Сервис-ориентированный Грид-подход к информационным задачам в пространствах данных виртуальных организаций

А.В.Жучков, А.В.Кравченко, Н.В.Твердохлебов

ИХФ РАН
alex@chph.ras.ru

Аннотация

В статье рассматриваются возможности применения Грид-технологии для построения платформы поддержки пространств данных. Приводится пример построения высокоуровневого сервиса, работающего на пространстве данных медицинской виртуальной организации.

1 Введение

Традиционно разработчики СУБД при реализации решений по управлению данными предлагают подход, направленный на максимальную структуризацию всех источников данных. Но, зачастую, алгоритмы интеграции разнородных данных на основе приведения к единой схеме данных требуют недопустимо высоких затрат. В тоже время в рамках большинства прикладных задач столь тесная интеграция - приведение всех источников к одной схеме или построение единой схемы над всеми источниками просто не требуется. И здесь становится актуальным подход к среде распределенных источников данных, как к пространству данных (ПД) [1,2].

С другой стороны, появившаяся значительно раньше концепция Grid (Грид) [3], как согласованная, открытая и стандартизованная среда, которая обеспечивает гибкое, безопасное и скоординированное разделение ресурсов в рамках виртуальных организаций, завоевала очень много серьёзных сторонников. Сегодня уже существует достаточно примеров работающих Грид-систем, правда, в основном в областях, требующих массированных вычислений [4,5]. Важно подчеркнуть, что концепция Грид понимает «ресурс» максимально широко и это даёт возможность Грид-технологам также оперировать и понятиями ПД, как разновидностью ресурсов Грид.

В связи с этим, интересно посмотреть на приведенные два подхода с точки зрения возможностей взаимного проникновения технологий.

2 Пространства данных – Виртуальные организации

Концепция ПД предполагает, что участники этого пространства – внешние по отношению к системе обработки данных, административно-распределенные и семантически гетерогенные источники данных могут сосуществовать с некоторой необходимой степенью связности: от простого перечня этих источников до серьёзной БД, объединяющей их в соответствии с некоторой схемой. При этом концепция ПД предусматривает возможность моделировать любой вид связи между участниками.

Такая широкая трактовка понятия пространства данных включает в область ПД также и контент современной WWW-паутины. Однако следует сказать, что в этой сфере уже созданы сильнейшие инструменты – прежде всего поисковые машины, обрабатывающие миллионы Web-документов и при этом осуществляющие миллиарды переходов по ссылкам. На данный момент поисковые сервисы (сервисы индексирования) являются основными сервисами, получившими практическое применение над средой слабосвязанных Web-документов.

Хаос WWW порождает и будет порождать все новые средства работы с такого рода данными. Они могут появляться и в рамках таких научных направлений как Data mining или Data warehouse, и в рамках собственно развития Web на пути привязки Web-данных к семантическим моделям и создании разноплановых метаданных. Хорошими примерами стали разработка стандарта RSS [6] или попытки компании Google перейти к управлению слабосвязанными данными в среде Web (протоколы Atom [7], GData [8]).

Однако, проблемы не только остаются, но и нарастают. Гигантский объем слабосвязанных данных, представленных в WWW, по мере своего гипербыстрого роста начинает терять информационную ценность. Свидетельство этому - рост доли P2P (peer-to-peer) в общем объёме Интернет-трафика. Складывается парадоксальная ситуация – фактическое отсутствие ценной

информации при резкой избыточности самих Web-данных. Это происходит как из-за низкого качества самого контента, так и из-за крайне малого числа сервисов, умеющих эффективно работать с подобным контентом.

Возможное решение этой актуальной проблемы видится нам в развитии направления, связанного с поддержкой концепции виртуальных организаций (ВО), концепции, которая по нашему убеждению становится фундаментальной для многих современных процессов вычислений и обработки данных.

Термин ВО стал особенно популярен в последнее время прежде всего благодаря набирающей практическую мощь концепции Грид. Но если мы будем следовать определению ВО, принятой в Грид-сообществе [9], как географически и административно распределенного динамического объединения ресурсов, служб и людей для решения различных задач (и прикладных, и научных), понимая, что при этом основной проблемой становится эффективное и согласованное разделение этих ресурсов, среди которых могут быть как вычислительные, так и информационные ресурсы и службы, тогда очевидно, что объединение в ВО может требовать весь диапазон интеграции данных – от объединения БД до максимально слабой связанности данных в ПД.

Однако очевидно, что даже при слабой связанности участников, когда зачастую требуется только знать, какие коллекции кто из них внес в это пространство, на самом деле требуется наличие достаточно серьёзного программного обеспечения для управления этими коллекциями, так как исходная слабая связанность данных нередко предполагает просто более сложные сервисы их совместной обработки.

Создание, управление и использование динамических ВО, разделяющих ресурсы, требует соответствующей технологии. Наиболее полно необходимая функциональность присутствует в программном обеспечении промежуточного уровня (middleware) Грид-систем и наличие этого программного обеспечения, безусловно, открывает дополнительные возможности и для работы в ПД.

