УДК 021:004.9
ББК 78.347.41(2Рос)

Мониторинг доступности внешних электронных ресурсов в библиотечной среде

Якшин М. М., Каленов Н. Е. (БЕН РАН, Москва)

В задачи академических библиотек как центров, обеспечивающих научной информацией ученых, в современных условиях входит предоставление своим пользователям ссылок на полнотекстовые ресурсы по тематике их исследований. Это могут быть как коммерческие полнотекстовые электронные издания, приобретаемые библиотеками, так и свободно распространяемые материалы. На своих сайтах библиотеки поддерживают различные системы ссылок на такие ресурсы. Чтобы обеспечить максимальный уровень сервиса пользователей, необходимо периодически контролировать доступность указанных ресурсов и при необходимости актуализировать ссылки на них. В статье рассматривается технология автоматического контроля доступности ссылок на ресурсы, реализованная в БЕН РАН.

Ключевые слова: доступность, качество обслуживания, коллекции ссылок, автоматическая проверка, актуализация ссылок, http.

In the modern context, academic libraries, acting as scientific information hubs, are sought to provide links to pertinent full-text resources to their users. These can be either commercial full-text electronic documents that the libraries are subscribed to, or any freely available content. Libraries maintain various link collections on their websites. To ensure best possible user service level, one should do periodic availability checks for referenced resources, and, if necessary, correct the links. This article overviews automatic resource availability checking technology, developed at LNS RAS.

Keywords: availability, service level, URL collections, automated availability checking, URL updates, http

В задачу современных научных библиотек входит предоставление доступа своим пользователям к электронным ресурсам. Как правило, такие ресурсы располагаются в IP-сетях (локальных или глобальных)

и, за редкими исключениями, для доступа используется протокол http или его вариант с шифрованием и проверкой подлинности – https.

Так как библиотека стремится повышать качество обслуживания читателей и по возможности предоставлять доступы к ресурсам все время своей работы, для всех таких электронных ресурсов (и локальных, расположенных в подконтрольных сетях библиотеки, и удаленных, расположенных на серверах сторонних организаций) становится актуальна проблема мониторинга доступности. Все используемое в сетях оборудование не является абсолютно надежным. Существует целый ряд проблем, из-за которых оборудование может перестать выполнять свои функции, и электронный ресурс станет недоступным для пользователей. Среди возможных причин этого, в частности, могут быть:

- ошибка в программном обеспечении;
- физический износ оборудования;
- неисправности систем охлаждения и кондиционирования и следующий за этим перегрев техники;
- ошибки на сетевых магистралях, в том числе физические, – повреждение передающей среды, обрыв кабеля, повреждение разъема и т.п.;
- параллельно выполняющиеся процессы, входящие в состояние «клинча»;
- солнечная активность и космические лучи, приводящие к нарушению работы подсистем ОЗУ.

Для подконтрольных локальных систем проблема мониторинга в целом хорошо изучена, и для ее решения существует большой ряд готовых решений, таких как Zabbix [4] или Nagios [5]; существует семейство протоколов мониторинга SNMP [6] и семейство стандартов принятия решений по инцидентам, происходящим внутри организации. В свою очередь, для систем, являющихся внешними по отношению к организации (в данном случае библиотеке), проблема все еще стоит остро. В целом все электронные ресурсы, на которые предоставляют ссылки библиотеки, можно разделить условно на 3 группы:

- электронные версии научных журналов и баз данных, доступные по ежегодной подписке библиотеке (далее – «журналы»);
- электронные версии монографий и прочих печатных изданий,

доступ к которым приобретен «навечно» (далее – «монографии»);
- тематические подборки ссылок на интернет-ресурсы в рамках определенных научных тематик, собранные и ранжируемые сотрудниками библиотеки (далее – «подборки»).

Такая группировка уместна, поскольку, как правило, все три группы таких ссылок обрабатываются тремя различными системами, формируются в результате независимых технологических процессов, курируются различными сотрудниками и обеспечиваются отдельными юридически значимыми соглашениями. При этом библиотека заинтересована в обеспечении непрерывной работы всех трех групп ссылок одинаково.

Процессы построения и поддержания этих групп ссылок могут существенно различаться. Рассмотрим подробнее (табл. 1), что является первоисточником информации для формирования ссылок, как они формируются, обновляются и что должно происходить при недоступности ссылки из группы.

Таблица 1.

