RCDL'1999–RCDL'2008: DL, VDL, Semantic Web/GRID, GRID…

© В.П. Шириков

Объединенный институт ядерных исследований (ОИЯИ)
shirikov@jinr.ru

Аннотация

Данная статья должна восприниматься как личный короткий авторский обзор достижений в области разработки и реализации электронных библиотек, представлявшихся участниками ежегодных конференций RCDL (в основном российскими) в последние 10 лет. Все оценки и соображения в тексте обзора приведены в соответствие с личным авторским пониманием известных технологий, необходимых для реализации электронных библиотек и в других областях e-Science и ее приложений.

Наша конференция – заключительная в 10-летнем цикле российских конференций по электронным библиотекам (DL). Еще на самой первой 1999-го года в разных докладах само понятие DL трактовалось по-разному: либо только как «зеркальное» Интернет-отображение ресурсов традиционных библиотек и средств их использования (как это было в докладе авторов из Технологического Университета в Мюнхене), либо уже в той расширенной трактовке, которая тогда была дана в докладе М.Р. Когаловского и преобладает сейчас: как хранилища (коллекции) знаний и данных общего и специального назначения, реализованные в сетевой среде с применением современных технологий и стандартов формирования, модификации и средств использования информации. Теме создания и использования информационной модели, способной отобразить структуру и семантику таких гетерогенных коллекций, на этой же конференции был посвящен доклад Л.А. Калиниченко "Integration of Heterogeneus Semistructured Data Models in the Canonical One": сразу заметим, что своеобразным логическим продолжением данной работы из последних стала доложенная с соавторами на конференции GRID`2008: "Application Driven Mediation Middleware Grid Infrastructure for Problem solving over multiple Heterogeneus Distributed Information Resources". К 1999-му году Web с первыми версиями Semantic Web был достаточно оснащен технологиями, минимально необходимыми для реализации распределенных информационных систем, в том числе виртуальных DL и коллекций разного типа, поэтому во второй половине 1990-х годов и в последующем выполнено много работ по Web-ориентированным системам DL, доложенных, в частности, на конференциях RCDL и «Научный сервис в сети Интернет» (можно упомянуть хотя бы работы В.А. Серебрякова и его соавторов по созданию инфраструктуры единого научного информационного пространства РАН). Среди необходимых или признанных технологий реализации в последние годы и российскими, и иностранными участниками RCDL (см. презентацию T. Risse для конференции RCDL'2007, http://rcdl2007.pereslavl.ru/en/doc/Risse_rcdl_tutorial.pdf), упоминаются следующие:

Grid, Pear-to-Pear, SOA, Semantic Web с его средствами применения онтологий. В указанной презентации подчеркивается тенденция перехода от понятия DL как интегрированной системы с централизованным управлением к динамически конфигурируемой федерации DL-сервисов и информационных коллекций, когда возникает понятие виртуальной DL как коллекции данных от разных контент-провайдеров без централизованного управления и с распределенными гетерогенными сервисами. Естественно, что предоставляемые каждым провайдером сервисы должны поддерживать поисковые средства для различных типов информации (разных типов мультимедийных данных, поиска по контексту, многоязыкового доступа), индексирование, аннотирование документов, предоставление регистров метаданных и ресурсов; в составе DMS виртуальной DL могут ресурсоемкие сервисы (например, для извлечения характеристических деталей из image/audio/video-документов, генерации метаданных, автоматического редактирования сложных документов). С учетом перечисленных выше требований и был реализован, в частности, проект BRICKS (http://www.brickscommunity.org), научным и техническим директором которого является T. Risse. В своей презентации он останавливается в основном на характеристике концептуальной модели и ее реализации в проекте, прикладной задачей стала интеграция ресурсов в общую и разделяемую DL, куда входят в качестве локальных материалы электронных музеев, архивы исторических документов и другие виды электронной памяти (цифрового наследия культурного многообразия: Building Resources for Integrated Cultural Knowledge Services); узлы (сайты) созданной по проекту сети BNnet оснащаются свободно доступными программными интерфейсами-«кирпичами» (bricks), через которые без централизованного управления они