Существует ещё одна важная особенность современных ВО, дающая дополнительные преимущества при работе с ПД. Все они, вне зависимости от корпоративного или межкорпоративного характера, формируются вокруг некоторой предметной области, связанной набором терминов или общей онтологией, знаниями, существующими в некоторой среде экспертов. Принадлежность наборов данных к ВО, вне зависимости от публичной доступности или отсутствия таковой, обязательно подчеркивает их принадлежность к некоторой «понятийной сети» [10].

Таким образом, рассмотрение ПД в контексте ВО открывает дополнительные конструктивные возможности. Инфраструктура поддержки ВО, их семантическая ориентация даёт неоспоримые преимущества при создании средств управления ПД перед сервисами, ориентированными на работу непосредственно на Web-платформе.

3 Платформы поддержки ПД – Грид ПО промежуточного уровня

Для поддержки ПД в работе [1] ее авторы предлагают создавать платформы поддержки пространств данных (Data Space Support Platform, далее DSSP). Цель поддержки ПД состоит в обеспечении базового набора функций над всеми источниками данных. Например, DSSP должна обеспечить поиск по ключевым словам, аналогично тому, что обеспечивают существующие поисковые системы в настольных компьютерах. При потребности в более сложных операциях, таких как запросы в реляционном стиле, анализ данных или мониторинг каких-либо источников, должна существовать возможность приложения дополнительных усилий к более тесной интеграции этих источников шаг за шагом.

DSSP, в отличие от СУБД, не имеет полного контроля над своими участниками, а разворачивает набор сервисов поверх их систем управления данными, соблюдая их потребности в автономии и, несмотря на это, DSSP должна поддерживать данные всех участников пространства без исключения. Средства каталогизации, интегрированного поиска, запрашивания и администрирования, безусловно, должны опираться на метаданные, которые продуцируются как собственно участниками пространства, так и самой DSSP.

В ситуации, когда мы рассматриваем ПД в контексте ВО, было бы удобно в качестве одной из возможных платформ для реализации DSSP использовать непосредственно middleware Грид.

Следует заметить, что это не единственный возможный подход к реализации DSSP. Например, технология виртуализации доступа к участникам ПД, основанная на программных модулях (wrappers), реализована в продукте IBM Web Sphere Information Integrator (WS II) [11]. Wrappers выступают в качестве интерфейсов между WS II и гетерогенными источниками данных.

Основным достижением Грид-технологов бесспорно можно считать разработку стандарта архитектуры программного обеспечения Грид - Open Grid Service Architecture (OGSA), которая рассматривает сервис (службу) в качестве основного объекта Грид [12]. Любое ПО в Грид-среде, должно являться доступным по сети компонентом (сервисом), обеспечивающим требуемую функциональность. Сервисы могут быть как статическими, так и создаваться динамически и, согласно OGSA, обладают состоянием - набором данных, ассоциированных с ним. Посредством удаленного обращения к интерфейсам (activities) сервиса потребитель получает определенный вид

обслуживания, в том числе информационного обслуживания. В ставшем де-факто стандартом ПО Globus Toolkit (GT) [13], начиная с версии 3.0, за управление данными отвечают сервисы OGSA-DAI [14]. Отметим, что специфика управления данными в научных приложениях привела к разработке Grid Wrapper для IBM Sphere II [15], также использующему сервис OGSA-DAI.

Покомпонентное сравнение требований DSSP и возможностей GT версии 4 приведено в Табл. 1.

компонент	DSSP	GT4
Разграничение доступа	Федеративные источники данных в ПД администрируются независимо, а DSSP при этом должна обеспечивать базовые методы взаимодействия, не имея полного контроля над данными. Таким образом , DSSP необходимо обладать компонентом, разграничивающим права доступа и делегирующим их различным участникам.	Имеет целый класс служб - службы инфраструктуры безопасности (Grid Security Infrastructure, GSI), которые обеспечивают безопасность как на границах, так и в самой ВО. При помощи GSI поддерживается однократная регистрация, делегирование полномочий и отображение прав доступа на разные локальные системы. GSI не является централизованной системой безопасности, она основана на инфраструктуре открытых ключей (Public Key Infrastructure, PKI), X.509 сертификатах и коммуникационном протоколе Secure Socket Layer (SSL) и использует службы делегирования прав для поддержания распределенной структуры ВО.
Управление хостинг-средой	Функциональность базовых компонент должна обеспечивать возможность управления особенностями хостинг-среды, опирающаяся на реализацию платформонезависимых интерфейсов.	Базовый сервис GT4 Work Space Management service (WSM) позволяет грид-клиентам динамически создавать и управлять своим рабочим пространством в рамках ВО (например создавать и управлять аккаунтами среды исполнения).
Каталог и просмотр	Необходимо хранить информацию обо всех участниках пространства и связях между ними. Для каждого участника каталог должен содержать схему источника, описательно-статистические данные, информацию о владельце, политику безопасности и доступа. Необходимо поддерживать базовый реестр элементов данных участников.	В инструментарии GT имеется целый класс высокоуровневых служб (Information Service) для идентификации грид-ресурсов: как вычислительных, так и данных. Основную роль в этом процессе играет Monitoring and Discovery Service (MDS), состоящий в свою очередь, из Index Service (IS), Trigger Service (TS) и Aggregator Framework. Именно MDS поддерживает реестр грид-ресурсов данной ВО: IS обнаруживает и собирает информацию о грид-ресурсах, позволяя запрашивать и получать данные ассоциированные с этими ресурсами; Aggregator Framework - служба, агрегирующая информационные потоки в MDS, получающая данные от источников агрегации, в роли которых может выступать любая WSRF или грид служба, эти данные могут генерироваться в процессе выполнения какого-то сервиса или как уведомление, затем эти информационные потоки перенаправляются в IS или TS, а конечными потребителями выступают либо клиенты, либо службы ВО.
Мониторинг	В рамках ПД источники данных должны быть изменяемыми, причем эти изменения могут осуществляться как компонентами DSSP, так и администраторами самого источника. Поэтому необходимо отслеживать изменение состояния и актуализировать эту информацию.	Мониторинг грид-ресурсов, доступных в рамках ВО осуществляет сервис MDS. Его компонента TS осуществляет мониторинг данных, ассоциированных с ресурсами данной ВО. Эти данные описывают состояние каждого ресурса в рамках ВО, и TS может предпринимать действия в ответ на изменение состояния ресурса. Из-за специфики задач мониторинга часто используются внешние программные средства, например MonALISA, представляющая глобальные ресурсы через удобный графический интерфейс.
Поиск	У пользователей должна быть возможность поиска любого элемента данных в ПД независимо от модели источника. Должен поддерживаться поиск по ключевым словам с возможностью просмотра	Компонент OGSA-DAI через сервис GDS (GridDataService) позволяет интегрировать в грид-среду разнородные источники данных, модель данных которых может быть как реляционная, так и XML. Эти данные могут находиться как под управлением СУБД, так и в виде файлов в локальных файловых системах грид-узлов. При помощи базового набора activities (функций) сервис GDS позволяет осуществлять как SQL запросы к реляционным СУБД, так и выполнение

