Библиотеки в "облаках": проблемы обмена данными

Е.В.Ковязина
Институт вычислительного моделирования СО РАН
elena@icm.krasn.ru

Аннотация

Проникновение облачных вычислений в библиотеки обусловлено растущими объемами данных, количеством обращений, потребностью в новых сервисах для читателей и еще целым рядом задач. Возросшее количество информационных систем и сервисов обострило проблему взаимообмена данными о документах, с которыми работают такие системы. Распределенное хранение данных, доступ к которым обеспечивался на основе стандартизованных протоколов и обменных форматов, сменилось «облачными» моделями, основанными на центрах обработки данных (ЦОД). Однако проблема обмена данными между такими системами остается для пользователя большой проблемой, зачастую неразрешимой.

GRID-вычисления. На начальном этапе библиотечной автоматизации корпоративные библиотечные системы строились как системы распределенные, в основе которых лежала концепция GRID-вычислений. Обратимся к определению: «GRID – система, которая связана с интеграцией, виртуализацией и управлением услугами и ресурсами в распределенной, гетерогенной среде, которая поддерживает коллекции пользователей и ресурсов (виртуальных организаций) в традиционных административных и организационных доменах (реальных организаций)». В таких системах каждый из хозяев доменов сам определял техническую и организационную структуру домена, состав и права доступа к информационным ресурсам, регламент работы и обслуживания пользователей и т.д., но в соответствии с оговоренными стандартами и правилами корпораций, в которых он состоял. Для корпоративных библиотечных систем были определены как обязательные стандарты сетевые протоколов Z39.50 и коммуникативный формат rusmarc. С помощью этих стандартных средств была решена не только проблема интеграции ресурсов, но и обмен данными между участниками корпораций на основе некоторой стандартной их структуры. Пользователю в таких системах обеспечивался доступ к данным через единый пользовательский интерфейс, с помощью которого он получал некоторый виртуальный информационный ресурс, выглядящий как единая база данных. Отметим, что поддержка подобных систем была достаточно обременительной для хозяев данных. Требовалось наличие и постоянное обновление парка вычислительной техники и программного обеспечения, а также содержание дорогостоящего квалифицированного IT-персонала. С ростом объемов данных проявлялись дополнительные проблемы, связанные с сохранностью данных, их архивированием и защитой от внешних атак [1-2]. Не секрет, что небольшие организации не могли позволить себе таких расходов, что тормозило дальнейшее распространение подобных систем. Сложившаяся ситуация создала предпосылки для предпочтения централизованных решений.

Облачные вычисления. Концепция облака появилась на фоне развития виртуализации в Интернет и повсеместного распространения web-сервисов. Точного

определения облачных вычислений все еще нет. Наиболее часто встречающееся определение, приведенное также и в Википедии: «Облачные вычисления это модель предоставления повсеместного и удобного сетевого доступа по требованию к общему пулу конфигурируемых вычислительных ресурсов (например, сетей, серверов, систем хранения, приложений и сервисов), которые могут быть оперативно предоставлены и освобождены с минимальными эксплуатационными затратами и/или обращениями к провайдеру услуг». Исторически облако - это симбиоз виртуализации и web-сервисов. Образно облако – это стена между пользователями и провайдерами, которая скрывает то, что происходит на стороне провайдера, предоставляя пользователю только необходимый ему набор услуг [3]. В печати появились публикации, посвященные разграничению grid- и облачных вычислений, выявлению принципиальной разницы между двумя технологиями, например, [4-7]. Отмечается, что разница между grid и облачными вычислениями достаточно условна, так как налицо взаимопроникновение технологий. Grid-системы зачастую реализуются через web-сервисы, а в облачных системах используются grid-решения [6-7]. Зачастую делается вывод, что grid и облачные системы имеют главное принципиальное отличие: grid – распределенная система, частью которой может быть и ваша библиотека, облако – система централизованная, имеющая стороннего хозяина, как правило, провайдера услуг. Централизация определяется, как правило, наличием в облачных системах Центров обработки данных (ЦОД), хорошо оснащенных вычислительной техникой, имеющих оборудованные помещения и квалифицированный персонал [3,8]. В ЦОД широко используется виртуализация, создающая у пользователя ощущение автономной работы. Однако, распространено также мнение, что облако характеризуется реализацией пользовательских задач посредством web-сервисов [8]. Такой способ работы позволяет защитить инфраструктуру облака и предоставить пользователям легкую для понимания среду работы. Если рассматривать ЦОД как узел grid-системы, то облако – реализация доступа пользователя к этому узлу. При таком подходе вопросы стандартизации взаимодействия между ЦОД, особенно в области взаимообмена данными, приобретает принципиальное значение для библиотек-пользователей. А развитие на базе ЦОД облачных сервисов могло бы стать логичным развитием традиционных корпоративных библиотечных систем.