взаимодействуют друг с другом и используют доступные ресурсы для работы с контентом и метаданными по принципу «равный с равным» (P2P); каждый узел может знать напрямую только одно подмножество других узлов, но если хочет использовать внешние по отношению к этому подмножеству ресурсы, то может послать запрос к одному из известных ему и тот сформвердирует запрос. T. Risse не акцентирует свое внимание на возможном применении своих средств в рамках Grid-структур, по существу это выглядит как Web-ориентированная реализация, сама по себе не нуждающаяся в явном виде в каких-то сервисах из набора Grid-middleware, поэтому упоминание им Grid-технологий в указанном списке можно понимать по-разному.

Моей целью не был обзор тех многочисленных работ российских авторов, которые решали чем-то похожие задачи с применением технологий из приведенного списка и докладывали их на прошедших конференциях RCDL, и я хочу только в основном остановиться на отношениях в развитии программных средств Grid-middleware и средств DL/VDL (т.е. как они сложились в прошедшие годы), поскольку их независимое использование уже сейчас в целом ряде случаев оказывается или неоправданным, или невозможным.

Так получилось, что хотя к тому же 1999-му году концепция Grid и его технологий была сформулирована и появились первые реализации его middleware в пакетах типа GT1, но она была представлена как концепция использования распределенных вычислительных ресурсов (ВР) и в значительной мере предусматривала только создание нового уровня системы распределенной потоковой обработки задач: Web как информационную систему и Grid как вычислительную систему объединяло на технологическом уровне только общее использование сетевых протоколов нижнего уровня, поэтому изначально системное Grid-middleware своими сервисами обеспечивало только поиск компьютерных вычислительных ресурсов для выполнения задачи (job, понятие не из области информационных систем) в соответствие с ее требованиями в сопровождающем ее JDL-файле. Конечно, большинство задач порождает или использует какие-то данные, сосредоточенные в DBMS или файловых системах, поэтому выполнение требований на предоставление таких данных задаче могло частично выполняться сервисом DMS из состава базового middleware, но в основном даже и сейчас работа с распределенными данными – это проблема приложений, пользующихся возможностями того сервисного программного обеспечения, которое "on top of Grid-middleware": к нему можно отнести, например, сервисы OGSA-DAI/DQP, изначально разработанные для пользователей британской архитектуры GridPP и AstroGrid и включенные в пакеты уровня GT3/GT4, а также использованные для работы с распределенными базами данных. Конечно, развитие технологий распределенных информационных систем в чем-то обязано той стандартизации, которая при создании Grid-архитектур была отработана для формирования VO/IVO и создания SOA, да и ПД (Пространств Данных). В материалах прошлогодней конференции RCDL'2007 хотелось бы в связи с этим отметить работу А.В. Жучкова, А.В. Кравченко и Н.В. Твердохлебова (авторов из ИХФ РАН) «Сервис-ориентированный Грид-подход к информационным задачам в пространствах данных виртуальных организаций». На примере перечня требований к системе DSSP (Data Space Support Platform), охарактеризованной в материалах ACM SIGMOD в 2005–2006 гг., авторы указанной работы подробно расписали те средства, которые заложены в рамках пакета GT4 и могли бы использоваться достаточно эффективно для реализации многих элементов перечня. Рассматривая пример разработанного ими сервиса высокого уровня для обеспечения информационной поддержки проведения исследований по разработке вакцин, они отмечали, что его реализация без использования Grid-технологий была бы затруднительна по двум причинам: вычислительная сложность процедуры лексического анализа (здесь средства CT4 позволяют распределить работу по подходящим узлам VO) и отсутствие необходимости в централизованном управлении. Заметим, что и в этой работе отмечается роль сервисов OGSA-DAI/DQP, независимо успешно применяемых также в совместных работах ИПИ РАН (В.Н. Захаров, Л.А. Калиниченко) и SAO РАН в рамках деятельности IVOA и RVO: эти работы также докладывались на наших конференциях RCDL и симпозиуме по онтологическому моделированию, упоминаемому в конце обзора. Кстати, значимость и признание полезности упомянутых сервисов подтверждается и тем, что они были использованы в продукте IBM Sphere Information Integrator для обеспечения доступа к виртуальной базе данных (как набору физических разнотипных СУБД) через универсальный SQL или XML-запрос, автоматически преобразуемый в формат, понимаемый той или иной физической СУБД: в этой работе помогают программные Java-посредники Wrappers между IBM Sphere и источниками данных. Ключевой идеей «виртуализации» реализации подобного "Data Grid" была естественная необходимость избавить прикладную задачу от необходимости знать – где физически размещены данные, в которых она нуждается, и как они структурированы: это отмечается и в докладе на RCDL'2007 авторов из ИФХ РАН, и независимо в моем обзоре на прошлогодней конференции «Научный сервис в сети Интернет».