	содержимого результатов и интерфейсом, позволяющим пользователю уточнять (усовершенствовать) свой запрос.	XQuery, XPath, XUpdate над XML данными, также базово поддерживаются activities архивирования и трансформации данных, передачи их другим службам или потокам, кеширование, индексирования и полнотекстового поиска.
Структури-рованные запросы	DSSP должны поддерживать запросы в стиле источников данных, входящих в ПД. При этом необходимо, чтобы такие запросы (запросы “в стиле БД”) поддерживались на основе общих интерфейсов. Данный компонент DSSP должен поддерживать как точные, так и приближенные семантические отображения, чтобы иметь возможность доступа к любому источнику данных, не зависимо от его модели.	Служба OGSA-DAI - *Distributed Query Processing* (DQP), позволяет реализовать механизм единого запроса к множеству источников данных и при этом оптимизировать запросы к различным СУБД. DQP состоит из двух классов сервисов: сервис Grid Distributed Query Service (GDQS) и сервисы Query Evaluator Services (QES), которые реализуются экземплярами GDS. Клиент взаимодействует исключительно с сервисом GDQS, к которому и поступает запрос на OQL. Сервис GDQS разлагает этот запрос на подзапросы, которые и передает соответствующем QES, которые взаимодействуют либо непосредственно с GDS, обслуживающем СУБД, либо с другим QES для оптимизации подзапроса. Данные полученные от GDS, обслуживающих источники данных агрегируются GDQS и возвращаются клиенту.
Запросы к метаданным	DSSP должны поддерживать различные виды метаданных: о местоположении данных, об уровне релевантности ответов на запросы, об источнике ответа или о том, как этот ответ был получен, о временных метках на элементах данных, которые участвовали в вычислении ответа, о зависимых элементах данных и др.	Базово работа с метаданными не поддерживается. В то же время те или иные ВО активно строят такие сервисы, опираясь на сервисы OGSA-DAI. Например, в [18] или в [19] предложен подход, связанный с созданием репозитория метаописаний, в котором метаданные различного применения хранятся в формате METS.
Локальное хранение и индексиро-вание	DSSP должны поддерживать создание запрашиваемых ассоциаций между объектами данных от разных участников, совершенствование доступа к источникам с ограниченными собственными средствами доступа. Необходимо поддерживать возможность кэширования отдельных элементов данных и обеспечивать возможности выполнения запросов без доступа к реальным источникам. Также для ПД нужна поддержка индексов сверх тех, которые обеспечиваются любым участником пространства.	Возможности сервиса OGSA-DAI предусматривают создание новых индексов данных, однако локальное хранение базово не поддерживается, если не считать, что все служебные информационные системы обслуживаются СУБД Postgres, которая может использоваться и для целей локального хранения данных. Однако, ВО могут использовать любые средства для данной цели, желательно только, что бы эти СУБД имели возможность обслуживаться базовым сервисом GDS.
Компонент репликации	DSSP должна содержать компоненту, строящую отображения между логическим именем элемента данных и физическим местоположением этих данных среди источников – участников пространства. Такие отображения, безусловно, должны снабжаться метаданными: временными метками, данными о происхождении этих элементов данных и др. Эти метаданные	Data Replication Service (DRS) и Replica Location Service (RLS) входят в набор базовых компонент по управлению данными грид в инструментарии GT4. RLS - сервис кэширования элементов данных. Именно этот сервис сопоставляет логическому имени файла адреса или местонахождения экземпляров этого файла на различных узлах ВО. Сервис DRS ведет реестр данных реплик в грид-сегменте и позволяет обнаруживать новые источники репликации одних и тех же данных и поддерживает передачу данных клиенту из множественных источников.