Использование ЦОД помогает небольшим библиотекам решить свои финансовые проблемы, избежав неподъемных трат на вычислительную технику, оборудование помещений и квалификацию персонала, что способствует популярности аутсорсинга, как коммерческой услуги. Широкое распространение в работе библиотек сервисов социальных сетей (электронной почты, виртуальных дисков-хранилищ, форумов, блогов и т.д.) создает предпосылки для развития таких сервисов в ЦОД. Если соотнести модели работы с ЦОД с облачными моделями предоставления услуг, то в российских библиотеках наиболее распространена работа в двух таких моделях:

1. ***PaaS (платформа как услуга).*** Библиотека арендует у провайдера (возможно у соседней более крупной библиотеки) вычислительную технику и хранилища данных. Программное обеспечение используется собственное. Потребитель при этом не управляет сетями, серверами, операционными системами и системами хранения данных (базовой инфраструктурой облака), но осуществляет контроль над развернутыми приложениями и, возможно, некоторыми параметрами конфигурации среды хостинга. То есть на

вычислительной технике провайдера устанавливается САБ, web-сервер и сайт библиотеки, вспомогательное программное обеспечение. Такой способ работы на практике достаточно далек от облачных сервисов и реализуется обычно через удаленный рабочий стол.

2. ***SaaS (программное обеспечение как услуга)***. При таком способе работы библиотека не покупает специализированное программное обеспечение, например, САБ, совсем, либо частично, что позволяет ей на этом сэкономить. Провайдер, владеющий ПО, размещает на своем сервере данные, к которым пользователи подключаются различными способами, в том числе и с помощью облачных сервисов. Эта модель часто используется при построении сводных каталогов и региональных корпоративных библиотечных систем. Уже сейчас на базе таких ЦОД делаются попытки реализовать некоторые дополнительные услуги как облачные сервисы. Стоимость программного обеспечения в этой модели обычно выше, чем плата за стандартное сопровождение собственного ПО, но существенно ниже, чем его покупка. Расширенный набор услуг, включающий и облачные, предоставляется за дополнительную плату.

У любой из этих моделей вне зависимости от развитой системы облачных сервисов есть ряд несомненных достоинств:

1. Доступность – подключиться можно из любой точки мира, где есть Интернет.
2. Гибкость – неограниченность вычислительных ресурсов за счет виртуализации.
3. Надежность – ЦОД, как правило, имеет резервные источники питания, охрану, профессиональных работников, регулярное резервирование данных, высокую пропускную способность Интернет-каналов, высокую устойчивость к вирусным и хакерским атакам.
4. Возможность сэкономить на покупке вычислительной техники, ПО и IT-персонале.
5. Качество оплаченного сопровождения существенно выше за счет высокой квалификации персонала и использования дополнительных услуг, таких как CRM и бесплатный круглосуточный телефон.

