ЕДИНАЯ СРЕДА РАСПРЕДЕЛЕННЫХ РЕСУРСОВ[1] (GRID) И ЦИФРОВЫЕ БИБЛИОТЕКИ

Жучков А.В., Арнаутов С.А.
Институт химической физики им. Н.Н.Семенова РАН (saa@ras.ru)

Приведены сведения о развитии нового подхода к созданию информационно-вычислительной инфраструктуры на основе Интернета[2]. Обсуждается роль и функции перспективных цифровых библиотек в этой среде.

Перспективы развития всемирной сети и цели создания ECP

Стремительное, лавинообразное развитие сетевых компьютерных технологий, Интернета не всегда позволяет задуматься об основных направлениях эволюции таких систем. Однако это вопрос первостепенной важности – и именно о нем задумались специалисты, объединившиеся в коммюнити GRID.

Для чего нужны сетевые компьютерные технологии? Какие старые задачи они позволяют решать более эффективно? Какие новые, ранее недоступные задачи становятся доступными? Понятно, что речь идет не о специфических задачах информатики или кибернетики – но о фундаментальных проблемах, стоящих перед нашей цивилизацией.

Главная долговременная цель развития единой среды распределенных ресурсов (ECP) - информационное обеспечение принятия решений [1]. Разработчики различают различные уровни – необходимое обеспечение требуется как для отдельных групп экспертов, занимающихся определенными проблемами, в составе отдельных кластеров понятийных сетей, так и для правительств отдельных стран или всемирных организаций (ООН, ЮНЕСКО и др.).

Термин "понятийная сеть" (knowledge network) можно определить, как совокупность информации, экспертов и знания по определенной дисциплине, которые могут быть использованы с применением компьютерных технологий для анализа различных проблем. Т.е. это люди, знания и инфраструктура. В случае научных сетей в число экспертов должны входить как профильные специалисты, так и компьютерщики, призванные организовать и обеспечить обработку информации. Инфраструктура включает ECP, информационно-ориентированный программный комплекс, модели и соответствующие приложения. Знание представлено, в том числе гипотезами (прогнозами) о возможных эффектах (например, о влиянии техногенных аэрозолей на изменение глобального климата) и интерпретацией существующих данных (например, социологических опросов)[3].

---

[1] Мы не считаем правильным вводить в русский язык еще одну кальку с английского слова, поэтому попытались предложить свой термин, отражающий суть дела. При этом мы отдаем себе отчет, что ничего “единого” создать на практике невозможно, а любое определение страдает неполнотой. Мы пытались лишь передать суть дела с помощью термина, который не вызывал бы ненужных аналогий с ранее предлагавшимися подходами.

[2] Разработчики GRID (термин не является аббревиатурой, это английское слово “сеть, решетка”) составляют отдельное коммюнити с собственной терминологией. Уже из названия можно увидеть, что они делают заявку на “второй Интернет” - т.е. их подход предельно глобален. В рамках такого подхода цифровые библиотеки рассматриваются как отдельная конкретная технология работы с информацией, которая потенциально полезна для решения некоторых частных задач в рамках значительно более общего подхода ECP. Цель данного сообщения – попытка понять и проанализировать их подходы с точки зрения коммюнити цифровых библиотек.

[3] Указанные проблемы рассматриваются в другой области, которую называют “knowledge management”. В рамках данного сообщения мы не имеем возможности более подробно останавливаться на этих вопросах.

Информационно-вычислительная инфраструктура

Работа по объединению всех ресурсов Интернета в единый интегрированный комплекс проводится в рамках создания единой среды распределенных ресурсов (ECP, GRID), которая составит информационно-вычислительную инфраструктуру будущего [1,2]. Предыстория и некоторые основные принципы изложены в работе [3].

GRID можно определить как исходно распределенную систему, которая сводит воедино данные, вычислительные мощности и ресурсы для визуализации. Основу, суть ECP составляют протоколы, сервисы, API (Application Programming Interfaces) и SDK (Software Development Kits)[4].