Говоря об информационных сервисах "on top of Grid-middleware", сосредоточенного в пакетах типа GT4 и обслуживающего разные Grid-архитектуры, нельзя не остановиться о том, что происходит в структуре EGEE (European Grid for E-sciencE), в рамках которой функционирует и развивается и структура RDIG (Russian Data Intensive Grid). Изначально прикладными областями, ради которых начинал развиваться EGEE, были определены науки о Земле, физика высоких энергий, биоинформатика и медици-

на, астрофизика. Сейчас примерно 68% пропускной способности EGEE и его ресурсов тратится на задачи, связанные с подготовкой к запуску в ЦЕРН (Женева) нового ускорителя LHC и основных экспериментальных установок на канале его пучков (т.е. на физику высоких энергий). Учитывая уникальность готовящихся экспериментов и заинтересованность в них стран-участниц ЦЕРН и его коллаборантов во многих странах мира (в том числе России), большинство существующих в них Grid-структур старается реализовать средства совместимости с EGEE и его базовым программным Grid-middleware (к июлю 2008 г. это стал набор gLite версии 3.1). Его разработка частично шла с учетом имевшегося в пакетах GT, но многое создано в рамках деятельности групп в ЦЕРН как головной организации по проекту EGEE, регулирующей все отношения с консорциумом из 91 представителей от институтов и стран, участвующих в проекте и комплектовании его ресурсов. Распределенные компьютерные ресурсы и средства хранения информации EGEE позволяют обрабатывать сейчас ежедневно более 100 000 задач, к моменту запуска LHC ожидается как минимум удвоение их возможностей. Будучи ориентированным разработчиками изначально на Linux-платформы, EGEE приспосабливается и к использованию Windows потребителями его ресурсов. Поскольку возможности EGEE и его базового gLite-middleware уже представили значительный интерес для разработчиков ресурсоемких информационных систем, хотелось бы обратить внимание на два связанных между собой проекта, которые имеют кураторов в группе EGE (ЦЕРН) и ESA (European Space Agency). Речь идет о проекте DILIGENT (A **DI**gital **L**ibrary **I**nfrastructure on **G**rid **EN**abled **T**echnology, http://www.diligentproject.org) и его системе сервисов "gCube on top of gLite", техническим и научным кооординатором которого является Донателла Кастелли из CNR-ISTI (Пиза, Италия), и его продолжении с 2008 г.: проекте D4Science (http://www.d4science.eu). Я хочу в данном обзоре как-то отобразить причину того интереса, который был проявлен к этим проектам со стороны руководства ПК RCDL'2008 и RCDL Steering Committee, планировавшего организацию тьюториала на нашей конференции.

