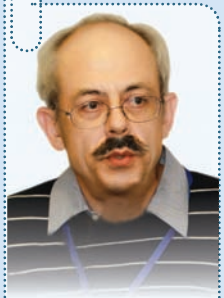
ЭЛЕКТРОННАЯ БИБЛИОТЕКА КАК ОСНОВА ВИРТУАЛЬНОЙ ИССЛЕДОВАТЕЛЬСКОЙ ИНФРАСТРУКТУРЫ

Когда-то давно было популярно такое определение университета: университет это библиотека и стоящие вокруг него здания. В прежние времена все знания действительно сосредотачивались в библиотеке, и вне её стен невозможно было заниматься наукой. Позднее их значение уменьшалось, особенно с внедрением компьютерной техники, и они стали снова осмысливать своё положение в новом мире. История развивается по спирали, и, возможно, сейчас библиотека (уже электронная) снова может начать играть основополагающую роль в научных исследованиях, как основа виртуальной исследовательской инфраструктуры (ВИИ)[1].

Авторы



Михаил Ефремович ШВАРЦМАН
начальник отдела исследования компьютерных систем Российской государственной библиотеки

Олег Павлович НАЙДИН
ведущий программист отдела исследования компьютерных систем Российской государственной библиотеки

Университет – это друзья и библиотека!
М. Ерёмин

ВИРТУАЛЬНАЯ ИССЛЕДОВАТЕЛЬСКАЯ ИНФРАСТРУКТУРА. ЧТО ЭТО И ДЛЯ КОГО?

Если мы посмотрим на темы докладов последних конференций по электронным библиотекам, то увидим, что сейчас практически никто не говорит о том, как создать электронную библиотеку. В основном, доклады посвящены интеграции электронных библиотек, и не только в межбиблиотечные, но и в образовательные, социальные и научные проекты. Этому способствует развитие web-сервисов, распространение технологии Linked Open Data, всё возрастающие возможности семантической обработки текста. Поскольку именно в библиотеках хранятся тексты, которые можно семантически обрабатывать, в библиотеках сосредоточены специалисты по обработке информации. Именно поэтому в последнее время стали появляться проекты исследовательских сообществ на базе университетских библиотек. Эти сообщества занимаются разными отраслями наук. Например, в Лейденском университете это сообщество социологов и экономистов, в проектах JISC (Единого комитета по информационным системам Великобритании) — биологи и филологи.

Хотя эта тема довольно активно обсуждается в последнее время, единого термина ещё так и не было выработано. Кроме Virtual-Research Environment (VRE) также употребляются термины «совместные э-исследовательские сообщества», «совместная виртуальная среда», collaboratory, «научный шлюз», «виртуальная организация», «виртуальное исследовательское сообщество».

Что же представляет собой эта ВИИ? Начнём с определения. *ВИИ — это платформа для интернет-ориентированной совместной рабочей среды, позволяющей использовать новые методы взаимодействия и новые способы обработки исследовательских данных и информации.* Определение это не единственное,

1. Работа выполнена в рамках гранта РФФИ 11-07-00750-а.

так же как и для электронной библиотеки, но в разных вариациях упоминаются возможности распределённой работы, разделения инструментов, опыта участников и информации между членами сообщества. Авторы попытались представить себе, какие задачи могла бы решать ВИИ по истории российской науки на примере электронной библиотеки трудов великих российских учёных XVIII – XIX вв. и какие бы сервисы для этого потребовались.

СОЗДАЁМ ВИИ

Обычно для создания ВИИ выбирается какое-либо портальное программное обеспечение. Чаще всего для этого используется **The VRE Toolkits для MS SharePoint** (рис. 1).

Этот инструмент используется для авторизации, ведения профилей, календарей и, что самое интересное, подключения различных web-сервисов, которые собственно и помогают организовать совместную научную деятельность. Посмотрим, какие сервисы нам потребуются.

Вначале нужно определить дисковое пространство, в котором будут храниться промежуточные результаты совместной работы участников ВИИ. Для этого можно использовать сервисы хранения данных:

- **Amazon Web Services** – отличается гибкой политикой тарификации;
- **Dropbox** – имеет хорошо сделанный API;
- **Google Drive** – умеет интегрироваться с другими сервисами Google;
- **YandexDisk** – предоставляет 10 ГБ бесплатного хранения.