Классификация ссылок

	Журналы	Монографии	Подборки ссылок
Первоисточник	Списки из договоров с издательствами	Списки из договоров с издательствами	Экспертные знания научных сотрудников библиотеки
Формирование ссылок	Для крупных издательств и распространителей электронных версий журналов (у которых имеются четкие алгоритмы формирования URL своих журналов) ссылки формируются автоматически или полуавтоматически специальным скриптом; для отдельных журналов, не входящих в крупные коллекции, они формируются вручную	Список монографий обрабатывается вручную или извлекается автоматически скриптом из файла, получаемого от поставщика	Вручную сотрудниками библиотеки

Окончание таблицы 1

Обновляемость	Типично – 1 раз в год для каждого поставщика	В соответствии с договором, но, как правило, ссылки на монографии можно считать условно неизменными	Как правило, нерегулярно в зависимости от задач, ставящихся обслуживаемыми научными сотрудниками
Возможные действия при недоступности	Инициирование контакта с издательством для прояснения обстоятельств, замена ссылки (возможно, с исправлением исходного механизма генерации ссылок)	Инициирование контакта с издательством для прояснения обстоятельств, при необходимости замена ссылки	Удаление ссылки из каталога, попытка найти новые ссылки по заданной тематике или связаться с организациями, представлявшими «пропавшие» ссылки, для выяснения

Таким образом, несмотря на значительно отличающиеся технологические процессы обработки, для всех трех групп можно выделить некие общие свойства:

- все группы работают со ссылками;
- все ссылки – внешние и неподконтрольные напрямую библиотечному персоналу;
- как правило, к ссылке привязан идентификатор какой-то сущности (журнала, документа, названия объекта в базе ссылок и т.д.) и для инициирования работ по недоступности нужно эту связь со ссылкой сохранить; одной лишь ссылки, как правило, недостаточно для восстановления информации о том, что это был за объект, в какой базе искать его описание, какая организация должна быть ответственна за его функционирование и т.д.

В БЕН РАН эти три группы ссылок присутствуют в следующих ресурсах, представленных на сайте Библиотеки (http://benran.ru):

- каталог журналов [3];
- каталог книг [1];
- раздел «Естественные науки в сети Интернет» [2].

Этими 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.

Чтобы минимизировать количество подобных обращений и оперативно принимать меры по исправлению ситуации с недоступными ресурсами, в БЕН РАН было принято решение организовать регулярный мониторинг всех предоставляемых ею ссылок (работа в этом направлении поддерживается грантом РФФИ (проект 16-07-00450)).

По состоянию на конец августа 2016 г. ресурсы БЕН РАН содержали следующее количество ссылок:

- каталог журналов: 7736 ссылок;
- каталог книг: 20831 ссылка;
- тематический каталог: 103 ссылки;

Для таких объемов ссылочной массы использовать готовые решения, принятые для мониторинга локальной сети, неэффективно. Такие решения требуют заводить каждый хост как отдельный объект в специальной базе данных. Если идти этим путем, то потребуется написание процедуры импорта этих ссылок 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 [4, 5], что в данном случае нецелесообразно.

В БЕН РАН было принято решение создать собственную систему скриптов, которая бы реализовывала описанную задачу. Как было показано выше, все три группы ресурсов обладают определенной общностью, поэтому архитектурно система скриптов выполнена в следующем виде:

1) скрипт, выгружающий данные из системы в системно-зависимом виде (таблицы, XML, JSON, YAML или другие форматы);

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

3) собственно единый скрипт проверки ссылок.

Для каждой строки массива пар скрипт проверки проводит ту же самую процедуру, которую проводит 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 мы можем обнаружить, что DNS-сервер не отвечает или отвечает, но имя не существует;
- производится подключение к серверу по TCP – на этом этапе мы можем 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 мы должны получить один из стандартизированных http-кодов ответов [8], но можем и не получить ничего или получить какую-то сетевую ошибку; наиболее часто встречающиеся коды HTTP:
  - 200 – OK, все в порядке;
  - 301 – Moved Permanently, ресурс изменил свой адрес на новый навсегда;
  - 302 – Found, ресурс изменил свой адрес временно;
  - 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 мы получили 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 этого алгоритма задействуется библиотека HTTP-клиента curl. Коды HTTP возвращаются численно (т.е. 200 или 302), а дополнительные ошибки соответствуют внутренней кодировке curl и отображаются, как curlXX: например, curl56. Результат работы скрипта выводится в таблицу следующего вида (табл. 2):

Таблица 2.

Вид результирующей таблицы

ID ресурса	Коды результатов проверки	Title последней полученной страницы (для визуального контроля соответствия содержимого)
http://vlib.org/	200	The WWW Virtual Library
http://www.img.ras.ru/	302;301;200	Институт молекулярной генетики РАН
http://bubl.ac.uk	404	Not Found

Вся система запускается с помощью cron [7] раз в неделю, результаты агрегируются с помощью стандартных утилит join, sort, uniq [7]. Результаты работы системы применительно к ссылкам в каталоге журналов представлены в табл. 3.

Таблица 3.