Одной из первых прикладных целей DILIGENT было создание сервисов для проекта SAPIR (Search in Audio Visual Content Using Peer-to-Pear IR) как части проекта Chorus (средства Multimedia Content Search Engines), т.е.для задачи создания в интересах этих проектов нового типа представления и поиска данных, отсутствовавших в традиционно используемых поисковых системах типа Google и Yandex, ограниченных средствами текстовой формы запросов и текстового описания данных, ассоциирующегося с мультимедийным контентом. Среди примеров желаемого, приводимых на сайтах этих проектов – поиск для пользователя мобильного устройства информации об объекте после ввода в качестве запроса фотографии этого объекта или поиск текста песни по ее мелодии (заметим кстати, что проблема нахождения в базе данных по образу достаточно не новая и для российских авторов, достаточно упомянуть из последних работ хотя бы сообщение на нашей RCDL'2008 от авторов из Петрозаводского университета «Система поиска в электронной коллекции изображений петроглифов Карелии»). Сама по себе проблема подготовки базы мультимедийных данных для того, чтобы по описаниям и представлениям их фрагментов можно было удовлетворять такие запросы, например, для модифицированного Web-приложения Flickr (см. http://www.flickr.com/tour по использованию цифрового архива фотографий и обмена ими) связана с обработкой каждой фотографии для представления и описания характерных деталей (черт лица, например), и это достаточно ресурсоемкая задача с вычислительной точки зрения. Во второй половине 2007 г. с этой целью с применением сервисов "gCube on top of gLite" был проведен на инфраструктуре EGEE 16-недельный прогон (data challenge) по обработке 37 млн фотографий из on-line базы данных Flickr, сгенерировано около 112 млн текстовых и image-объектов (примерно 5TB данных); ресурс-брокером в EGEE PPS (Pre-Production Service) было пропущено более 44 тыс. задач (jobs), в каждой обрабатывалось примерно 1000 фотографий. Составленная коллекция (см. https://twiki.cern.ch/twiki/bin/view/DILIGENT/DiligentFlickrDL) и должна использоваться проектом SAPIR для развития новой крупномасштабной технологии контекстно-ориентированной классификации и выборки данных, расширяющей пределы возможностей традиционных поисковых механизмов. Почему DILIGENT попал в сферу интересов группы EGE ЦЕРН? Да может хотя бы потому, что исходная информация о событиях, регистрировавшихся на экспериментальных установках при ускорителях, изначально представлялась в виде обычных фотографий и затем их электронных подобий, а астрономия и астрофизика до сих пор используют фотографии, которые ежемесячно публикуются в издании CERN Courier или размещаются на сайте www.sai.msu.su/apod, но это уже мои домыслы… Во всяком случае, в программе прошедшей в октябре 2007 г. международной конференции EGEE'2007 была и секция по проекту DILIGENT как средству для интеграции Grid и DL, что, как отмечалось, заложит основы нового поколения инфраструктуры использования знаний в электронизированной науке (e-Science) для разных областей исследований и в промышленности. Интерес ESA к EGEE и сервисам типа "gCube on top of gLite" также понятен: накапливается спутниковая информация о Земле, лет через 10 это примерно 12 Петабайт данных, нужна оперативная информация для мониторинга и систем охраны окружающей среды, контроля состояния ее ресурсов и результатов их использования, а это, в частности, одна из главных задач упомянутого проекта D4Science, базирующегося на использовании сервисов DILIGENT. Отмечу сразу, что на прошлой конференции RCDL'2007 был представлен родственный по теме доклад Е.Б. Кудашева и А.Н. Филонова «Технология и стандарты интеграции сервисов и баз данных дистанционного исследования Земли из космоса», и сейчас

на RCDL'2008 в докладе тех же авторов делается обзор текущего состояния развития технологий в области распределенных информационных систем для данных дистанционного зондирования, а также возможные пути развития подобных систем в ближайшие годы, технические аспекты реализации систем трех поколений: от проекта первого поколения INTAS IRIS с участием авторов доклада до инициативы GMES HMA, находящейся в стадии реализации и сфере интересов и NASA, и ESA, и прикладной науки о Земле.

Теперь немного более подробно о сервисах gCube (см. материалы по ссылкам http://www.gcube-system.org/architecture/overview.html и http://www.gcube-system.org/architecture/services/services.htm).

В самом общем определении gCube позволяет исследователям динамически, по требованию (on-demand) создавать информационно-вычислительные среды (Virtual Research Environments, VREs), агрегируя и формируя контент-ресурсы, прикладные сервисы и компьютерные ресурсы как за счет собственных у прикладных проектов, так и за счет имеющихся в EGEE. В составе gCube достаточно средств для мониторирования использования разделяемых ресурсов с гарантией их оптимального распределения и эксплуатации, а также легкого создания Web-порталов для VREs, через которые пользователи могут иметь доступ к контенту и сервисам; предоставляется также набор типовых для DL функций (поиск, аннотирование, формирование, визуализация документов и др.).

