Библиогрид - электронные библиотеки как средство управления информацией в грид-среде

Жучков А.В., ИХФ РАН alex@umos.ru

Маркарова Т.С., ГНПБ им. К.Д. Ушинского

Твердохлебов Н.В., ИХФ РАН

Арнаутов С.А. ИХФ РАН

Аннотация

Описано современное состояние работ, проводимых в проекте Библиогрид, в результате которых программное обеспечение среднего уровня основе Globus Toolkit 4/OGSA-DAI было дополнено новым инструментарием для работы с цифровыми сущностями, информационными объектами, коллекциями и онтологиями, что позволило реализовать основные функции электронных библиотек для участников виртуальных организаций в грид-среде.

1 Введение

Основные элементы концепции и реализации проекта Библиогрид были описаны в [1]. В этой публикации были представлены причины, которые инициировали разработку интегрирующей среды на основе грид-технологий и сервис-ориентированной архитектуры, а также некоторые новые возможности, которые предоставляет среда Библиогрид для пользователей электронных библиотек (ЭБ).

В настоящей статье более подробно описаны механизмы организации информационного пространства Библиогрид и грид-сервисы для формирования научных персональных и корпоративных коллекций данных с использованием разноуровневых метаданных (в том числе предметно-ориентированных онтологий) и распределенных разнородных федеративно-администрируемых информационных ресурсов.

Программно-техническую и архитектурную основу Библиогрид составляют программное обеспечение промежуточного уровня (middleware) Globus Toolkit 4 (GT) [2] и OGSA-DAI [3], которые используются в наиболее крупных и разработанных грид-сетях в Европе [4] и США [5]. Использование в Библиогрид таких программно-технических и архитектурных решений позволяет обеспечить взаимодействие пользователей Библиогрид с информационно-вычислительными ресурсами основных мировых научных центров, включая ресурсы экспериментальных данных, информационные хранилища и, что важно для научных коллабораций, вычислительные ресурсы. Тестирование такого взаимодействия с европейскими ресурсами в целях медико-биологических исследований было осуществлено в рамках совместных работ по проектам ассоциации HealthGrid [6, 7].

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

2 Виртуальные организации и электронные библиотеки

Человек есть по определению существо общественное. Практически все виды его активности протекают в составе группы (коллектива, коллаборации). И потому, что современное научное оборудование в весьма короткие сроки позволяет получать огромные массивы экспериментальных данных, требующих квалифицированной обработки и анализа, и потому, что решение сложных проблем любого типа (научных, прикладных, социальных, политических и т.д.) требует коллективного разума группы экспертов. Взаимодействие между экспертами включает обмен данными, информацией и знаниями. Для организации такой работы создаются виртуальные организации (ВО) [8], деятельность которых неразрывно связана с использованием современных информационно-коммуникационных технологий [9].

Возможности ВО определяются общим для данной группы экспертов информационным пространством, которое включает в себя как уже имеющиеся по данной проблеме данные, информацию и знания, так и новое знание, полученное в ходе работы данной ВО. Хранятся эти ресурсы в виде цифровых сущностей (digital entities [10]), информационных объектов, коллекций и онтологий (заметим, что формы и методы хранения знаний могут быть и иными). Инструментарий, необходимый для создания информационного пространства ВО разрабатывается как в рамках грид-технологий, так и в ЭБ.

Сильными сторонами грид-технологий являются развитые сервисы безопасности (что существенно для чувствительных научных данных, например, медицинских [11]), возможность использования огромных вычислительных мощностей грид-инфраструктур, а также комплекс технологий управления распределенными данными, включая их хранение в удаленных хранилищах, поддержку единого пространства имен, каталог для хранения информации о данных, механизмы для обеспечения предпочтительных механизмов доступа к данным.