	призваны помочь в разрешении коллизий противоречивости данных при работе с федеративными коллекциями в рамках ПД.	
Компонент раскрытия	Этот компонент DSSP должен обнаруживать участников пространства данных, создавать связи и оказывать помощь в совершенствовании построенных связей. Также одной из функций данного компонента должно являться создание начальной классификации на основе моделей данных и контента. В дальнейшем, эти связи должны проверяться и уточняться человеком.	Базовый информационный сервис MDS позволяет обнаруживать грид-ресурсы (сервисы) ВО и формирует реестр этих сервисов, который допускает простейшую классификацию сервисов на сервисы работы с данными, вычислительные сервисы и т.д.
Компонент расширения	У некоторых участников ПД могут отсутствовать существенные функции управления данными. У DSSP должны иметься средства наполнения такого участника дополнительными возможностями, такими как схема, каталог, поиск по ключевым словами и мониторинг обновлений. Заметим, что может оказаться необходимо обеспечивать эти расширения "по месту", поскольку могут иметься существующие приложения или потоки данных, рассчитанные на имеющиеся форматы или справочные структуры.	Инструментарий GT4 содержит целый класс библиотек и инструментальных компонент, реализующих общую исполняющую среду (Common Runtime). Эти компоненты общей исполняющей среды обеспечивают web и pre-web службам возможность работы, независимо от среды исполнения (hosting container), а также позволяют наращивать функциональные возможности web-служб. Так как усиление функционала служб при помощи C, Java, Python WS Core, не нарушает их WSRF спецификации, то такие службы могут быть временными и запускаться на любом выбранном грид-узле, входящем в ВО, "по месту" при помощи протокола доступа к управлению ресурсами (Grid Resource Allocation and Management, GRAM) и службы Gatekeeper, которые обеспечивают безопасное создание удаленных процессов и управление ими.
Надежная передача данных	В DSSP не предполагается выделения отдельной компоненты передачи данных. Предполагается, что сама функция передачи и связанное с ней QoS реализуются средствами телекоммуникационной среды, поверх которой и развернуто ПД.	В инструментарии GT4 этому уделено серьёзное внимание. Существуют базовые сервисы надежной передачи файловых данных – Reliable File Transfer service (RFT) и GridFTP. RFT является сервисом передачи данных, базирующемся на стандарте SOAP over HTTP, и обеспечивает расширенные функциональные возможности протоколу GridFTP. GridFTP базируется на популярном в сети Интернет протоколе FTP. Адаптация его к грид-среде осуществлена за счет интеграции поддержки GSI-аутентификации, реализации частичного доступа к файлам, доступа к содержимому файла со смещением, поддержка маркеров успешной записи блока байтов на диск и проч. Сервис RFT позволяет контролировать, и расширяет, функции управления передачей данных по протоколу GridFTP.

Табл.1

Из приведенной выше таблицы видно, что хотя инструментарий GT4 и имеет большинство необходимых компонент, но не реализует всего функционала DSSP, описанного в [1, 2]. Однако этот подход обладает мощным механизмом расширения, позволяющим наращивать необходимый функционал. Его можно наращивать как за счёт создания новых сервисов более высокого уровня, так и за счёт увеличения функциональности набора базовых сервисов. Например, можно использовать механизм расширений, встроенный в OGSA-DAI. Этот

механизм заключается в описании новых, дополнительных действий (activities), реализующих заданную функциональность и в возможности выстраивать из них цепочки, за счёт организации информационных потоков внутри сервиса. Причём в Грид-среде такие действия могут выполняться как одновременно (параллельно), так и последовательно.

Следует заметить, что схема взаимодействия сервисов через документо-ориентированный интерфейс отнюдь не единственная. Сервисы могут осуществлять промежуточное хранение данных, в том числе служебных, создавать временные информационные объекты, временные экземпляры служб, и через них взаимодействовать друг с другом, а могут создавать потоки данных, как например, при создании Library Proxy Service в проекте Библиогрид [16] – сервиса, работающего по технологии OGSA с Web-контентом.

GT исходно создавался как среда исполнения сервисов и, следовательно, ключевыми инструментами в этой среде являлись протокол удалённого запуска процедур (GRAM) и такие средства, как Community Scheduler Framework, а также очень мощные средства безопасности, включающие аутентификацию, авторизацию и делегирование.. Однако, будучи дополненной средствами управления данными на базовом уровне (OGSA-DAI), механизмом их взаимодействия и развития, а также набором сервисов более высокого уровня, middleware Грид может рассматриваться, как инструмент создания прототипа DSSP. Такие Грид-сегменты также могут рассматриваться и как полигон для оценки компромиссов в терминах качества и эффективности при сосуществовании источников данных с различными моделями хранения.

4 Сервисы для работы с ПД ВО в Грид-среде

Очевидно, что на базовом уровне следует создавать лишь общие и очевидно простые (безальтернативные) с точки зрения реализации действия. В тоже время, для эффективного создания серьёзных прикладных технологий в ПД таких средств явно недостаточно. Практика поддержки реальных ВО говорит о необходимости создания так называемых многоцелевых инфраструктурных (intermediate) сервисов промежуточного уровня [17], опирающихся на базовые сервисы и являющихся в свою очередь необходимым строительным материалом для высокоуровневых сервисов. Рассмотрим несколько групп таких сервисов.