Для пользователя обращение к ECP будет в некотором роде аналогично использованию электроэнергии посредством обычной розетки[5] - единый интерфейс будет предоставлять доступ ко всем необходимым ресурсам так, словно мы имеем дело с одним огромным метакомпьютером. Все задачи, как традиционные для обычных компьютеров (управление процессами, памятью, файловой системой, вводом/выводом и пр.), так и принципиально новые (учет, контроль, способ доступа и распределение ресурсов, обеспечение безопасности, совместная работа над набором данных в реальном масштабе времени и пр.) будет решать специализированный комплекс программного обеспечения (например, разрабатываемый в проекте Globus [4,5]) на базе соответствующей аппаратной инфраструктуры.

Перечислим некоторые принципиальные задачи, которые необходимо решить при создании ECP:

- обеспечение интероперабельности в глобальной программно-аппаратной инфраструктуре ECP;
- диспетчеризация, включая идентификацию доступных ресурсов, статистика использования и загрузки ресурсов и пр.;
- система безопасности и контроля доступа, в т.ч. гибкое регулирование объема прав и привилегий пользователей;
- обращение к наборам данных в удаленных архивах (включая протоколы, которые необходимо использовать для работы с гетерогенными источниками данных) и др.

Эволюция информации в вычислительных сетях

Стремительное развитие глобальных информационных и вычислительных сетей ведет к изменению фундаментальных парадигм обработки данных. Некоторые тенденции проиллюстрированы в таблице 1.

Таблица 1

Локальные системы	Распределенные системы	ECP
Системы хранения данных	Распределенная обработка данных	Системы анализа данных
Обработка больших объемов данных непосредственно в хранилищах	Среда для "открытия" информации (СПИ[6])	Гибкая информационная политика на основе "обратной связи"
Цифровые библиотеки	Вычисления с интенсивным использованием информации (information-based computing)	"Понятийные сети" (knowledge networking)

Кратко эти изменения можно охарактеризовать как переход к распределенным ресурсам и создание инфраструктуры для свободного доступа к любым ресурсам сети.

[4] Найти строгое формальное определение ECP в литературе нам не удалось. Предложенное краткое описание отражает, на наш взгляд, суть и наиболее существенные черты.

[5] Во всяком случае, именно на это претендуют разработчики ECP – ведь термином “power grid” обозначаются обычные электрические сети.

[6] Термин, используемый в публикациях сообщества ECP, очень напоминает широко известный термин “информационно-поисковая система” - однако это отчетливо другой термин. Мы попытались перевести английский термин “information discovery system” как “система поиска информации” (СПИ) - осознавая при этом неполноту и неточность такого перевода.

В приводимой ниже таблице представлена эволюция методов обработки данных.

Таблица 2

Опция	Эволюция опции		
Способ именования (организации) данных	Файловая система (UNIX)	LDAP (Lightweight Directory Access Protocol)	Каталог (база) метаданных
Способ хранения данных	Локальный жесткий диск	Архивное хранение	Интегрированные базы данных и архивы
Доступ к данным	Вручную	Интегрированные архивы на базе файловых систем	Унифицированный доступ к файловым системам, архивам и базам данных
Использование данных	Локальные приложения	Распределенные объекты	Понятийные сети
Публикация данных	Репозитории	Цифровые библиотеки	Федеративные информационные репозитории (федеративные цифровые библиотеки)
Представление данных	Визуализация в конкретном приложении	Системы управляемых пользователем потоков данных (user-managed data flow systems)	Координированная презентация (coordinated presentation), Java Beans

Программное обеспечение

Базовое программное обеспечение, необходимое для создания инфраструктуры для обработки больших объемов информации включает в себя:

- создание в рамках ECP согласованной объектной среды, позволяющей поддерживать выполнение необходимых приложений и сервисов в реальном масштабе времени;
- создание СПИ, обеспечивающих поиск данных по атрибутам, выделение метаданных (metadata mining), семантическую интероперабельность, общие онтологии и улучшенные системы аннотирования данных;
- разработку технологий цифровых библиотек, поддерживающих публикацию, каталогизирование и хранение наборов данных;
- создание систем управления данными, снабженных каталогами системных метаданных для обеспечения интероперабельности между объектами и ресурсами в ECP;
- создание механизма т.н. "брокера ресурсов" (storage resource broker)[7], обеспечивающего единый механизм доступа к гетерогенным источникам данных;
- создание комплексов "база данных+хранилище", как основу для специализированных коллекций;
- создание архивных систем для долговременного хранения данных.

Речь идет фактически о разработке программного обеспечения промежуточного уровня (среднего, middleware) с тем, чтобы спрятать сложность взаимодействия и интеграции различных распределенных гетерогенных ресурсов, сохраняя возможности. При этом весьма важно, чтобы “прозрачными” оставались три подсистемы [6]:

- подсистема именования – поскольку уникальные имена практически невозможны, используются атрибуты (метаданные). СПИ, выполняя запросы по этим атрибутам, осуществляет поиск в каталоге метаданных (information discovery catalog).

[7] Разработчики ECP сознательно используют термин “брокер”, заимствуя его из технологии CORBA (более широко – сообщества OMG). При этом они подчеркивают, что ECP решает фактически те же задачи, что и CORBA – но на более высоком уровне. В состав интегрируемых ресурсов могут при этом входить и ресурсы, созданные с использованием этой технологии.

- подсистема нахождения, расположения (location) ресурса. Расположение конкретного ресурса может описываться в виде UNIX-атрибутов и храниться в упомянутом каталоге. При этом становится возможным разнесенное хранение данных и метаданных. Очевидно, что в каталоге должны храниться и данные о протоколе для доступа к конкретному набору данных.
- подсистема (конверсии) протоколов, которая должна осуществлять преобразование протоколов с целью обеспечения интероперабельности. Преобразование может осуществляться с помощью специальных серверов, установленных в каждом репозитории, что позволяет автоматизировать эти процессы.

Интеграция ECP, технологий обработки больших объемов информации и технологий цифровых библиотек в перспективе приведет к появлению уникальной инфраструктуры для доступа, анализа и генерации информации.

Цифровые библиотеки в ECP

Разработчики ECP отводят цифровым библиотекам (ЦБ), по нашему впечатлению, достаточно скромное место. Они рассматривают технологии ЦБ как достаточно частные инструменты, эффективные для решения ряда задач хранения, обработки, поиска и генерации информации. Ниже изложены именно их взгляды, причем они могут выглядеть достаточно нетрадиционно с точки зрения членов коммюнити ЦБ (в т.ч. и с точки зрения терминологии).

Место ЦБ в ECP и их структура

ЦБ начинались в значительной степени как локальное приложение. В распределенных системах (в качестве критерия которых можно рассматривать разнородность ресурсов) говорят о федеративных цифровых библиотеках [7], а в ECP предполагается создание "понятийных сетей". Подчеркнем, что это именно эволюция, каждая новая стадия которой включает в себя предыдущую.

В структуре понятийной сети ЦБ можно рассматривать, как полностью автоматизированные комплексы, предназначенные для "аннотированного хранения" данных, причем пользователем ЦБ может быть как человек, так и приложение на удаленном компьютере.