Проблемы взаимообмена данными. Однако у таких систем, как и у соответствующих облачных моделей, есть недостатки, отмеченные в публикациях [9-11]. Практика показывает, что такой способ работы не приносит ожидаемой экономии средств, а в ряде случаев ведет и к дополнительным финансовым расходам [10-11]. Однако главным недостатком, на котором хотелось бы остановиться подробнее, является отсутствие интероперабельности – нет набора универсальных стандартов и интерфейсов, что увеличивает зависимость от поставщика. И в силу конкурентной среды в Интернет создается ощущение, что поставщики услуг кровно заинтересованы в сохранении такого порядка вещей. То есть, нет тех самых «согласованных стандартов и правил», по которым работали корпоративные библиотечные системы вне web-среды. Особенно заметным в настоящей практической работе библиотек является отсутствие таких стандартов при взаимообмене данными между существующими и проектируемыми виртуальными ЦОД.

Как это сказывается на повседневной работе библиотек? Можно выделить, по крайней мере, два крупных направления работы библиотек, где указанные обстоятельства играют очень большую роль, существенно замедляя и усложняя работу.

I. ***Наукометрия и проекты интеграции библиографических ресурсов.*** Множество научных и образовательных библиотек России вовлечено в крупные федеральные проекты. Наиболее значимыми из них являются Карта российской науки, Российский индекс научного цитирования (РИНЦ) и Электронный каталог библиотек сети

образования и науки (ЭКБСОН). Каждый из этих проектов располагает некоторыми собственными данными о публикациях российских ученых, полученными из различных источников, связанных с проектами договорами о поставке информации. К сожалению, эти данные не отличаются необходимой полнотой и точностью. Однако реформа науки и образования отводит создаваемым ресурсам очень важную роль в будущем распределении грантов и финансирования, что вынуждает организации очень ответственно и активно подходить к тому, как они там отражены. Требуется сверка и пополнение данных, как по публикациям, так и по персоналиям. В силу аккумуляции в библиотеке данных о трудах сотрудников эта работа часто поручается библиотекам. Данные проектов в части персоналий и публикаций повторяют друг друга, однако никаких средств загрузки и выгрузки данных системы не предоставляют. Выливается это в бесконечный ввод вручную одних и тех же данных в разные ресурсы. Если добавить к этой работе еще и собственную библиографическую базу трудов сотрудников, а также базу данных ученого секретаря, то объем работы становится еще больше. Справедливости ради следует отметить, что под давлением научной общественности средства выгрузки данных были обещаны разработчиками проектов, как и обмен данными между проектами. Однако этот вопрос, по-видимому, не является сколько-нибудь значимым и его решение постоянно откладывается. Это вынуждает библиотеки прилагать собственные усилия для решения проблемы бесконечного дублирования работы. Разработаны и распространяются в библиотечном сообществе средства конверсии в системы автоматизации библиотек (САБ) данных, выгруженных из международных индексов научного цитирования Web of Science и Scopus, благо средства выгрузки данных есть в этих системах достаточно давно. Отметим, что этот процесс не был быстрым и гладким, требовал усилий и квалификации. Пользователи РИНЦ, оплатившие доступ к Science Index, получили возможность выгрузить данные в xml-формате. Можно приступить к работе над их конверсией, потому что хотя бы прояснилась структура хранимых в РИНЦ данных. Хотелось бы, чтобы решена была и проблема загрузки данных, хотя бы на платной основе. Непонятно, почему специалист может ввести данные о публикации в систему вручную, но не может загрузить их, например, из rusmarc, и при необходимости отредактировать. Проект ЭКБСОН предоставляет сервис загрузки данных в САБ пользователя, однако формат и структура данных пока неизвестны. На очереди Карта российской науки. Неужели и здесь сверка и пополнение данных будут также производиться вручную.

II. ***Электронные библиотечные системы (ЭБС) как фактор книгообеспеченности студентов.*** Одним из базовых требований для лицензирования образовательной деятельности является наличие в библиотеке ЭБС, собственной и/или продукта сторонних разработчиков. Система книгообеспеченности студентов предполагает и учет литературы из имеющихся в библиотеке ЭБС. Предполагалось, что проблема может быть решена с помощью одной из многочисленных дискавери-систем, предлагаемых как отдельный продукт разработчиками ЭБС. Однако такие системы, как правило, нацелены на сквозной поиск данных и не предназначены для их обмена. К тому же они очень дороги. Поэтому на практике библиографическое описание каждой книги дублируется в электронный каталог библиотеки с помощью табличной выгрузки, предлагаемой разработчиками ЭБС. Данные выгружаются в той структуре, в какой они хранятся в ЭБС. Дальше идет привычная работа по конверсии выгруженных табличных данных в САБ с

разбором и редактированием каждого поля. И так для всех ЭБС, которыми располагает библиотека.