1. Сервисы, поддерживающие семантические связи в ПД ВО.

Как уже отмечалось выше, в ПД требуется семантическая интеграция до того, как могут быть обеспечены какие-либо прочие услуги. Однако, единая глобальная схема, которой соответствуют все данные, привычная в технологиях СУБД, скорее всего отсутствует. В то же время, система должна знать некоторые точные взаимосвязи между концептами предметной области, вокруг которой формируется ВО. Формализация такого знания требует предварительной работы экспертов, иногда значительной.

Примерами подобных сервисов могут служить сервисы создания и работы с онтологиями [18]. В ситуации, когда концепты онтологии, описывающие некоторую предметную область, указывают на информационные объекты в ПД, можно говорить о том, что сервисы работы с такими структурами, по сути, задают глобальную схему интеграции данных.

Безусловно, по мере функционирования ВО исходная онтология должна претерпевать изменения, эволюционировать. В этом отношении интерес может представлять подход, ориентированный на сервисы создания авторских онтологий [19], которые дают возможность уйти от абсолютизации исходных схем.

2. Сервисы - анализаторы, устанавливающие связи между данными, составляющими коллекции ВО.

Может быть, точнее следовало бы говорить, что связи между этими данными не были выявлены во всём их многообразии на предварительном этапе или мы исходно пренебрегали частью этих связей ради некоторых целей. Например потому, что не зная всего многообразия запросов, мы не хотели тратить ресурсы на хранение и поддержку этих связей.

3. Сервисы поиска и агрегации данных.

В эту группу попадают сервисы формирования информационной структуры запросов на поиск информации в пространстве слабо-связанных данных и собственно сервисы добычи и агрегации данных, прежде всего из внешних для ВО источников, обеспечивающие создание актуальных отображений по всему профилю связанных данных в соответствии с критериями отбора и агрегации.

Примером такого сервиса в GT4 может служить The Grid Distributed Query Service (GDQS), интеграционный сервис от OGSA-DAI, поддерживающий OQL в качестве сквозного языка запросов.

4. Сервисы хранения.

В качестве одного из примеров укажем сервисы-репликаторы, кэширующие данные участников ВО и обеспечивающие эффективную и надёжную работу с ними. Другим примером могут служить сервисы для создания и хранения различных метаданных, например сервис Репозитория Метаописаний (РМО) и сервис создания авторских коллекций данных [20].

5. Передача данных.

Эти сервисы призваны обеспечить как эффективную и надёжную передачу данных, так и возможность передачи сегментов одного элемента данных из разных источников, использование

многих хостов для передачи одних данных, использование сертификатов в системе аутентификации, осуществление многосторонних передач, частичное обращение к файлам, передачу файла по частям и др.

6. Сервисы лексического анализа текстов.

В этом направлении ведётся достаточно большое количество работ, но наибольший успех достигнут в направлении, предусматривающем выделение информационных блоков, соответствующих некоторым наперёд заданным схемам. Например, в следующем разделе будет подробно описан пример, в котором такой сервис выполняет выделение данных, связанных с конкретными факторами риска, выявленными в результате клинического применения некоторого фармпрепарата.

7. Сервисы фильтрации.

Эти сервисы могут быть очень разнообразными, в том числе могут использоваться для отбора необходимой пользователю информации из слабо-связанных данных по семантически-обусловленным критериям, или работать с внутренними хранилищами информации, такими как PMO или хранилища промежуточных реплик. Примером таких сервисов могут быть сервисы персонализации медицинской информации в медико-биологических ВО.

В некоторых случаях фильтр позволяет уточнять семантическую модель. В зависимости от запроса фильтр разбивает пространство данных на классы и выделяет ссылками связи между ними, далее строится подмножество классов (domain-model), которое зависит от качества самого запроса и степени компетентности запрашивающего. Такой динамически фильтрующий сервис над пространством данных позволяет как улучшать качество ответов, так и продуцировать новую ценную информацию.

8. Визуализация.

Это сервисы представления части данных в удобном для понимания и анализа виде, таком который удовлетворяет критерию comprehensibility – минимизации интеллектуальных усилий на понимание. Очевидно, что одну и ту же информацию надо представлять различным образом специалисту, для которого актуален принцип "взглянул - и сразу стало ясно" и учащемуся, которому важнее принцип Декарта о минимизации умственных усилий, затрачиваемых на единицу прочно усваиваемых знаний, умений и навыков.

9. Сервисы, отвечающие за учёт параметров качества (QoS) и за реализацию параметров качества.

Например, это могут быть анализаторы, определяющие степень качества ответа при поиске и запрашивании в ПД. Дело в том, что работа в ПД, в отличие от традиционных СУБД, в обязательном порядке порождает коллизии или неточности. Эти коллизии могут проявляться в различных вариациях, например, недостоверность ответа на запрос к ПД, различие результатов поиска во времени, связанные с недоступностью одного из участников пространства, определение степени актуализации ответов и релевантности их запросу и др. Для разрешения всех вышеописанных коллизий необходимо вводить критерии или метрики, определяющие качество. Критериями QoS могут служить как время обработки запроса, точность ответа, степень релевантности его, полнота ответа так и другие.