Любая библиотека по своей функциональности представляет собой информационно-поисковую систему и включает в себя как непосредственно информационные ресурсы (возможно хранящиеся распределенно), так и службы (сервисы) доступа к информационным ресурсам и административные службы (поддержки и хранения ресурсов и т.п.). ЭБ предназначены обеспечивать эти же самые функциональности. ЭБ, создаваемые как приложения верхнего уровня, содержат механизмы создания коллекций, просмотра и открытия информации. Соответствующий набор функций включает создание схем данных, массовую загрузку метаданных, импорт-экспорт метаданных, управление иерархической структурой коллекции [10]. Пополнение набора сервисов Библиогрид указанными функциональностями даст возможность научным ВО сочетать преимущества обоих технологий при работе с большими объемами информации и позволит не только создавать ЭБ посредством сервисов грид, но и использовать ЭБ как средство управления информацией ВО в грид-среде. Первым этапом в этой работе стала разработка грид-сервиса, обеспечивающего формирование коллекций информационных объектов, цифровые сущности которых хранятся в распределенных разнородных информационных ресурсах (реляционные и XML базы данных, файлы, Web-страницы и т.д.).

3 Объединение цифровых сущностей и информационных объектов в коллекции

Создание коллекций научных данных является одним из важнейших элементов научной работы. Коллекции могут быть как индивидуальными, принадлежащими одному эксперту («авторский набор» или персональная коллекция), так и федеративными, создаваемыми в ходе совместной деятельности коллаборации экспертов. Достоинства грид-технологий наиболее полно проявляются при совместной работе экспертов. Первой возможностью, которую предоставляют сервисы Библиогрид для такой работы являются механизмы аутентификации и авторизации участников совместной деятельности, обеспечивающие согласованную работу исследователей над общими информационными объектами и возможность сохранения их информационной целостности при распределенном хранении наборов данных нижних информационных уровней.

Библиогрид, как было описано в [1], основывается на следующих базовых концепциях: информационный объект, репозиторий метаописаний и коллекции данных.

Информационный объект (ИО) состоит из цифровых сущностей (содержательных данных), соответствующих метаданных (семантических ярлыков) и методов (activities), представляющих собой программный код, определяющий допустимые действия, которые могут быть произведены над данными и их метаданными, включая интерфейс нижнего уровня для доступа к ним. В Библиогрид точкой доступа к ИО в информационных ресурсах ВО является грид-сервис данных (Grid Data Service, GDS). Функциональность GDS позволяет эффективно оперировать с распределенными гетерогенными данными и включает поддержку различных моделей хранения данных (реляционная, файловая система, XML, и др.) и Различных СУБД (MySQL, PostgreSQL, Xindice, eXist, и др.). С точки зрения сервисов OGSA-DAI каждый ИО представлен набором XML файлов с описывающими его метаданными, которые определяют основные характеристики цифровых сущностей и параметры методов (activities) для доступа к ним, включая права доступа.

Вышеназванные метаданные полностью определяют ИО с точки зрения возможности осуществления доступа к его цифровым сущностям. Но для объединение ИО в коллекцию этого уровня метаданных недостаточно, необходимы расширенные метаданные, описывающие коллекцию в целом.

Для реализации такой возможности в ходе работ по развитию Библиогрид был разработан и реализован специализированный сервис, позволяющий создавать репозиторий метаописаний (РМО), в котором в формате METS [12] хранятся расширенные метаданные, описывающие характеристики коллекции ИО, то есть логически связанного набора ИО (с точки зрения автора данной коллекции). В формате METS определены три вида метаданных – описательные, структурные и административные, что позволяет описывать все виды цифровых объектов. При этом формат содержит не определенные предписания о содержании метаданных, а лишь базовые схемы ИО. Собственно цифровые сущности (текст, изображение, видео и пр.) могут храниться либо в составе METS-описания в виде XML-файла, либо отдельно (что описывается соответствующей ссылкой). Важно, что структурные и описательные метаданные хранящегося в РМО метаописания коллекции рассматриваются как часть контента и могут модифицироваться и пополняться как пользователями, так и автоматически сервисами, на

основе анализа доступной информации. Эти метаданные, по сути своей, являются набором семантических ярлыков (в другой терминологии – описательными, структурными и административными метаданными) и семантических связей, которые эксперты присваивают цифровым сущностям (то есть ИО) в ходе работы над проблемой. Таким образом создается осмысленный контекст, в котором цифровые сущности (отдельные ИО) превращаются в персональные или корпоративные коллекции.

