Библиогрид – основные элементы концепции и реализации

Жучков А.В.
ИХФ РАН, АНО «ТЦ «Наука и общество»
alex@umos.ru

Аннотация

В статье рассматриваются основные положения принятого в проекте «Библиогрид» подхода к созданию технологической интеграционной среды для построения электронных библиотек, базирующейся на распределенных федеративных принципах.
Приводятся некоторые результаты, связанные с созданием исследовательского прототипа.

1 Введение

При обсуждении вопроса о создании фонда электронных документов или электронной библиотеки (ЭБ) внутри библиотеки традиционной, сегодня чаще преобладает «библиоцентричный» подход, базирующийся на переносе в компьютерную среду традиционных библиотечных технологий [1]. Включение в такой фонд разно уровневых объектов нарушает его важнейшие системные свойства, поэтому в центре внимания становятся целостные издания, представляющие собой экземпляры хранения, но не структурированная информация, представленная в них. Несомненно, такой подход имеет право на существование. Он позволяет найти место традиционным библиотекам в существующем цифровом информационном пространстве, не превратив их в дальнейшем в «Музеи книги». В то же время, присущий этому подходу консерватизм, немасштабируемость, отсутствие возможности аккумулировать максимально широкий круг разнородной информации и желание сделать ЭБ максимально удобным средством для читателя, предоставляя различные возможности для потребителей информации самостоятельно работать с содержанием электронных экземпляров хранения, заставляют обратить внимание на более «технократический» подход, который диктуется колоссальным развитием современных компьютерных технологий. В частности речь идёт о концепции ГРИД, которую можно определить как интегрированную единую среду распределенных ресурсов [2]. С точки зрения ЭБ в грид-подходе привлекает прежде всего возможность реализации концепции построения ЭБ как объединения разнородных, распределённых, федеративных коллекций.

Наиболее широко применяемое для создания такой среды программное обеспечение среднего уровня (middleware), разработанное альянсом крупнейших американских университетов и распространяемое в открытом коде - Globus Toolkit [3] включает в себя уже сегодня практически все базовые технологии, необходимые для создания реально работающих приложений, в том числе связанных с хранением и обработкой больших массивов данных.

Участники проекта «Библиогрид» [4] (ТЦ «Наука и общество» (ЮМОС), РГБ, ГИВЦ Роскультуры и ряд университетских библиотек) рассчитывают, что подход, ориентированный на применение грид-технологий успешно дополнит складывающуюся библиотечную парадигму и позволит эффективно решить целый ряд задач, среди которых:

- Создание технологической интеграционной среды, базирующейся на распределенных федеративных принципах. Обеспечение интероперабельности в гетерогенной компьютерной среде, и среди гетерогенных распределенных информационных ресурсов.
- Обеспечение технологического единства с информационными пространствами, создаваемыми в проектах Евросоюза, включая возможность участия в таких проектах.
- Эффективное использование имеющихся аппаратных ресурсов, на основе динамического использования всех простаивающих и недогруженных ресурсов, без создания дополнительных.Повышение безопасности при работе с информацией в публичных сетях.Стимулирование разнообразия и конкуренции программных средств: ОС, СУБД, средств автоматизации, организации поиска и представления информации.

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

- Осуществление «сквозного» поиска в гетерогенных ресурсах (прежде всего в БД, но и на веб-сайтах и др. источниках информационных ресурсов).
- Развитие таких передовых информационных технологий, как: электронная репрезентация данных без возможности их полноценного сохранения в месте получения, динамическое создание и персонализация личного информационного пространства, управление знаниями (включая генерацию нового знания).
- Демонстрация возможностей интеграции разнородных объектов культуры, стимулирование формирования стандартов метаданных объектов культуры.
- Создание и развитие множества грид-сервисов для ЭБ.

2 Общие положения

Проект представляет собой последовательность научно-технических разработок, имеющих целью обеспечить *виртуальным организациям* (ВО) *пользователей* возможность работы с *информационными объектами* (ИО) посредством множества *сервисов* (служб).

Под ВО понимается динамическое объединение пользователей, ресурсов, служб. ВО однозначно определяет политики безопасности, доступа и выполнения обязательств [5].

Важно заметить, что технология работы через ВО является принципиальной для функционирования системы безопасности в грид-сегментах. Любой пользователь должен стать участником какой либо ВО, получить сертификат доверия от сертификационного центра, который котируется владельцем ресурсов.

