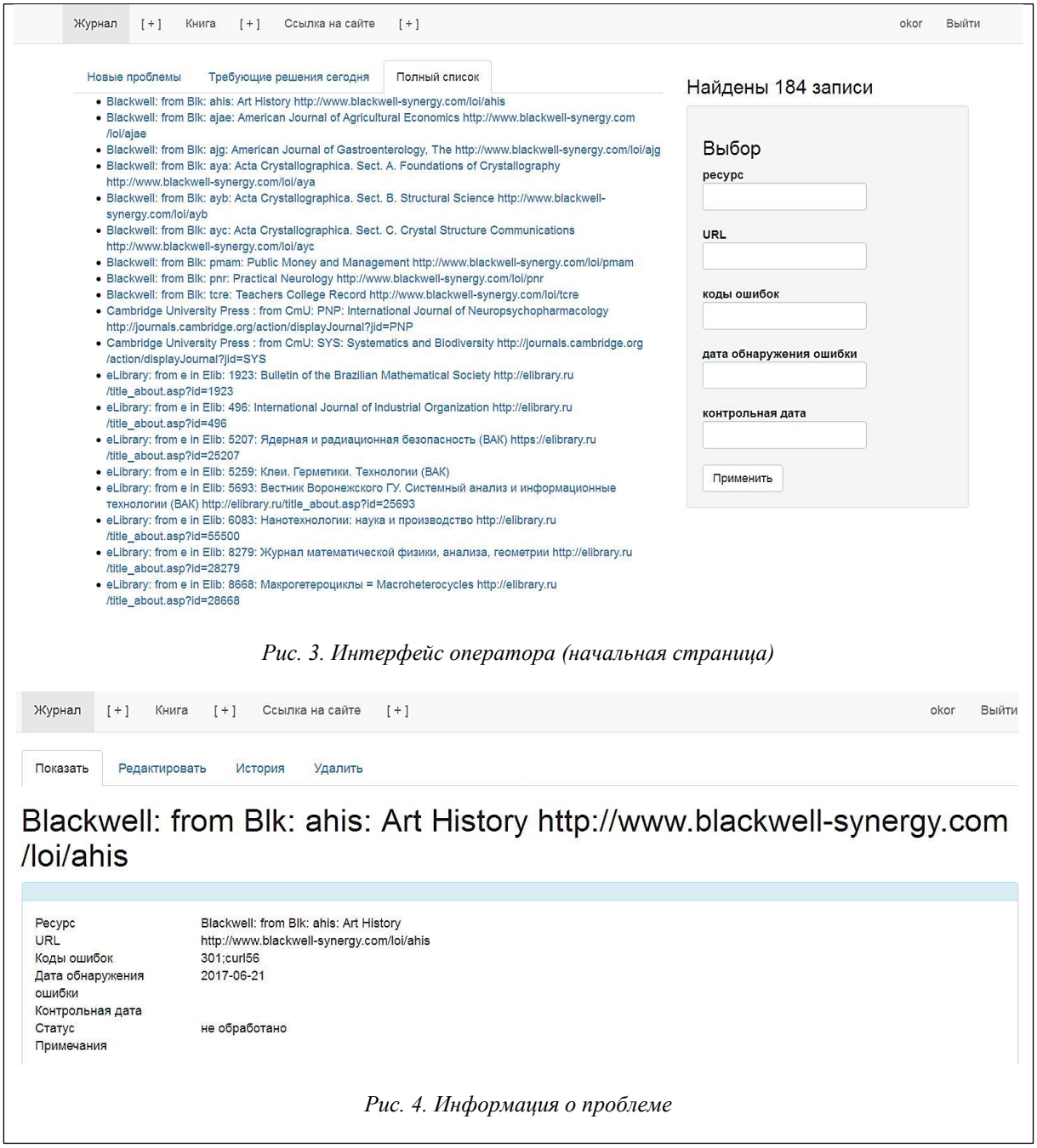

Рис. 3. Интерфейс оператора (начальная страница)

Рис. 4. Информация о проблеме

Оператор проверяет ее доступность и далее действует в соответствии с логикой, представленной на рисунке 2. Закончив работу со списком «Новые проблемы», оператор переходит к списку проблем, «требующих решения сегодня». В этот список выводятся записи, в поле «контрольная дата» которых содержится текущая дата, а поле «статус записи» имеет значение «направлен запрос» или «ожидание решения».

Наряду с пакетной загрузкой информации система допускает ручной ввод данных о проблемных ссылках, обнаруженных в промежутке автоматического опроса сайтов. Для ввода новой записи необходимо нажать на значок «+» рядом с названием потока, после чего откроется соответствующая страница (рис. 5).