У каждого сервиса есть свои достоинства и недостатки, а выбор будет зависеть от задач.

Поскольку мы работаем в команде, то нам необходим сервис управления совместной работой. Помимо широко известных календарей «Яндекса» и Google можно обратить внимание на сервис **TRELLO** (www.trello.com). В нём присутствует доска объявлений, которая предоставляет немало возможностей для организации совместной работы над каким-либо проектом. Есть возможность объединять заметки в группы, помечать их цветными метками и прикреплять файлы, изображения, списки дел и ссылки. Также есть возможность объединять пользователей в группы и добавлять людей к заметке, чтобы они её просмотрели. Помимо web-интерфейса TRELLO имеет API (рис. 2). Для примера мы попробовали использовать этот сервис для планирования работ по созданию нашей библиотеки.

Определив задачи, нам нужно выбрать сервисы для совместной работы над документами, которые будут создаваться в процессе нашей совместной работы. Трудно соперничать с сервисами Google по совместному редактированию текста таблиц и презентаций, поэтому об этом можно не говорить подробно. Хотим только обратить внимание на **Google Fusion Tables** (www.google.com/drive/start/apps.html#fusiontables) и **Googlecharttools** (www.google-developers.appspot.com/chart). Эти сервисы позволяют создавать таблицы из различных наборов данных. В них есть гибкие инструменты для объединения таблиц, форматирования их содержимого, анализа данных и их визуализации с помощью построения диаграмм различных видов. Это бывает очень удобно при групповой работе – при заполнении (изменении) таблицы одним из участников работы все остальные сразу видят изменённую картинку. Так, например, заполняя таблицу в docs.google.com/spreadsheet об иерархической структуре Российской Академии наук, все сразу видят получающуюся иерархию (рис. 3). К тому же можно использовать систему геопозиционирования для изображения табличных данных.

Для представления табличных данных на карте также можно использовать сервис **ArcGIS** (www.esri. ▶

Рис. 1. Скриншот The VRE Toolkits

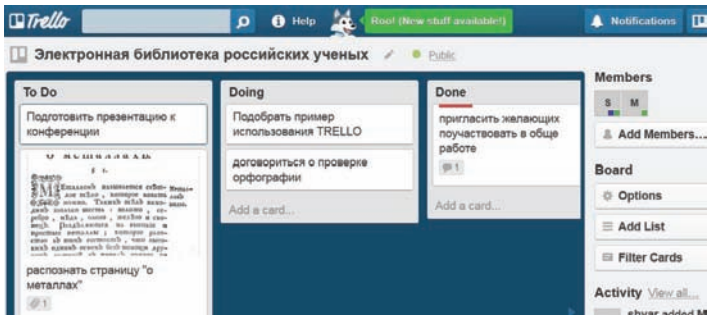


Рис. 2. Скриншот сервиса TRELLO

Рис. 3. Метод построения иерархической структуры в онлайне

com/software/arcgis). Это система для организации, анализа и отображения любых статистических данных, имеющих географическую привязку. Помимо инструментов анализа и визуализации данных приложение включает в себя средства интеграции с сайтами. К сожалению, в нём нет поддержки русского языка. Для взаимодействия предоставляется web-интерфейс, есть API для Java и Python. Для примера мы попробовали представить на карте, как распределяются места рождения и смерти авторов книг, включённых в нашу электронную библиотеку. Для этого была составлена таблица случайным образом отобранных авторов (рис. 4). Общая картина нас не удивила. Как и сейчас, места рождения (зелёные кружки) разбросаны по всей России, а места смерти (красные кружки) находятся в основном в Москве, Санкт-Петербурге, на Чёрном море и в Западной Европе.