Доступ к любому ИО конкретной коллекции осуществляется посредством экземпляра GDS, который включает в себя метаописание данного ИО. Важно отметить, что доступ к хранящимся в РМО наборам метаописаний, определяющих коллекции ИО, осуществляется посредством такого же самого механизма, что и доступ к самим коллекциям, а именно, с помощью экземпляра GDS, то есть совокупность метаописаний коллекции ИО также является коллекцией. Такая реализация доступа к метаописаниям возможна потому, что они представляют собой XML-документы и хранятся в XML базе данных (eXist), что позволяет оперировать с ними стандартными средствами OGSA-DAI. Необходимо заметить, что уже первые эксперименты показали, что для данной задачи необходима более мощная СУБД, чем eXist и в дальнейшем предполагается использовать для хранения РМО СУБД Sedna [21].

Добавление ИО к коллекции состоит во внесении его метаданных в хранящееся в РМО метаописание данной коллекции, что осуществляется посредством соответствующего экземпляра GDS. Очевидно, что ИО может входить в несколько коллекций, при этом возможно гибкое изменение как ассоциированных с ним метаданных, так и методов взаимодействия с ним. Отметим, что функциональность OGSA-DAI позволяет реализовывать распределенные гетерогенные РМО, а также получать доступ к информационным ресурсам вне грид (например, к базам данных, доступным по Интернет).

Таким образом, описанные сервисы Библиогрид реализуют механизмы создания коллекций, просмотра их каталогов и открытия содержащейся в них информации и вместе с цифровыми сущностями информационных объектов представляют собой ЭБ в среде грид.

Предложенная технология виртуализирует обработку данных при работе с коллекциями ИО. Процесс взаимодействия показан на рис. 1. Пользователь посылает в РМО запрос, чтобы найти метаданные, соответствующие его требованиям. Сервисы OGSA-DAI находят соответствующий набор ИО необходимой коллекции и далее либо возвращают набор метаданных этих ИО пользователю, либо GDS взаимодействует с ресурсами данных, используя метаданные в ИО и представляет пользователю необходимые цифровые сущности в соответствии с запросом вместе с сопутствующими метаданными (рис. 1).

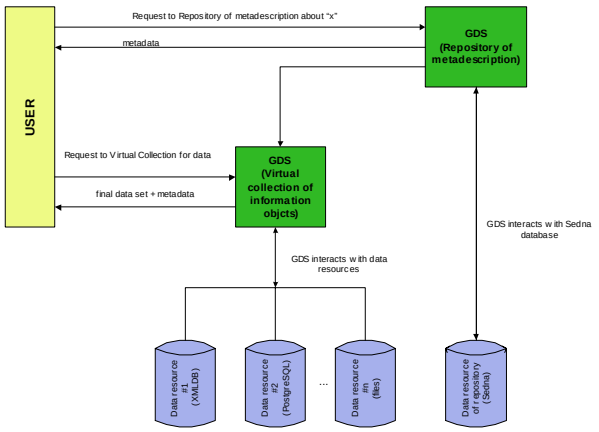

Рис. 1. Использование метаописаний коллекции при взаимодействии с ресурсами данных через GDS

Каждый ИО содержит информацию, где размещены его собственные цифровые сущности и как их получить при использовании GDS. Следовательно, любой запрос будет выполнен в виртуальном информационном пространстве, которое определено метаданными, сохраненными в РМО. Сервисы OGSA-DAI выполняют поиск по всем доступным распределенным гетерогенным ресурсам - различным базам данных, веб-сайтам, и т.д., и пользователь может сконцентрироваться на предмете исследования, не вдаваясь в технические детали поиска данных.

GUI, который используется в Библиогрид в качестве интерфейса между пользователями и GDS, был унаследован от предыдущей версии (не грид-ориентированной) GUI ВО “Вакцины нового поколения”. Этот GUI обеспечивает Windows-интерфейс к интегрированному информационному пространству, а также включает инструментальные средства для построения предметно-ориентированных онтологий. Эти средства основаны на специализированных грид-сервисах, которые были разработаны в рамках проекта BiblioGrid для работы с онтологиями в грид-среде [13, 14].