Сертификационный центр, являясь доверенной стороной, выдаёт и отзывает сертификаты, построенные на базе асимметричной криптографии, поддерживает репозитории для действительных сертификатов, а так же формирует списки отозванных сертификатов. Пользователь может быть участником разных ВО, в зависимости от того к каким ресурсам он предполагает получить доступ.

ИО предназначен для агрегирования контента, методанных и сюжетов (сервисов). В качестве ИО могут выступать, например, записи в базе данных электронных библиотек диссертаций, базе данных читателей библиотек, базе данных классификаторов (УДК, ББК и пр.), базе данных статей научных журналов и т.д. ИО описываются методанными, являющимися расширением формата METS [6] и представляются в виде XML-файлов. Основные требования к архитектуре ИО - это гибкость и расширяемость. Отдельно выделяются только системные метаданные, а описательные и структурные рассматриваются как часть контента.

Репозиторий поддерживает, обеспечивая хранение и использование, ИО на базе разделения (обобществления) и согласованного использования разнообразных ресурсов, в т.ч. ИО, включая объединение в коллекции, авторские подборки, резервное копирование и целый ряд других вопросов. Характерными свойствами репозитория является его распределённость и федеративность частей.

Грид-подход и применяемое программное обеспечение предусматривают очень высокую степень виртуализации ресурсов, что означает возможность создания репозитория в грид-сегменте физически на любых доступных подходящих ресурсах сегмента.

Взаимодействие ВО с репозиторием происходит посредством множества *сервисов* управления и доступа к ИО и коллекциям. Сервис (служба) - это доступный по сети компонент среды программного обеспечения среднего уровня, обеспечивающий заданную функциональность. При этом следует обсуждать лишь открытую архитектуру служб и их реализацию в открытом коде. Обязательным требованием службы является поддержка интероперабельности и виртуализации.

Службы реализуют один или несколько интерфейсов, каждый из которых определяет набор операций, активизируемых путем обмена определенной последовательностью сообщений. Служба стандартным образом описывается на некотором расширении языка WSDL. Службы могут создаваться и ликвидироваться динамически, при этом она обладает описанием внутреннего состояния в течение времени жизни.

Сюжеты ИО реализуются как ссылки на распределенные сервисы, выполняемые посредством механизмов HTTP GET/POST или связывания SOAP.

Службы характеризуются (типизируются) возможностями, которые они предлагают. Их следует разделять на системные и прикладные (высокоуровневые). Системные службы могут являться строительными блоками, которые могут быть использованы при разработке разнообразных высокоуровневых служб. Состав системных служб несколько меняется от версии к версии, функциональной стандартизации, к сожалению, пока не происходит, однако, прежде всего следует выделить:

- Службы безопасности. Авторизация, аутентификация и делегирование, в сочетании с билинговыми сервисами высокого уровня позволяют ВО проводить политику доступа к различным ресурсам, тщательно отслеживая права доступа и безопасность для большого количества пользователей.
- Службы управления заданиями. В своей основе являются реализациями грид-протокола GRAM.
- Службы управления данными, среди которых отметим сервисы организации реплик и сервисы OGSA-DAI [7]. Поддерживаются сервисом GridFTP.
- Информационные службы грид-сегмента.

Важно заметить, что на уровне системных служб доступны Common Runtime Components, такие как Java, C, Python, XIO и др.

Создание на базе системных служб удобного набора высокоуровневых сервисов, таких например как сервис доступа к репозиторию, сервис администрирования и других является одной из основных целей настоящего проекта.

3 Реализация

Реализация приведённой концепции на данном этапе предусматривала создание исследовательского прототипа.

Прототип собирался на распределённых гетерогенных ресурсах сети ЮМОС [8]. По мимо чисто телекоммуникационной поддержки ЮМОС выступает в проекте в качестве провайдера базового middleware, то есть отвечает за администрирование системных грид-служб, а так же предоставляет свой сертификационный центр (CA) и поддерживает LDAP сервера ВО.

Повышенное внимание к системе безопасности связано прежде всего с тем, что грид-сегменты не являются традиционными клиент-серверными системами, в них участники ВО могут получить полный доступ над имеющимися ресурсами сегмента и только использование инфраструктуры отрытых ключей PKI (Public Key Infrastructure), представляющей собой интегрированный набор криптографических служб и инструментов, встроенных в middleware, повышает безопасность работы с грид-среде до необходимого уровня.