Не все виды сервисов были изначально подробно описаны в указанных выше разделах сайта, например, сервис vdlgenerator (Virtual Library Creation and Management), дающий возможность пользователям или их сообществам создавать собственные DLs: он позволяет специфицировать набор критериев, отражающих ожидаемые характеристики новой DL, и затем определяет сервисы и информационные источники, нужные для обеспечения для обеспечения требуемых свойств DL и ее контента. Само формирование и функционирование DL обеспечивается выбором доступных DHNs (Diligent Hosting Nodes), Grid-узлов: они, как уже было сказано, берутся из собственных компьютерных ресурсов проектов или получаются от сервисов gLite (его брокера) с учетом требований на свойства (Property requirements) нужного ресурса (тип операционной системы, мощность CPU, размер дисковой памяти, наличие нужного software: например, конкретной библиотеки программ общего или специального назначения). Ключевую роль играют такие сервисы общего назначения (Collective layer services), как Keeper (с его распределяемым набором программных компонент, нужных для создания и поддержки функционирования динамических DL) и DIS (Distributed Information System): все ресурсы и их состояние регистрируются здесь и их статус при необходимости опрашивается и корректируется другими сервисами системы через универсальный API-интерфейс.

Возвращаясь к теме обзора и итога работ, выполненных российскими участниками конференций RCDL и опубликованных в трудах, доступных на сайтах этих конференций, нельзя не отметить, что многие их важные итоговые работы были доложены в мае этого года в Звенигороде на симпозиуме «Онтологическое моделирование: состояние и направления исследований и применения», организованном ИПИ РАН и Московской секцией ACM SIGMOD, поэтому судить о состоянии дел на текущий момент по проблематике RCDL нельзя без учета материалов этого симпозиума по ссылкам:

http://synthesis.ipi.ac.ru/synthesis/ontologyprogram
http://synthesis.ipi.ac.ru/synthesis/publications
http://www.cemi.rssi.ru/mei/articles/koga08-1.pdf

В заключение хочется отметить следующее личное впечатление: как-то подзабытой оказалась та концепция Semantic Grid, которая была сформулирована еще в 1999 г. Keith G.J effry и затем детализована в 2002 г. в статье David De Rouge et al "The Semantic Grid: a Future e-Science Infrastructure" (http://www.semanticgrid.org/documents/semgrid-journal/semgrid-journal.pdf), где авторы предсказывают, что программная среда компьютеризованной науки и все Grids должны будут включать в себя трехуровневую систему сервисов:

Data/Computation Services, средства размещения данных и их транспортировки между обрабатывающими программами, обеспечение вычислительных и сетевых ресурсов;

Information Services, средства представления, запоминания и доступа к информации, управление ею;

Knowledge Services, средства накопления, представления, обновления, «публикации» (сетевого распространения) знаний для помощи ученому в его исследовательском процессе.

Все положения демонстрируются детальным формализованным примером цикла полной автоматизации обработки экспериментальных данных в сетевой компьютерно-аппаратной среде (от начала поступления данных на анализ до подведения итогов результата обработки научным сообществом) с применением конкретного перечня сервисов каждого из указанных уровней, подчеркивается роль семиуровневой системы онтологий для нормального функционирования всей клиент-сервисной структуры приведенного примера.

Что-то до сих пор не видно и сейчас в нашей практике примеров подобной реализации.

RCDL``1999–RCDL'2008: DL, VDL, Semantic Web/GRID, GRID…

V.P. Shirikov

This article must be imaged as private author's short review of researches in the field of digital libraries, presented by RCDL community and some foreign specialists for our RCDL's in the last 10-years period. Of course, author's estimations in this review were done in the main in accordance with private understanding of technologies, nesessary to be taken in account for realizations of digital libraries and in other fields of e-Science and its applications.