4 Библиогрид в ГНПБ им. Ушинского. От Тезауруса к онтологиям в грид-среде

Новым примером применения разработанных грид-сервисов могут служить работы, проводимые в ГНПБ им. Ушинского.

Важной особенностью организации информационного пространства ГНПБ является наличие формализованного отраслевого тезауруса, который вбирает в себя понятийно-

терминологический аппарат такой интегрированной области человеческого знания как педагогика (педагогическая наука). Русско-английский тезаурус по педагогике и образованию ГНПБ им. К.Д. Ушинского, разработан на основе тезауруса ЮНЕСКО МБП по образованию и представляет собой терминологический словарь иерархического типа, в котором термины истолковываются посредством других родственных и близких по смыслу терминов, которые определенным образом подчиняются главному слову. Тезаурус имеет достаточно строгую, но в то же время динамичную структуру и состоит из крупных логико-семантических полей, которые подразделяются на более мелкие фасеты, состоящие из терминов (дескрипторов) [15].

Базовыми единицами тезауруса являются термины, когда речь идет о тезаурусе как о таковом, и дескрипторы, когда речь идет об информационно-поисковом тезаурусе. Дескриптор - это слово (термин), которое выбирается из ряда синонимических терминов в качестве представителя этого ряда (предмета информационно-поискового запроса) в информационно-поисковых системах. Перечень дескрипторов формирует поисковый образ документа [16]. Лексическая наполненность и семантический потенциал логико-понятийных полей данного тезауруса по педагогике позволяют ему функционировать в качестве информационно-поискового языка, но и в равной мере представлять достаточно интегрированную и постоянно обновляющуюся педагогическую терминологию.

Говоря о системе терминов, мы подразумеваем сложную систему семантических отношений между ними. Онтологическая иерархия заложена в семантике самого термина и самодостаточна для выявления всех реально существующих или потенциальных связей и отношений между понятиями, событиями, явлениями, предметами, номинируемыми данным термином. Характерно, что такие подробные и полные классификации явились итогом соединения философско-логических исследований о природе знания с прагматическими запросами информатики.

Таким образом, тезаурус ГНПБ - это дискретный парадигматический инструмент со своим метаязыком, который может служить лингвистической моделью целостного знания, выраженного в научных текстах, словарь с концептуальным входом и фиксированными семантическими связями между его единицами. Дескрипторы-термины, составляющие логико-понятийные поля тезауруса, могут служить вербальными элементами поискового образа документа (педагогической информации) в электронной базе данных и одновременно их можно использовать как концепты базовой онтологии, являющейся основой для построения персонального информационного пространства.

Очевидно, что данный отраслевой тезаурус является ценнейшей информацией, необходимой (и используемой) для организации информационного пространства ГНПБ. Однако, существующая реализация тезауруса жестко связана с имеющейся в ГНПБ программной оболочкой. Для использования тезауруса в среде Библиогрид он был переведен в XML форму и затем, для использования в качестве основы онтологии, записан в стандарте OWL, который является одним из наиболее используемых и перспективных сегодня для формализованной записи онтологий [17]. В результате отраслевой тезаурус стал самостоятельным информационным объектом, который можно применять в различных информационно-поисковых системах по данной предметной области. Использование OWL позволяет упростить процесс поиска данных с использованием семантических критериев, возложив необходимость описания предметной области в целом и контекста поиска полностью на авторов и на систему поиска.

Дальнейшая загрузка представленного в OWL тезауруса в РМО Библиогрид позволяет в полной мере реализовать преимущества коллективной работы ВО над общим информационным полем. Для этого сервисом используются специальные механизмы OWL. К ним, в частности, относятся механизм описания версий онтологий и механизмы агрегирования данных, содержащихся в онтологиях. Первые позволяют описывать ссылки на предыдущие версии онтологии и изменения, сделанные в данной версии по отношению к предыдущей. Это позволяет прослеживать эволюцию онтологий, а так же оперировать разновидностями онтологий разных авторов для одной и той же предметной области. Механизмы агрегирования данных решают задачи объединения различных онтологий, размещенных в среде Библиогрид, а также в Интернет вне грид-среды (но уже без отслеживания авторских прав по грид-сертификату), в том числе организацию ссылок из одной онтологии на классы и объекты другой онтологии. Все это позволяет организовывать набор онтологий и связанных с ними содержательных данных и метаданных в единую понятийную сеть ВО. На этой основе как пользователи сервиса, так и сервисные программные системы (например, контекстного поиска) будут способны ориентироваться в понятийной сети и осуществлять различные необходимые им интеллектуальные операции над доступной информацией [18].