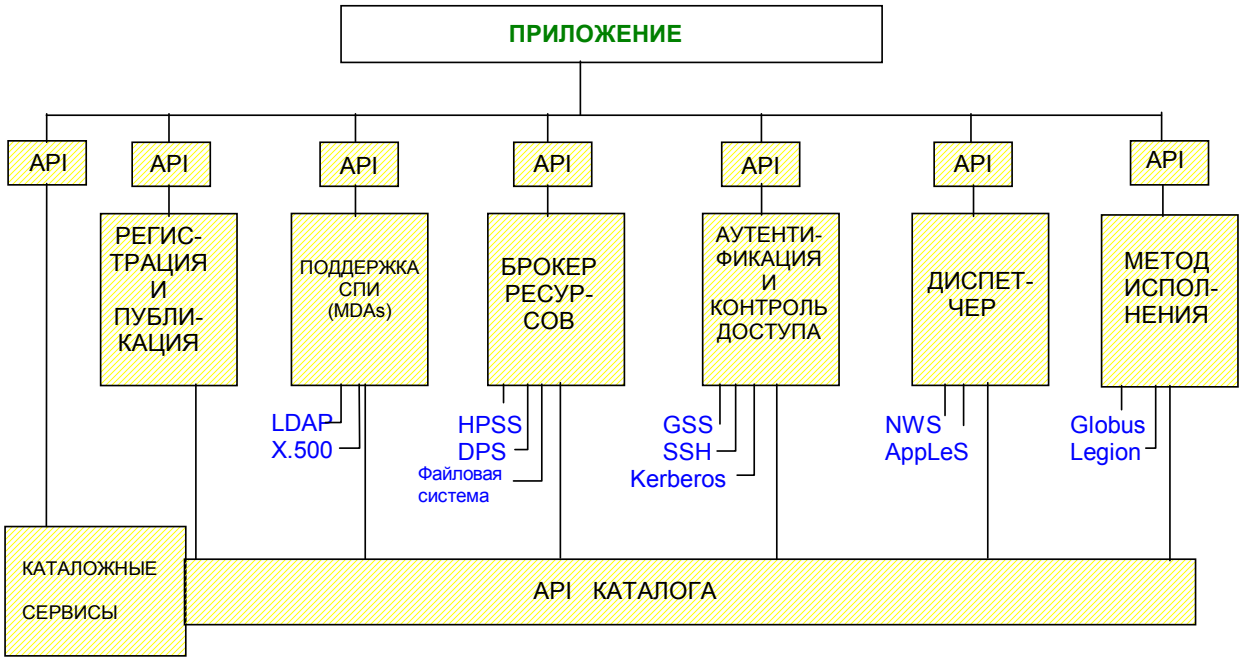
Очевидно, тут уместно будет сказать о том, что сама по себе информация, хранящаяся в репозиториях (архивах) является только лишь набором битов, комбинацией данных и метаданных, сгенерированных с использованием адекватного языка разметки. Как именно конкретный пользователь (приложение) будет использовать эту информацию, определяется пользователем. В этом смысле очень точным представляется термин “информационный контекст”, предложенный в [6]. Под ним понимаются возможные точки зрения, аспекты рассмотрения данного оригинального набора данных.

ЦБ предоставляют достаточно мощный и совершенный набор инструментов (сервисов) для манипуляции с наборами данных и включают:

- публикацию/регистрацию новых наборов данных;
- база метаданных для поиска данных по атрибутам;
- доступ к гетерогенным ресурсам посредством брокера ресурсов;
- контроль аутентификации и доступа;
- диспетчирование вычислительных ресурсов и ресурсов ввода/вывода;
- распределенное исполнение сервисов ЦБ.

Соответствующие сервисы могут регистрироваться в специальной базе данных и вызываться для обработки любого набора данных, хранящегося в библиотеке (ее репозитории). Комбинация возможности доступа к данным через базу метаданных, каталожных сервисов, зарегистрированного набора методов

обработки данных позволяет решать практически все задачи для создания среды обработки данных в ECP. На рисунке приведена возможная архитектура ЦБ [1].



Из приведенной схемы (а также таблицы 2) следует, что сообщество ECP отделяет задачи публикации, хранения и "открытия" информации от собственно ЦБ. Фактически, речь идет о "модульной системе", когда все необходимые блоки функционируют отдельно и федеративно: когда СПИ существует отдельно, а архив данных – отдельно и т.д. В этой связи возникает вопрос о соответствии такой модели - модели ЦБ. У нас вообще сложилось впечатление, что коммюнити ECP не выделяет ЦБ в самостоятельную сущность.