ДОСТУП К СТОРОННИМ ОНЛАЙН-БАЗАМ ДАННЫХ

Большим преимуществом современных научных библиотек является их хорошая обеспеченность доступом к онлайновым научным базам данных, на которые они тратят сейчас существенную часть бюджета. Поэтому в большой степени основой для ВИИ могут являться библиотека и библиотекари. Именно здесь можно получить экспертную оценку большинства источников и сформировать массив научной информации, являющийся базой для любого научного исследования. Первоначальный массив может быть собран при помощи библиографических менеджеров типа **Zotero, Mendeley, Citavi**. Наибольшей функциональностью обладает Citavi. В него можно загрузить данные по протоколу Z39.50 из крупнейших библиотек мира, проведя предварительный поиск по элементам библиографического описания. У программы много достоинств (можно оставлять комментарии, просматривать оригиналы, создавать структуру будущей научной работы и т.п.), однако неясно, как её можно подключить к порталу, поэтому можно посоветовать использовать сервис Mendeley. Для сервиса разработан Web-Importer, который позволяет добавлять записи, найденные в базах крупных информационных провайдеров. Загруженные записи можно редактировать, снабжать метками, связывать с полными текстами. Для сервиса разработан API, что позволяет использовать его в портале ВИИ. На рис. 5 приведён пример загрузки записей из Springer и Elsevier по теме «history of Russian science».

Любой группе учёных будет очень полезно получать информацию о новых поступлениях в онлайновые научные базы. Для этого нужно установить в портале своего научного коллектива блок для получения RSS[2] и настроить его соответствующим образом. Многие крупные издатели и агрегаторы научной информации предоставляют сервис отправки по RSS списков новых поступлений по заданной теме. Например, на сайте научной библиотеки СПбГУ можно найти 20 последних публикаций, включённых в БД Scopus (www.library.spbu.ru/news/sps). Если тема, интересующая коллектив учёных, освещается на различных сайтах, то можно при помощи сервиса **Yahoo! Pipes** (pipes.yahoo.com/pipes) сначала собрать полезные новости со всех возмож-

Рис. 4. Распределение мест рождения и смерти

Рис. 5. Пример использования Mendeley

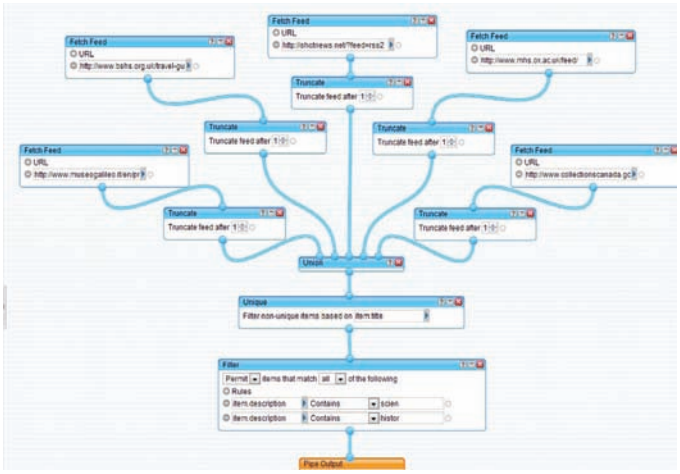


Рис. 6. Скриншот настройки контейнера

2. RSS – семейство XML-форматов, предназначенных для описания лент новостей, анонсов статей, изменений в блогах и т.п. Информация из различных источников, представленная в формате RSS, может быть собрана, обработана и представлена пользователю в удобном для него виде специальными программами-агрегаторами, или онлайн-сервисами.

ных поставщиков и затем отобрать по заданным ключевым словам действительно интересную информацию. Yahoo! Pipes — это бесплатный сервис для создания конвейеров обработки информации. Конвейер принимает на входе одну или несколько RSS-лент и обрабатывает их в соответствии с блоками в конвейере (множество основных блоков уже реализовано).

На выходе мы можем получить ещё одну отфильтрованную RSS-ленту, JSON или KML. Есть возможность сделать общедоступными результаты работы конвейнера. В качестве примера можно посмотреть, как выглядит конвейнер для сбора новостей с ряда сайтов, посвящённых истории науки, и фильтрации собранных результатов по терминам histor и scien. Скриншот настройки контейнера показан на рис. 6. Результат работы этого контейнера в виде RSS доступен по ссылке www.pipes.yahoo.com/pipes/pipe.run?_id=6d82c07ce0bd77abc5dfae5eb6cb4193&_render=rss.