Данная инфраструктура предназначена для создания и развертывания приложений, применяющих шифрование с открытым ключом (класс криптографических методов, использующих двуключевые шифры), а также для управления ими. С помощью технологии PKI пользователь генерирует пару ключей (private key и public key), сохраняет их на ключевом носителе, формирует запрос на сертификат в электронном виде и отправляет его в CA. При работе использовались сертификаты стандарта X.509.

В качестве ПО среднего уровня использовалось ПО Globus версии 3.2. В процессе работы на этом этапе версии ПО несколько раз повышались до версии 4.0, однако принципиальных сложностей с заменой не наблюдалось.

В качестве основного решения определяющего политику безопасности было использовано базовое грид-решение, основанное на использовании Community Authorization Service (CAS).

Репозиторий метаданных формата METS был реализован с использованием свободно распространяемого ПО Fedora [9]. Следует заметить, что это ПО уже достаточно хорошо себя зарекомендовало в качестве репозитория в целом ряде библиотек. Удачным примером может служить национальная библиотека Эстонии. Однако требованиям данного проекта это ПО в значительной мере не удовлетворяет, так в частности оно не представляет распределённого хранилища, тем более виртуально организованного. Все указатели на контент задаются в явном виде. В связи с этим в последующих реализациях планируется создать этот компонент среды, а так же сервис доступа к распределённому хранилищу XML методанных на базе сервисов, встроенных в грид-платформу.

В состав распределённого репозитория вошли структурированные БД МНТП «Вакцины нового поколения», коллекции диссертаций РГБ, авторские библиографические коллекции, тематически составленные из доступных по подписке РФФИ источников, таких как Elsevier, Blackwell, Kluwer и др.

Основной технологический процесс создания ЭБ был связан с наполнением репозитория метаданных. Понятную сложность представляло описание именно структурированных источников данных. По сути, для каждой коллекции приходилось создавать небольшое ПО, автоматизирующее этот процесс.

В качестве интерфейсной части использовалось ПО «Gazel», ранее разработанное в рамках проекта «Вакцины нового поколения и медицинские диагностические системы будущего». Это ПО позволяет достаточно удобно применять помимо рубрикаторов, классификаторов и словарного поиска механизмы онтологий для семантической интеграции источников информации [10].

ПО «Gazel» является достаточно интеллектуальным приложением, в том смысле, что оно поддерживает средства создания, редактирования, многоязыковой поддержки и другие возможности работы с онтологиями, а так же возможности привязки различных структурных компонентов данных, представленных в распределённых коллекциях к концептам онтологических структур [11]. Однако и сами онтологии, являясь отражением взгляда конкретного учёного или группы экспертов на часть понятийного пространства должны являться частью контента распределённой ЭБ. На последующих этапах предполагается организовать хранение частных онтологий участников ВО в репозитории ИО.

Как уже отмечалось выше, основной целью проекта является создание в грид-среде специализированных сервисов для работы с ЭБ. Для этого в использованном middleware существует три возможности:

- Использование возможностей базового грид-сервиса, предназначенного для управления данными – Grid Data Service (GDS).

Схема организации рабочего процесса данного сервиса полностью идентична схеме, принятой в Open Grid Service Architecture (OGSA) (рис. 1). Сервис создаётся соответствующей фабрикой (Factory), сведения о которой находятся в информационном регистре OGSA-DAI (1a – запрос о источнике данных, 1b – регистр возвращает заголовок фабрики, 2a – запрос к фабрике на доступ

к БД, 2b – фабрика создаёт сервис, 2c – возвращает заголовок сервиса). Сервис погружен в контейнер, управляющий его жизненным циклом. Он получает запрос на языках SQL, Xpath, Xquery и др. (3a) и передает XML-потоки данных другим сервисам (3d) и пользователям (3c). Взаимодействие с сервисом возможно из среды JAVA через набор Activity, среди которых работа с реляционными, XML и другими базами данных, действия по доставке, преобразованию и индексированию данных и другие.