Система прошла опытную эксплуатацию и в течение месяца работает в БЕН РАН в технологическом режиме. За это время было скорректировано более 200 изменившихся ссылок на журналы и книги, направлены запросы поставщикам, удалены ссылки на несуществующие ресурсы. Общее количество некорректных ссылок уменьшилось почти на 25 % – с 719 в начале работы системы до 441 к моменту написания данной статьи.

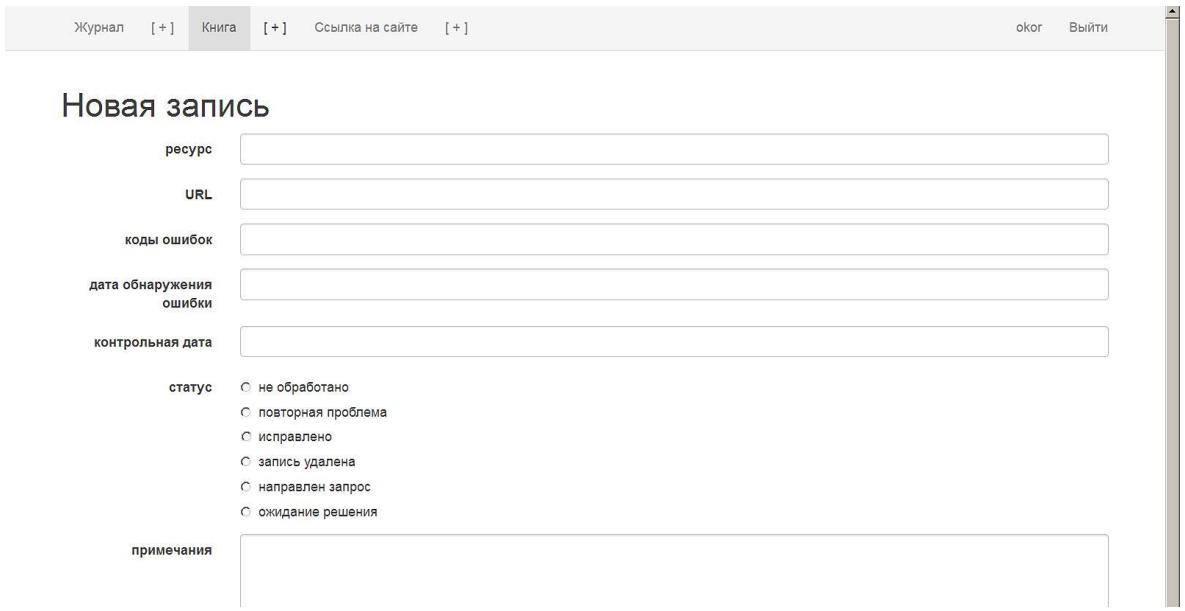

Рис. 5. Окно для ручного ввода данных об ошибке

Литература

1. Власова С.А., Каленов Н.Е. Информатика в академической библиотеке // Системы и средства информатики. 2016. Т. 26. № 3. С. 162–178.

2. Власова С.А., Каленов Н.Е. Роль каталогов научных библиотек в задачах информационного сопровождения научных исследований // Информационные процессы. 2014. Т. 14. № 3. С. 232–241. URL: http://www.jip.ru (дата обращения: 24.07.2017).

3. Соловьева Т.Н. Ссылки в Интернет-каталоге журналов БЕН РАН // Информационное обеспечение науки: новые технологии: сб. науч. тр. 2015. С. 249–253.

4. Глушановский А.В., Каленов Н.Е. Библиотека по естественным наукам РАН как политематический центр предоставления электронной информации для ученых и специалистов РАН // Межотраслевая информационная служба. 2014. Т. 4. № 169. С. 16–18.

5. Andrea Dalle Vacche, Stefano Kewan Lee. Mastering Zabbix. Packt Publishing Ltd., 2013, 358 p.

6. David Josephsen. Building a Monitoring Infrastructure with Nagios. Prentice Hall., 2007, 600 p.

7. Douglas Mauro, Kevin Schmidt. Essential SNMP. O'Reilly Media, Inc., 2005, 442 p.

8. Eleen Frisch. Essential System Administration: Tools and Techniques for Linux and Unix Administration. O'Reilly Media Inc., 2002, 1178 p.

9. Якшин М.М. Развитие платформы SciRus // Информационное обеспечение науки: новые технологии: сб. науч. тр. 2015. С. 203–207.

10. Fielding R. et al. RFC 2616: Hypertext Transfer Protocol HTTP/1.1. The Internet Society, 1999. URL: https://tools.ietf.org/html/rfc2616 (дата обращения: 24.07.2017).