Тезаурус по педагогике в XML форме загруженный в РМО Библиогрид может являться интеграционной основой формирования ЭБ ГНПБ, поскольку определяет пространство концептов предметной области в виде предметно-ориентированной онтологии, а каждый концепт тезауруса (онтологии) связан с коллекцией ИО в описанном выше смысле. Конечно, на начальном этапе формирования ЭБ коллекции представлены в

РМО только метаданными и ссылками, а цифровые сущности, как правило, отсутствуют. Связывание концептов с конкретными источниками данных и пополнение коллекций и ИО метаданными является прерогативой администраторов ЭБ (с точки зрения формирования общей части ЭБ) и пользователей (с точки зрения формирования их персональных коллекций).

ГНПБ располагает значительными электронными ресурсами, включая электронный каталог библиотеки, каталог «Труды Российской академии образования», полнотекстовую базу авторефератов диссертаций по педагогике и народному образованию, Электронный журнал: «Образование: исследовано в мире», а также каталог Интернет-адресов дополнительных авторитетных баз данных и альтернативных источников библиографической и иной информации в области образования, педагогики и психологии [19]. Однако оцифрованного фонда очевидно не хватает и использование читателями библиотеки созданного сервиса онтологии будет являться стимулирующим фактором к ведению такой важной деятельности с пониманием реальной потребности и приоритетов.

Главной целью проводимых работ является предоставление сервисов поиска информации и формирования персональных коллекций научным работникам, занятым написанием диссертаций, созданием методических материалов по педагогике и т.п. С помощью сервисов грид-среды они получают такую возможность вместе с возможностью доступа к распределенным информационным ресурсам библиотеки, хранящимся в электронной форме и внешним относительно библиотеки ресурсам, но включённым в ресурсное пространство грид-сегмента. При этом основные преимущества грид-среды, такие как информационная безопасность, гибкое управление правами доступа к любым ресурсам ВО на основе использования цифровых сертификатов и технологии «открытых ключей», «прозрачный» доступ с единой точкой входа ко всем информационным ресурсам ВО благодаря использованию технологии OGSA-DAI, а также использование сервисов контентного билинга и другие значительно расширяют возможности ЭБ ГНПБ.

5 Заключение

Целью проекта Библиогрид является создание методов и инструментов, которые позволяли бы экспертам, входящим в состав ВО создавать собственное информационное пространство, включая модель предметной области (концептуальное пространство), на основе сервисов, оперирующих набором ИО в грид-среде [20]. Важно, что частные онтологии (как элементы концептуального пространства), созданные одним экспертом, могут эффективно аггрегироваться в более широкие онтологии верхнего уровня, разрабатываемые всеми членами ВО. При этом они остаются увязанными с метаданными более низких уровней и в конечном итоге с ИО и цифровыми сущностями, распределёнными по грид-сегменту.

Работы настоящего этапа проекта Библиогрид позволили развить как инфраструктурные сервисы грид-среды, за счёт создания сервиса поддержки РМО, так и высокоуровневые (прикладные), за счет привязки сервиса онтологий и сервиса поддержки коллекций ИО к сервису РМО.

Таким образом, программный комплекс среднего уровня на основе Globus Toolkit 4/OGSA-DAI в Библиогрид оснащён сегодня удобным инструментарием для работы с цифровыми сущностями, ИО, коллекциями и онтологиями. Он содержит достаточный набор сервисов, позволяющих реализовывать основные функции ЭБ при работе с распределенным гетерогенным контентом. Для грид-среды это означает появление новой возможности – использования ЭБ, как высокоэффективной системы управления информацией.