Различные виды несогласованностей предусматривают разные способы разрешения, однако основной вариант – это анализ происхождения данных, с которыми оперирует компонент поиска и запрашивания, на каждом этапе формирования ответа.

Данная статья не претендует ни на описание полноты набора сервисов, ни на выбор их базового набора. Более того, понятно, что чаще с практической точки зрения бывают востребованы сервисы более высокого уровня, сочетающие названные выше. Учитывая, что архитектура сервисной среды организована таким образом, чтобы можно было выстраивать композиции из нескольких сервисов, используя механизм обмена XML документами, как это принято в OGSA-DAI, то задача построения высокоуровневых сервисов резко упрощается. При этом, безусловно, важно следовать стандартам обмена информацией, например для сервисов работы с онтологиями в качестве межсервисного стандартизованного интерфейса вполне применим OWL [20].

В следующем разделе рассмотрим пример сервиса высокого уровня, использующего несколько intermediate сервисов, в соответствии со схемой, показанной на Рис.1.

Следует заметить, что реализация этого высокоуровневого сервиса без использования Грид-технологий была бы затруднительна по двум причинам: первая – это вычислительная сложность процедуры лексического анализа, а стандартные средства GT4 позволяют распределять эту задачу по наиболее подходящим узлам ВО; второе – отсутствие необходимости в централизованном диспетчере.

Этот высокоуровневый сервис разработан для обеспечения информационной поддержки проведения исследований по разработке вакцин, в том числе для исследований in silico, в Russian ChemBioGrid [21]. Сервис осуществляет поиск и обработку информации о клинических испытаниях фармпрепаратов с целью выявления новых свойств, обобщения клинического опыта применения и доведения его лечащих врачей и пациентов. Однако очевидно, что он может применяться и в ВО целого ряда других понятийных сетей.

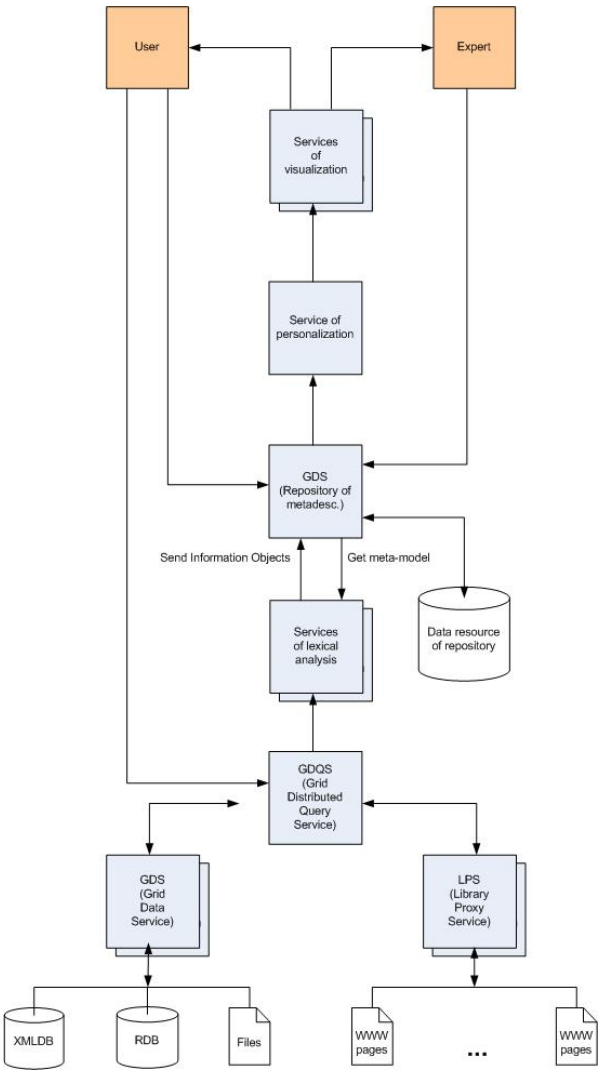

Рис.1

5 Пример из области доказательной фармакологии

Опыт клинического применения лекарственных препаратов накапливается в виде научных статей, отчетов и презентаций в многочисленных и разнообразных информационных ресурсах. О восстребованности этой информации свидетельствует, например, активное использование Web-сайта ClinicalTrials.gov, который принимает около 31000 посетителей ежедневно.

Одна из серьёзных проблем использования этой информации заключается в том, что представление информации в существующих электронных ресурсах опирается на информационные структуры, повторяющие структуру текстовых документов, регламентирующих состав и форму описаний фармпрепаратов и лечебного процесса. При этом поиск информации в ПД осуществляется обычными поисковыми машинами, и в результате пользователь получает крайне сложные для анализа и восприятия данные.

Создание новых высокоэффективных информационных сервисов и наряду с этим персонализация процесса анализа данных является одним из путей снижения рисков, связанных с неадекватным медикаментозным лечением, и увеличения безопасности пациентов. Мы полагаем, что одним из преимуществ доказательной фармакологии будет являться именно возможность персонификации медикаментозной терапии благодаря тому, что врач, фармацевт, пациент получат доступ к агрегированной актуальной информации по клиническому применению лекарственных препаратов.