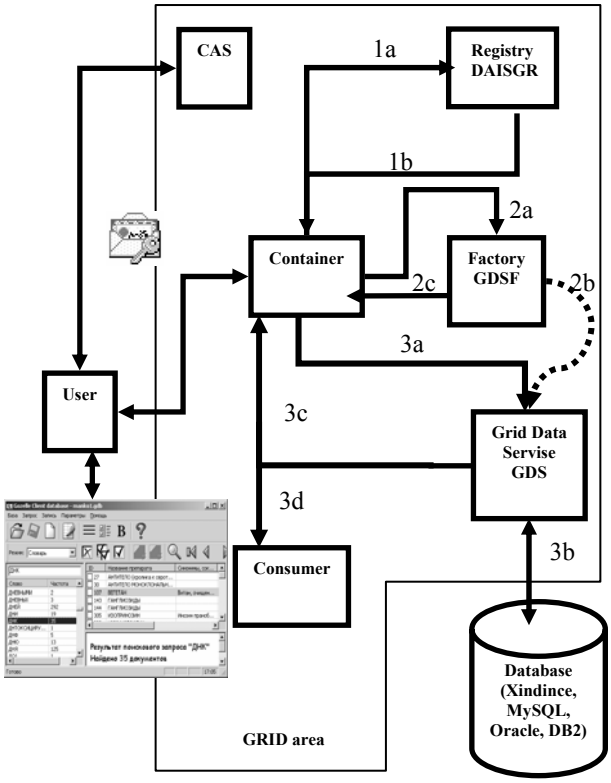

Рис.1

Следует отметить, что показанный рабочий процесса по размещению данных в виртуальном распределённом репозитории, реализованный в базовом ПО OGSA-DAI отвечает вышеприведённым требованиям доступа к информации, таким как: виртуальность, распределённость, гетерогенность, асинхронность работы, гибкость и масштабируемость.

- Наращивание возможностей GDS за счёт создания новых Activity этого сервиса.
- Создание других специализированных высокоуровневых сервисов.

Все эти возможности использовались в процессе создания прототипа, однако в конечном итоге, если не рассматривать как конечную задачу задачу расширения грид-среды, как среды программирования, для создания ЭБ, необходимую пользователям функциональность реализовывали именно специализированные высокоуровневые сервисы.

На рис.2 показан в качестве примера рабочий процесс созданного в рамках проекта нового высокоуровневого сервиса доступа к источникам полнотекстовых ресурсов, представленных в БД ведущих научных издательств (Elsevier, EBSCO и др.).

Видно, что этот сервис реализуются полностью с использованием архитектуры OGSA на базе существующего в middleware инструментария (контейнеры, сервисы безопасности мониторинга и т. п.).

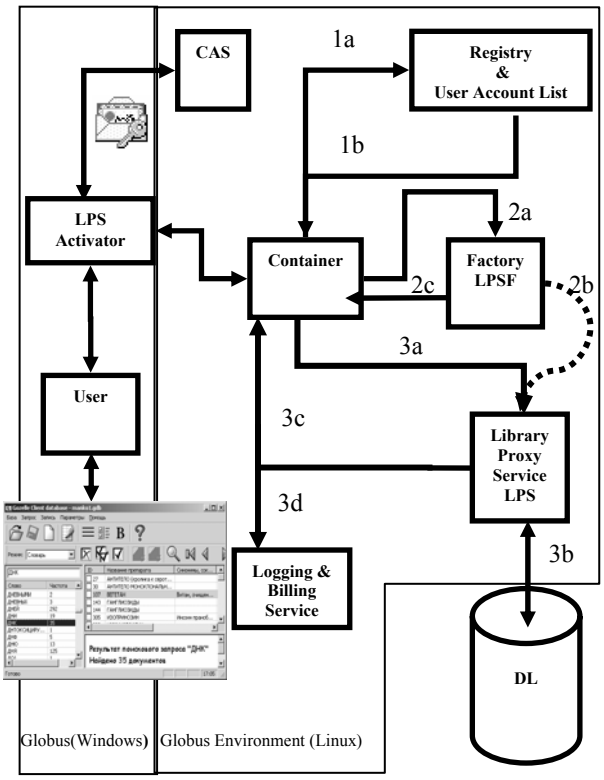

Рис.2

Данный сервис работает как на Linux, так и на Windows платформах, поэтому он имеет Windows часть, называемую активатором, которая по требованию и с учётом возможностей предъявленного сертификата доверия находит в общем регистре необходимую фабрику. Фабрика Library Proxy Service, создаёт сервис LPS, существующий в контейнере только на необходимое время жизни. Сервис формирует запрос во внешнюю ЭБ и реализует обратный информационный поток в интерфейсное ПО с использованием языков XML-HTML.

Следует заметить, что существующий сервис может порождать другие сервисы как это видно на рисунке, например делающие соответствующие записи в БД биллинга. Созданные сервисы могут быть использованы другими ВО.