Важно, что на данном этапе проекта к ранее созданным и функционирующим ВО «Новое поколение вакцин», ВО «Химическая и биохимическая физика», где использование грид-среды было вызвано прежде всего необходимостью проведения сложных вычислений, увязанных на интенсивные взаимодействия с распределёнными гетерогенными данными присоединилась ВО Государственной научно-педагогической библиотеки, целью участников которой является развитие ЭБ и предоставление нового уровня сервисов читателям библиотеки. Такая работа более чётко формирует новые задачи развития сервисов ЭБ в грид-среде.

Литература

[1] Жучков А.В. Библиогрид – основные элементы концепции и реализации // Электронные библиотеки: перспективные методы и технологии. Труды 7-й Всероссийской научной конференции (RCDL’2005). Ярославль, ЯрГУ, 2005, с.31-36.

[2] Globus Project. http://www.globus.org .

[3] OGSA-DAI Project. http://www.ogsadai.org .

[4] The Enabling Grids for E-Science Project. http://www.eu-egee.org .

[5] The TeraGrid Project. http://www.teragrid.org .

[6] Черный А.А., Трушкин К.А., Боковой В.А., Яновский А.К., Твердохлебов Н.В., Жучков А.В., Лысов Ю.П. Система распределенного хранения и анализа геномной информации // Молекулярная биология. 2004 г., т. 38, №1, сс. 104-109.

[7] The HealthGrid White Paper. Ed. By Breton V., Dean K. Studies in health technology and informatics, IOS Press, 2005; 112:249-321.
[8] Foster I., Kesselman C. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Pub., San Francisco, CA (1999)
[9] Mowshowitz A. Virtual Organization: Toward a Theory of Societal Transformation Stimulated by Information Technology // Westport, CT: Quorum Books, 2002, 264 pp. ISBN 1-56720-501-1.
[10] Moore R., Rajasekar A., Wan M.: Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Sharing, Publishing, and Archiving Data. Proceedings of the IEEE, Vol. 93, # 3 (2005) 578-588.
[11] Satou K., et al.: An Integrated System for Distributed Bioinformatics Environment on Grids. In: Konagaya A., Satou K. (eds.): Life Science Grid 2004. Lecture Notes in Computer Science, Vol. 3370. Springer-Verlag, Berlin Heidelberg New York (2005) 8–19.
[12] Metadata Encoding & Transmission Standard (METS), http://www.loc.gov/standards/mets
[13] Joutchkov A., et al.: Grid-Based Onto-Technologies Provide an Effective Instrument for Biomedical Research. Studies in Health Technology and Informatics, 112 (2005) 37-46.
[14] Жучков А.В, и др. Интеграция и поиск информации в гетерогенных динамических информационных массивах с помощью онтологий // Труды 6-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" (RCDL'2004), Пущино, 2004.
[15] Плотникова В.А., и др. Современная библиотека образовательного учреждения: Пособие для системы доп.проф. образования. – М.: Федерация Интернет Образования, 2005, с.45.
[16] Краткий словарь лингвистических терминов - М.: Русский язык, 1995, с.31.
[17] Когаловский М.Р. Стандарты XML и электронные библиотеки // Электронные библиотеки, 2003, Т. 6, Вып. 2.
[18] Кафтанников И.Л., Коровин С.Е. Перспективы использования web-онтологий в учебном процессе // Educational Technology & Society , 2003, 6(3):134-138.
[19] Электронные ресурсы ГНПБ им. К.Д. Ушинского, http://www.gnpbu.ru/katalog/kat_0.htm .
[20] Joutchkov A., et al.: Libraries of Strategies and Ontology-driven Subject Area Models as "Corner Stones" in Grid Development. Methods of Information in Medicine, Vol. 44, #2, (2005) 249-252.
[21] Интернет-портал Sedna – Native XML DBMS, http://www.modis.ispras.ru/Development/sedna.htm

BIBLIOGRIG – DIGITAL LIBRARIES AS TOOLS TO CONTROL OF INFORMATION IN GRID ENVIRONMENT

A. Zhuchkov, T. Markarova, N. Tverdokhlebov, S. Arnautov

Current state of Bibliogrid project is described. Middleware Globus Toolkit 4 and OGSA-DAI have been extended by new tools to operate with digital entities, information objects, data collections and ontologies. These results allowed to implement basic functions of digital libraries for participants of virtual organizations in Grid environment.