ОБРАБОТКА И АНАЛИЗ СОБСТВЕННОГО КОНТЕНТА

Помимо информации из сторонних источников, нам нужно будет обрабатывать информацию из собственной электронной библиотеки. Учитывая, что у нас находятся отсканированные издания XVIII—XIX вв., задача их качественного распознавания является очень важной, тяжёлой и требующей коллективной работы. Пока сервисов распознавания текстов немного, а бесплатных практически нет. Нам удалось найти только **Virtual Transcription Laboratory** (www.wlt.synat.pcss.pl/wlt-web/index.xhtml). Цель этого проекта — предоставить web-сервис для загрузки изображений, распознавания их при помощи программного обеспечения **Tesseract** и коллективного редактирования результатов распознавания. Пример работы представлен на рис. 7. К сожалению, сервис пока не предоставляет API, но авторы говорят, что работают над этим.

После того как мы получили распознанные тексты исследуемых документов, нам уже можно приступать к их анализу. Какие же сервисы анализа текстов существуют сейчас? Интересный пример такого сервиса предлагает команда разработчиков Google **Ngram Viewer Team**, подразделение Google Research. Весь массив книг, оцифрованных и распознанных Google, был разбит на так называемые *n*-граммы, т.е. словосочетания — сочетания по *n* слов, идущих в тексте подряд (*n* может быть от 1 до 5). Для каждой *n*-граммы сочетания записывается, в какие документы она входит и какова дата публикации каждого из этих документов. Пользователь сервиса задаёт поисковый запрос, в котором указывает, какие *n*-граммы его интересуют, и интервал времени, а сервис выдаёт график, на котором показано отношение количества запрошенных *n*-грамм к общему количеству *n*-грамм с таким же значением *n* в общем массиве Googlebooks за каждый год из указанного отрезка

Рис. 7. Интерфейс Virtual Transcription Laboratory

Рис. 8. Google Ngram Viewer. Сравнение популярности Бутлерова, Менделеева и Ломоносова

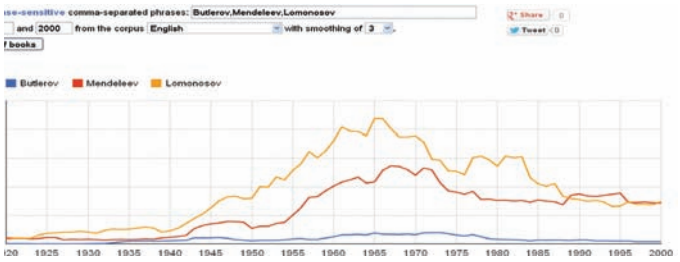


Рис. 9. Google Ngram Viewer. Сравнение Butlerov, Mendeleev и Lomonosov

Рис. 10. Результаты работы сервиса Annie

времени. Например, нам интересно сравнить, насколько были популярны А.М. Бутлеров, Д.И. Менделеев и М.В. Ломоносов в России и за рубежом. Задав поиск на русском языке в массиве российских книг, мы видим, что Ломоносов всегда был самым популярным, а популярность Бутлерова и Менделеева была примерно равна до 1972 г., а потом популярность Менделеева стала выше (рис. 8 и 9).

Если же мы посмотрим на результат поиска на английском языке среди массива книг на английском, то увидим, что популярность Ломоносова была выше только до 1985 г., а потом стала ниже или сравнима с популярностью Менделеева. А популярность Бутлерова всегда была существенно ниже остальных. Авторы ни в коей мере не претендуют на интерпретацию этих фактов, но уверены, что многим историкам такой инструмент может быть очень интересен.

Для семантического анализа текста можно использовать сервис **Annie** (www.aktors.org/technologies/annie). Это приложение с открытым исходным кодом для анализа текстов, основанное на GATE. Основной его задачей является распознавание объектов в тексте. Annie умеет находить людей, месторасположения, организации, даты, адреса, деньги и проценты. В данный момент, кроме английского, поддерживаются болгарский, румынский, бенгальский, греческий, испанский, шведский, немецкий, итальянский и французский языки и происходит адаптация к арабскому, китайскому и русскому. Для работы с приложением существуют web-интерфейс и API. На вход приложению подаётся анализируемый текст или адрес страницы в Интернете. Результатом является размеченный текст. На примере (рис. 10) мы видим выделенные разными цветами имена, даты, названия мест.