В качестве языка запросов на данном этапе применялся язык SQL, а при дальнейших исследованиях планируется применять уже

используемые в грид-приложениях реализации языков Xquery и OQL. При этом некоторые высокоуровневые интеграционные сервисы, например The Grid Distributed Query Service (GDQS), поддерживающий OQL в качестве сквозного языка запросов, разрабатываются в рамках других европейских проектов [12].

4 Заключение

Предлагаемый подход, основанный на применении современных грид-технологий, позволяет уже сегодня осуществлять практические шаги по построению ЭБ, как объединения федеративных распределенных коллекций.

Среда грид, в особенности использованная OGSA архитектура являются достаточно удобной платформой развития подобных ЭБ и, прежде всего для создания ЭБ больших распределённых корпоративных проектов, в т. ч. научных, со сложившийся или очевидной структурой ВО.

Она представляет прекрасный полигон для выбора и создания информационных сервисов, в том числе работающих в БД и коллекциях, не имеющих публичного доступа.

Реализация конкретных проектов требует большой работы по формированию метоописаний данных. Однако, по всей видимости, это неизбежный процесс. ВО должна обладать возможностями и пониманием необходимости проведения такой работы.

Работы в рамках данного проекта можно рассматривать и как создание и развитие среды программирования для создания ЭБ. Действительно, все службы и компоненты middleware доступны из JAVA среды, а следовательно все вновь созданные сервисы или функциональные дополнения базовых представляют собой новые дополнительные объекты в библиотеках Runtime среды. Конечно, необходимость использования только JAVA для такого программирования можно воспринимать как достаточно жёсткое ограничение, однако оно связано только с развитием работ по проекту OGSA-DAI.

Важно заметить, что целью данного этапа не являлось продемонстрировать эффективность подхода, тем более её оценивать количественно в сравнении со скажем привычным клиент-серверным подходом к построению ЭБ. Цель заключается в том, что бы показать принципиальную возможность создания ЭБ в грид-среде. Некоторые преимущества такого подхода очевидны:

- Использование большого числа готовых решений базового middleware для построения ЭБ.
- Поддержка одного из приоритетных направлений мировой IT-политики и вовлечение в этот процесс ведущих библиотек страны.
- И, по всей видимости главное - это организация работы ЭБ в среде, ориентированной прежде всего на высокопроизводительные вычисления, среде удобной для работы научных коллективов, где ЭБ рассматриваются прежде всего как среда обработки данных [13].

Очень удачным примером, иллюстрирующим такие преимущества, может служить система распределенного хранения и анализа геномной информации [14], построенная с использованием настоящего подхода.

В соответствии с международными стандартами, принятыми в науках о жизни, некоторые особо крупные базы данных зеркалируются на региональных серверах. В первую очередь, это самые большие БД нуклеотидов (EMBL, GenBank) и аминокислот (SWISS-PROT) с еженедельным пополнением до 500 Мб сжатых данных. С 1995 г. на сервере нашего сегмента также зеркалируется ряд таких БД. Созданы средства для поиска и анализа генетической информации. Эти средства используют известный алгоритм BLAST и соответствующее ПО, доступное бесплатно для различных платформ. Однако, это ПО реализует только непосредственно поиск гомологий не принимая во внимание имеющиеся для каждой последовательности информацию, находящуюся в текстах научных статей. Разработанные нами средства оперируют и этой информацией и последовательностями, что позволяет получить максимум релевантных откликов во время поиска и одновременно значительно снизить количество документов, не относящихся реально к делу.

Другим, похожим примером может служить технология протеомных исследований, основанная на использовании новейших комплексных приборов, применяемых в биохимических и медико-биологических исследованиях, таких как, например, используемый в Институте химической физики РАН масс-спектрометр Finnigan LTQ FT, который формирует результаты структурного анализа ферментов со скоростью до нескольких мегабайт в секунду. При этом необходимо принимать во внимание, что таких приборов несколько и объем данных, получаемых за неделю будет сопоставим с сегодняшним объемом генетической базы данных SWISS-PROT (более 300 Гигабайт). Приборы такого класса стоят очень дорого и, как правило, используются совместно несколькими организациями. Анализ результатов опирается на использование ранее накопленных тематических баз данных, связь с которыми из среды обработки данных осуществляется через публичный Интернет, а сами результаты исследований на различных приборах могут комплексироваться и сопоставляться с результатами моделирования. Используемый подход не только позволяет организовывать распределённое хранение вновь полученных данных с задействованием простаивающих мощностей, их безопасность и целостность, но также предоставляет эффективный доступ к данным различным группам исследователей непосредственно из среды экспериментальных исследований. При этом управление данными не может сводиться только к

организации системы управления репликами. Анализ информации, связанной с проводимым экспериментом, полученной из репозитория ЭБ позволяет значительно сократить сценарий исследований.