Результаты проверки корректности ссылок в каталоге журналов

Количество ссылок	Код(ы) результатов проверки
2748	302;302;200
1632	302;200
1563	200
631	301;200
439	301;302;200
374	400
254	302;302;302;200
27	301;302;302;200
14	404
14	301;404
13	301;301;302;302;curl56
9	302;302;500
9	301;301;200
3	302;301;200
2	301;301;301;200
1	302;404
1	302;302;404
1	301;302;302;302;302;302;200
1	301;302;301;200

В правой колонке таблицы – результаты, в левой колонке таблицы – количество ссылок с такими результатами. Рассмотрим

наиболее типовые примеры и расшифровку результатов работы проверочного скрипта:

- 302;302;200 – на первоначальный запрос ссылки система предложила перейти по второй ссылке и ответила с кодом HTTP [8] 302 Found, что означает временный редирект на вторую ссылку; скрипт прошел по второй ссылке, где обнаружил третью ссылку (с таким же кодом 302); пройдя по третьей ссылке, скрипт наконец-то получил искомую страницу с успешным кодом получения HTTP 200 OK. Как правило, такой сценарий характерен для систем, требующих установления какой-нибудь сессии и использующих для этого промежуточные страницы (например elibrary.ru). Такие ссылки следует считать работающими правильно;
- 302;200 – аналогичная ситуация, но на один редирект меньше. Ссылка с точки зрения пользователя работает, поводов вмешиваться нет;
- 200 – наилучшая ситуация, ссылка работает сразу же, пользователь получает запрошенный документ;
- 301; 200 – скрипт получил редирект с кодом 301 Moved Permanently. Согласно стандарту [8] это означает, что ресурс навсегда изменил свой адрес, и всем, кто хранил ссылки, нужно впредь использовать новую ссылку. На практике же, к сожалению, далеко не все следуют стандарту и иногда используют 301 и для «временных» редиректов, для которых по стандарту нужно использовать 302. Эти ссылки желательно просмотреть вручную (скорее всего, там обнаружится некая системность) и принять решение после анализа: это может быть как легитимная ситуация перманентной замены ссылки (например, одно издательство объединилось с другим, и все ссылки теперь должны приводить на сайт объединенного издательства), так и ошибка конфигурации сервера, с которой придется мириться;
- 400 – ответ с кодом HTTP 400 Bad Request. Это, как правило, означает серьезную проблему в формировании URL – какие-то недопустимые символы, пробелы и т.п. Так как (см. выше) первоначальное формирование ссылок для журналов в большинстве случаев автоматическое или полуавтоматическое, такой код ошибки будет означать системную проблему в скрипте формирования ссылок на журналы/книги;

- 404 – ответ с кодом 404 Not Found означает, что ресурс более не может быть найден по данному URL, но при этом домен (и, скорее всего, организация-поставщик) существует, сеть функционирует нормально, удаленный сервер настроен на прием запросов – но запрошенного объекта нет. Ситуация требует ручного вмешательства и выяснения, что же произошло с объектом;
- 301; 404 – комбинация из перманентного редиректа 301 Moved Permanently и 404 Not Found. Как правило, это означает, что объекты переехали с сайта одной организации на сайт другой (преемницы), но через какое-то время исчезли (ввиду реорганизации сайта или чего-то такого). Ситуация требует ручного анализа и поиска объектов (если они еще существуют) или удаления ссылки;
- 301;301;302;302;curl56 – одна из наиболее сложных ситуаций. Скрипт прошел по 4 редиректам, два из которых перманентные (301)? и два – временные (302), но в итоге работа с пятым по счету запросом закончилась совсем плохо (код ошибки curl56: Failure in receiving network data), т.е. проблемы даже не на уровне протокола HTTP, а со связанностью сети в указанном направлении. Необходимо вмешательство и анализ проблемы;
- 302;302;500 – после двух редиректов третий сервер ответил 500 Internal Server Error. Обычно это означает какую-то серьезную проблему с сервером приложений на удаленной стороне. Проблема вполне может быть временной (но требовать вмешательства системного администратора), поэтому ее как минимум нужно перепроверить вручную и попытаться связаться с организацией.

Первичные проверки по всем трем группам ссылок БЕН РАН показали следующие результаты (табл. 4).

Таблица 4.

Сводные результаты проверки качества ссылок

	Всего ссылок	Проблем	Доля ошибок
Журналы	7736	426	5,50%
Монографии	20831	185	0,88%
Подборки ссылок	103	8	7,77%

В настоящее время ведется работа над выяснением причин недоступности ресурсов и исправлением обнаруженных проблем.

Литература

1. Власова С. А. Новая версия каталога книг и продолжающихся изданий Библиотеки по естественным наукам РАН / С. А. Власова, Н. Е. Каленов // Информационное обеспечение науки: новые технологии : сб. науч. тр. / отв. ред. П. П. Трескова ; сост. О. А. Оганова, М. А. Уласовец. Екатеринбург, 2014. С. 122–127.

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

3. Соловьева Т. Н. Ссылки в интернет-каталоге журналов БЕН РАН / Т. Н. Соловьева // Информационное обеспечение науки: новые технологии : сб. науч.тр. / Н. Е. Каленов, В. А. Цветкова (ред.). Москва,2015. С. 249–253.

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

5. David Josephsen. Building a Monitoring Infrastructure with Nagios / David Josephsen . Prentice Hall. 2007. 255 p.

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

7. Æleen Frisch. Essential System Administration: Tools and Techniques for Linux and Unix Administration / Æleen Frisch. O'Reilly Media. Inc. 2002. 1178 p.

8. RFC 2616: Hypertext Transfer Protocol HTTP/1.1[Electronic resource] / R. Fielding, [et al]. The Internet Society. 1999. https://tools.ietf.org/html/rfc2616.