Выводы. Регулярная работа по конверсии выгруженных данных из каждой новой российской информационной системы и невозможность самостоятельной загрузки данных даже при оплаченных правах на их ввод порождают в среде библиотечных специалистов острую тоску по стандартизованным Z39.59 и rusmarc.

- Необходим единый для всех систем, содержащих данные о публикациях, формат загрузки-выгрузки данных
- Единый и строго определенный формат загрузки-выгрузки данных, как и сама возможность внешней загрузки-выгрузки, должны стать обязательным требованием для всех информационных систем, содержащих библиографию и данные о персоналиях.

Литература

1. Хрусталев, Е.Ю. Методический подход к проектированию сервисов упрощенной интеграции распределенных IT-ресурсов / Е.Ю.Хрусталев, А.А.Чумичкин / Информационные ресурсы России. – 2012. – №3. – с.2-6.
2. Ковязина Е.В. Перспективы развития автоматизации библиотек / Е.В.Ковязина // Научные и технические библиотеки. – 2011. - №2 – с.89-92.
3. Ильин, В.А. Больше данных, хороших и разных! / В.А.Ильин, В.Е.Велихов // В мире науки. – 2014. - №2. – с.38-44.
4. Sanchati, R. Cloud Computing in Digital and University Libraries [Текст] [Электронный ресурс] / R. Sanchati, G. Kulkarni // Global Journal of Computer Science and Technology. – 2011. – Vol. XI, Iss. XII, ver. 1.0. - с. 37-41. - URL: http://computerresearch.org/stpr/index.php/gjcst/article/viewFile/860/765.
5. Kaushik, A. Application of Cloud Computing in Libraries [Текст] [Электронный ресурс] / A. Kaushik, A. Kumar // International Journal of Information Dissemination and Technology. – 2013. – Vol. 3(4). – с.270-273. – URL: http://www.ijidt.com/index.php/ijidt/article/viewFile/3.4.9/pdf.
6. Myerson, J. V. Cloud Computing versus grid computing [Текст] [Электронный ресурс] / J.M. Myerson. - IBM, 2009. – URL: http://www.ibm.com/developerworks/library/wa-cloudgrid/wa-cloudgrid-pdf.pdf.
7. Hashemi, S.M. Cloud Computing Vs. Grid Computing [Текст] [Электронный ресурс] / S.M.Hashemi, A.K. Bardsiri // ARPN Journal of Systems and Software. – 2012. - Vol. 2, № 5. – с. 188-194. – URL: http://scientific-journals.org/journalofsystemsandsoftware/archive/vol2no5/vol2no5_4.pdf.
8. Сафонов В.О. Платформа облачных вычислений Microsoft Windows Azure / В.О.Сафонов. – М.: Национальный открытый университет «ИНТУИТ»: БИНОМ. Лаборатория знаний, 2013. – 234 с. – (Основы информационных технологий).
9. Интероперабельность в облачных вычислениях [Текст] [Электронный ресурс] / Е.Е.Журавлев [и др.] // Журнал радиоэлектроники. – 2013. - №9. – с.1-63. – URL: http://razinkin.16mb.com/publications/clouds.

10. Новиков И. Облачные вычисления: на пороге перемен [Текст] [Электронный ресурс] / И.Новиков // PC Magazine/RE. – 2011. - №4. - URL: http://www.pcmag.ru/solutions/detail.php?ID=44441. (таблица цен вендоров)

11. Емельянов И. Миф о дешевизне облачных решений [Текст] [Электронный ресурс] / И.Емельянов // Компьютерра. – 2013. - №10. – URL: http://www.computerra.ru/cio/5574