Для построения таких сервисов, на предварительном этапе, мы предлагаем задавать семантическую схему (мета-модель) интересующей нас информации. Наиболее удобной формой этой мета-модели является XML-представление информационных структур, которые используют набор тегов для организации семантики отношений между данными и метаданными в описании фармпрепарата. Этот набор тегов определяет критерии поиска и выделения данных из информационных источников.

Вводить описанные мета-модели можно как при помощи уже хорошо апробированных инструментариев работы с онтологиями[18], так и используя библиотечные тезариусы различных предметных областей, или в простейших случаях, используя шаблоны.

При накоплении данных о клиническом применении фармпрепарата могут использоваться различные мета-модели в зависимости от задачи, для которой осуществляется информационная поддержка. Близким примером могут служить стандарты в области формирования стратегий лечения [22], также опирающиеся на XML-схемы.

Как правило, XML-описания такого рода включают в себя значительное количество тэгов, описываемых в DTD, однако в данном примере мы ограничимся рассмотрением набора тегов, относящихся к клиническому применению лекарственных препаратов.

Мета-модель, приведенная на рис. 2, позволяет создавать информационные объекты (ИО), содержащие данные по клиническому опыту, связанному с возникновением побочных эффектов при медикаментозной терапии.

```
<Drug Name>
    <Adverse Event>
            <Risk Factor >
            <Data> … </Data >
            </RiskFactor >
    </AdverseEvent>
            <URI > ..... </URI >
</DrugName>
```

Рис. 2.

Модель предусматривает задание наименования препарата (DrugName), конкретного побочного эффекта (AdverseEvent) и фактора риска (RiskFactor). Базовый набор данных для такой модели содержится в описании препарата в Регистре лекарственных средств. Например, для побочных эффектов при применении варфарина описаны следующие основные побочные эффекты: кровотечение (bleeding), дисфункция печени (liver disfunction), диарея (diarea), некроз кожи (skin necrosis). Основные факторы риска, как правило, также известны. Например, для варфарина это возраст, артериальная гипертензия, дисфункция почек или печени.

Важно заметить также, что использование XML-представления позволяет естественным образом наращивать описание фармпрепарата. Например, если список известных побочных эффектов расширяется экспертом или сервисом лексического анализа, то все сервисы работы с ИО будут работать с этими данными точно так же, как и с первоначальными, базовыми.

Поскольку целью сбора, интеграции и представления информации пользователю является персонализация знаний о клиническом применении лекарственных препаратов, то ИО должен включать данные о параметрах пациента или выборки пациентов, для которых получены эти данные. На рис. 3 показан пример информационного объекта, включающего параметрические данные, полученные в результате анализа одной из статей, посвященных опыту клинического применения варфарина при антикоагуляционной терапии и возникших при этом побочных эффектах.

<Drug Name>Warfarin
<Adverse Event>Bleeding
<Risk Factor >Age
<Value>53</Value>
<Data>A 53-year-old man experienced two episodes of skin necrosis on his left flank and buttock, following the initiation of warfarin therapy for acute thrombophlebitis and after a dose adjustment. Seven days after the initiation of warfarin therapy, an area of erythema surrounded by a halo was noted on the left thigh of a 79-year-old woman.**</Data >**
<Value>70**</Value>**
<Data>A 70-year-old man was given a warfarin dose of 10 mg daily that was reduced to 2.5 mg daily. It was discontinued when bullous violaceous lesions were discovered on his lower left leg and foot.**</Data> …**
</Risk Factor >
</Adverse Event>
<URI>http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7030072&dopt=Abstract**</URI >**
</Drug Name>

Рис. 3.

Основной особенностью поисковой части сервиса (GDQS + GDS + LPS) следует считать необходимость постоянной актуализации и пополнения той информации, которая соответствует запросу (см. рис. 1). В этом смысле сервис исполняет роль агента, следящего за возможными источниками в ПД ВО, Интернет и в доступных корпоративных информационных системах. ИО, выделенные сервисами лексического анализа из данных, найденных поисковыми сервисами в информационном пространстве в соответствии с заданной мета-моделью, сохраняются в РМО и могут в любой момент быть доступны участнику ВО. Представление этих ИО в подходящем для потребителя виде может осуществляться как через специализированные сервисы визуализации – например сервисы работы с онтологиями, так и стандартными средствами браузера – наложением соответствующих XSL-схем.

Накапливаемые данные и метаданные доступны не только пользователям – врачам, пациентам и фармацевтам, но также и экспертам - участникам ВО. Эти эксперты, используя те же сервисы, могут (и должны) осуществлять уточнение, проверку непротиворечивости и более полную формализацию накопленных метаописаний. В связи с этим наличие и возможность использования различных сервисов персонализации и визуализации информации является важным преимуществом описанной технологии в целом.