Это принципиальный вопрос, над котором необходимо думать и на который надо дать убедительный ответ. Нам представляется, что атрибуты ЦБ, позволяющие говорить об особом качестве и отдельной области исследований, следующие:

- многовековая научная и культурная традиция библиотек, как хранилищ знаний человечества;
- функция публикации новых знаний. Здесь мы очевидным образом "посягаем" на область, которая традиционно существовала отдельно от библиотек – издательское дело. Нам представляется, что развитие новых информационных технологий ставит на повестку дня объединение этих областей[8];
- наличие развитой системы генерации метаданных (каталогизации, аннотирования, и пр.).

Кроме того, "модульная система" фактически реализуется в программном обеспечении промежуточного уровня, в то время как пользователь (приложение) общается с метакомпьютером через единый интерфейс. ЦБ же предполагает концептуальное и логическое единство, поддерживаемое онтологией, схемой, семантикой, административной структурой и пр., позволяющее эффективно интегрировать разнородные информационные ресурсы.

Еще раз подчеркнем, что проблема ЦБ существенно по-разному рассматривается в сообществах исследователей, развивающих ECP и ЦБ. Можно, как нам представляется, с сожалением констатировать, что связи между ними недостаточно тесные и содержательные. В тоже время такая координация и взаимодополняемость существуют. Например, сообщество ECP исследует проблемы создания

[8] Этому вопросу посвящено сообщение "Роль и место виртуальных цифровых библиотек в Интернете" (см. сборник докладов данной конференции).

инфраструктуры обработки данных, технологии связывания различных коллекций (наборов данных). А сообщество разработчиков ЦБ исследует проблемы метаданных, разрабатывает технологии поиска конкретных цифровых объектов в репозиториях ЦБ, включая метаданные и “помощники поиска” (finding aids).

В настоящее время в России коллаборацией НИИЯФ МГУ , ИТЭФ , ИПМ и АНО ТЦ “Наука и общество” на базе Южной Московской Опорной Сети создан российский сегмент ECP (RGRID). Этот сегмент зарегистрирован в Европейской Комиссии в рамках проекта EU Data GRID. Осенью этого года физики, биологи и химики, использующие технологии ECP в своей работе, планируют провести учредительную конференцию Ассоциации пользователей. Цель создания Ассоциации – разработка конкретных проблем создания понятийных сетей на базе ECP, выработка стандартов, создание и поддержание информационных ресурсов.

Литература

1. The Grid: Blueprint for a New Computing Infrastructure. Ed. by I.Foster and C.Kesselman. Morgan Kaufmann Pub., San Francisco, CA. 1999.
2. В.Коваленко, Д.Корягин Вычислительная инфраструктура будущего. Отрытые системы, 1999, N11-12. http://www.osp.ru/os/1999/11-12/045.htm
3. I.Foster. Internet Computing and the Emerging Grid. Nature, December, 7, 2000.
4. I.Foster, C.Kesselman. Globus: A metacomputing infrastructure toolkit. Int. J. Supercomput. Appl., 1997. http://www.globus.org
5. I.Foster, C.Kesselman. The Globus project: A progress report. In Proc. Heterogneous Computing Workshop, p.4-18. Los Alamos, CA: IEEE Computer Society Press, 1998.
6. A.Rajasekar, R.Marciano, R.W.Moore “Collection Based Persistent Achives”. Proc. Of the 16 IEEE Simposium on MassStorag Systems, 1999.
7. W.Arms. Digital Library. The MIT Press. 2000

The Uniform Environment Of The Distributed Resources (Grid)

And Digital Libraries

Alexey V.Zhuchkov, Sergei A.Arnautov

N.Semenov Institute of Chemical Physics RAS

ul.Kosygina, 4, 119991 Moscow, Russia *(saa@ras.ru)*

Information about development of the new approach in creation of an information infrastructure (named as GRID in English) on the basis of the Internet is given. The role and functions of perspective digital libraries in this environment is discussed.