Приведённые примеры показывают эффективность использования ЭБ участниками научных ВО, которые получают возможность работать с разнородными распределёнными данными непосредственно в среде, ориентированной на высокопроизводительные вычисления, увязывая информацию из ЭБ (авторских коллекций данных) с различными исследованиями, в том числе исследованиями in silico.

Участники проекта рассчитывают на то, что используемый в проекте подход будет востребован при построении корпоративных информационных систем или крупных информационных проектов.

Исследовательская компонента концепции предполагает активное сотрудничество с любыми заинтересованными субъектами и дает возможность на равных условиях участвовать в текущих и будущих международных проектах.

Литература

[1] Майстрович Т.В. Российская Национальная электронная библиотека: задачи и принципы организации// Библиотековедение.-2005.-№1 – с. 44-52.

[2] Интернет-портал по грид-технологиям. – (http://www.gridclub.ru).

[3] L. Ferreira, V. Berstis, J. Armstrong et all. Introduction to Grid Computing with Globus. IBM, 2002., а так же The Globus Alliance. (http://www.globus.org).

[4] А.В. Жучков. Проект "БиблиоГрид" и его технологические особенности. Международная конференция "Электронный век культуры". Сочи, 6-10 сентября 2004 г., [Электронный ресурс]: Материалы конф. - Электрон. дан. - М.: РГБ, 2004. - 1 электрон. опт. диск (CD-ROM).

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

[6] Metadata Encoding & Transmission Standard (METS). – (http://www.loc.gov/standards/mets/).

[7] Open Grid Services Architecture Data Access and Integration OGSA-DAI. – (http://www.ogsadai.org.uk).

[8] А.В. Жучков. ЮМОС - новые возможности старой сети /Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса [Электронный ресурс]: Материалы конф. - Электрон. дан. - М.: ГПНТБ России, 2004. - 1 электрон. опт. диск (CD-ROM). - Систем. требования: IBM PC, Windows 95 и выше. - ISBN 5-85638-091-6. - № гос. регистрации 0320400576, 1500 экз.

[9] The Flexible Extensible Digital Object and Repository Architecture (Fedora). – (http://www.fedora.info).

[10] А.В. Жучков, С.А. Арнаутов, Н.В. Твердохлебов, С.В. Голицын, И.Г. Стриж. 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, 29 сентября - 1 октября 2004 г., с.82-85.

[11] Alexei Joutchkov, Nikolay Tverdokhlebov, Irina Strizh, Sergey Arnautov, Sergey Golitsyn Grid-Based Onto-Technologies Provide an Effective Instrument for Biomedical Research. / Technology and Informatics #112/ From Grid to Healthgrid/ IOS Press, 2005, p.37-47 (ISSN 0926-9630, ISBN 1-58603-510-X)

[12] M. Nedim Alpdemir, Arijit Mukherjee, Anastasios Gounaris, Norman W. Paton, Paul Watson, Alvaro A.A. Fernandes, Desmond J. Fitzgerald. OGSA-DQP: A Service for Distributed Querying on the Grid.//E. Bertino et al. (Eds.): Springer-Verlag Berlin Heidelberg 2004, pp. 858–861 (EDBT 2004, LNCS 2992)

[13] Арнаутов С., Жучков А., Цифровые библиотеки в распределенной среде. «Открытые системы», 2001, №2, с.46-8.

[14] А. А. Черный, К. А. Трушкин, В. А. Боковой, А. К. Яновский, Н. В. Твердохлебов, А. В. Жучков, Ю. П. Лысов. Система распределенного хранения и анализа геномной информации // Молекулярная биология. 2004 г., т. 38, #1, сс. 104-109.

BIBLIOGRIG – BASIC IDEAS OF CONCEPTION AND IMPLEMENTATION

A. Zhuchkov

In this paper core ideas are considered of an approach to development of technological integrative environment that is based on federative distributed principles and that is used in the BiblioGrid project by digital libraries development.

There are also some results described considering an experimental prototype implementation.