6 Заключение

Из приведённого примера видно, что сервис-ориентированная грид-технология предоставляет новые возможности, которых не было в сетях, организованных по схеме peer-to-peer или клиент-сервер. Функционирование и взаимодействие сервисов близко к технологии мульти-агентных систем, а роль интеллектуальных агентов выполняют грид-сервисы. При этом, в данной задаче они имеют ряд преимуществ перед web-сервисами. Среди них: возможность реализации функциональности поиска данных, не ограниченной набором процедур, реализованных на сервере хранилища данных; возможность проводить анализ как в публичных, так и в корпоративных сетях; возможность продолжения работы сервисов-агентов поиска и сбора данных и после выполнения конкретного запроса; встроенная возможность передачи прав доступа к данным от пользователя ко всей последовательности грид-сервисов с помощью цифрового сертификата. Последнее особенно важно для e-Health ввиду чрезвычайной чувствительности медико-биологических данных к несанкционированному доступу.

Сервис-ориентированный подход к слабо связанным массивам данных как к пространству данных позволяет уже сегодня создавать сервисы нового уровня, оперирующие не только БД или

метаданными, но и работающие непосредственно с Web-данными и другими слабо-структурированными ресурсами. При этом ощущается серьёзная потребность применения доработанных технологий СУБД для пространств данных, что в сумме позволяет с новых позиций решать проблему интеграции федеративных гетерогенных информационных ресурсов.

Литература

[1] M. Franklin, A. Halevy and D. Maier: >From Databases to Dataspaces: A New Abstraction for Information Management. ACM SIGMOD Record 34, No. 4 (December 2005), pp.27-33.

[2] M. Franklin, A. Halevy and D. Maier: Principles of dataspace systems. ymposium on Principles of Database Systems archive. Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems table of contents Chicago, IL, USA 2006 ISBN:1-59593-318-2

[3] Foster, C. Kesselman. The Grid: Blueprint for a New Computing Infrastructure // Morgan Kaufmann Pub., San Francisco, CA. 1999.

[4] EGEE Web site: http://www.eu-egee.org/

[5] OSG Web site: http://www.opensciencegrid.org/

[6] RSS in Wiki Web site: http://en.wikipedia.org/wiki/RSS_(file_format)

[7] Atom Web site: http://www.atomenabled.org/

[8] GData Web site: http://code.google.com/apis/gdata/index.html

[9] Foster I., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations International Journal of High Performance Computing Applications, 15 (3). 200-222. 2001.

[10] Жучков А.В., Твердохлебов Н.В., Арнаутов С.А., Голицын С. От информационной системы проекта (учреждения) к электронной библиотеке в понятийной сети. Труды V Всероссийской объединенной конференции «Технологии информационного общества – интернет и современное общество». Санкт-Петербург, 25-29 ноября 2002 г., с.91-94.

[11] IBM Web siteb: http://www-306.ibm.com/software/data/integration/

[12] Foster I., Kesselman C., Nick J., Tuecke S. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration Global Grid Forum, June 22, 2002

[13] GT Web site: http://globus.org/toolkit/docs/4.0/

[14] OGSA-DAI Web site: http://www.ogsadai.org/

[15] IBM Web site: http://www-128.ibm.com/developerworks/grid/library/gr-feddata/

[16] Жучков А.В. Библиогрид – основные элементы концепции и реализации/ Эктронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL2005). Ярославль, 4-6 октября 2005 г. – Ярославль: Ярославский государственный университет им.П.Г.Демидова, 2005, с 31-37. ISBN 5-8397-0400-8 УДК 681.3 ББК 32.973 Э 45

[17] А.В. Жучков, Н.В. Твердохлебов. Разделение ролей в предоставлении грид-услуг – новые возможности для разработки сервисов и инновационной деятельности. / Распределенные вычисления и Грид-Технологии в науке и образовании: Тез. докл. 2-й междунар. конф. (Дубна, 26-30 июня 2006 г..). – Дубна: ОИЯИ , 2006. – 151-152 сс. ISBN 5-9530-0117-7

[18] Joutchkov A., et al. Ontology-driven data integration and datamining in dynamic heterogeneous data sources. In Proc. of : “Digital Libraries: Advanced methods and Technologies. Digital Collections”. Sixth National Russian Research Conference, Pushchino, 29.09 – 01.10, 2004. Institute of Mathematical Problems of Biology RAS.

[19] Joutchkov A. et al. Grid-based Onto-Tecnologies provide an effective instrument for Biomedical research. From Grid to Healthgrid. T. Solomonides et al (Eds.). IOS Press, 2005, pp.37-46.

[20] Когаловский М.Р. Стандарты XML и электронные библиотеки // Электронные библиотеки, 2003, Т. 6, Вып. 2.

[21] Zhuchkov A. et al. Advancing of Russian ChemBioGrid by bringing Data Management tools into collaborative environment. Challenges and Opportunities of HealthGrids. V. Hernandes et al. (Eds). IOS Press, 2006, pp. 179-186.

[22] ASTM E2210-02 Standard Specification for Guideline Elements Model (GEM)-Document Model for Clinical Practice Guidelines..

SERVICE-ORIENTED GRID-APPROACH TO MAINTAIN DATA SPACES OF VIRTUAL ORGANIZATIONS

A. Zhuchkov, A. Kravchenko, N. Tverdokhlebov

This article covers the adaptation of some facilities of Grid-technology in order to form the Data Spaces support platform. It shows also an example of the constructing of a high-level service which operates in the subject-oriented Data Space of a medical Virtual Organization.