Несколько большим функционалом обладает сервис семантического анализа текста **OpenCalais** (www.opencalais.com). Это приложение для анализа текстов любой сложности: оно извлекает из текста информацию о различных объектах (используя технологии обработки естественных языков и машинного обучения) и создаёт из них теги. Также приложение умеет находить конструкции, отражающие некоторые события или взаимосвязи между объектами. Приложение может быть использовано для построения облаков тегов для различных материалов портала, а также для формирования записей в формате RDF. Приложение имеет как web-интерфейс (поддерживает только английский язык), так API (поддерживает английский, испанский и французский языки). Для иллюстрации возможностей сервиса мы взяли текст из англоязычной «Википедии» о пребывании М.В. Ломоносова в Марбурге (рис. 11). На рисунке мы видим, что организации, имена, события и связи между ними выделены, определены семейные связи. Текст преобразован в RDF, снабжён тегами и может быть подвергнут дальнейшей семантической обработке.

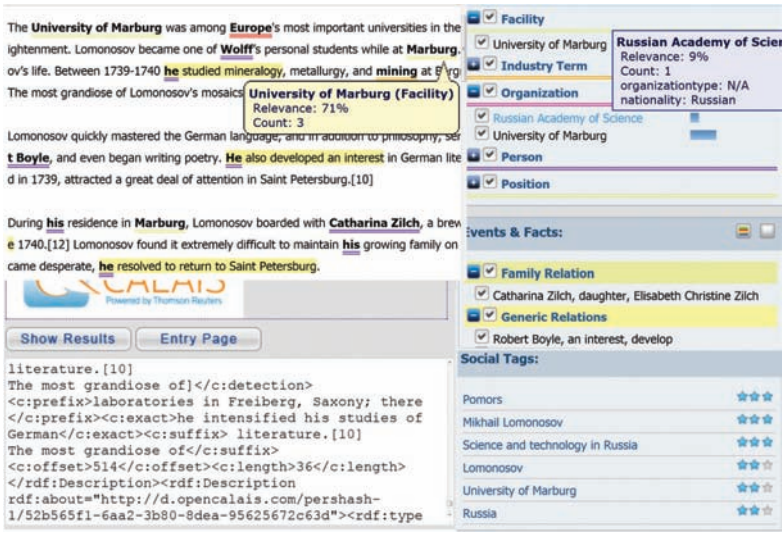


Рис. 11. Пример работы с OpenCalais

ПУБЛИКАЦИИ В ОТКРЫТОМ ДОСТУПЕ

Одна из задач исследователя в современном мире — опубликовать результаты своей работы так, чтобы они могли быть максимально удобно использованы другими учёными. Поэтому нашу электронную библиотеку трудов российских учёных, обогащённую исследователями ВИИ, хорошо бы опубликовать как Linked Open Data. Одним из основных преимуществ современных технологий является возможность исследователей связывать полученные ими результаты с результатами других коллективов и предоставлять свои данные для использования всем желающим. Для этого потребуется технология **Linked Open Data** (www.linkeddata.org). Таким образом публикуют свои данные многие зарубежные библиотеки. Так, например, OCLC, British Library, Bibliotheque nationale de France, Europeana и многие другие уже опубликовали свои каталоги как Linked Open Data. Недавно было объявлено, что и **Virtual International Authority File (VIAF)** позволяет использовать свои данные по этой же технологии. Это особенно важно для нас, поскольку позволяет устанавливать связи от нашей электронной библиотеки к нормативным записям о великих русских учёных, получать сведения о написании их фамилий на разных языках и, соответственно, полный список их трудов, изданных на этих языках. Поскольку в задачу авторов не входит детальное описание этой технологии, то напомним только, что основными особенностями таких данных является то, что они представлены в виде триплетов «субъект — предикат — объект», каждая часть триплета — это URL, ведущий к месту публикации этой части, и весь массив находится в открытом доступе. Для поиска в таких массивах разработаны специальные инструменты.

Наша база российских учёных XVIII—XIX вв. пока ещё не имеет возможности публикации своих данных в виде Linked Open Data. Надеемся, в будущем мы это сделаем с помощью **DR2 Server** (www.d2rq.org/d2r-server).

Это бесплатный сервлет[3] для Apache Tomcat, являющийся прослойкой между реляционными базами данных и семантической паутиной. Для демонстрации полезности Linked Open Data в деле изучения истории российской науки мы выбрали данные, которые неизвестные исследователи уже опубликовали как Linked Open Data в DBpedia (www.wiki.dbpedia.org). Мы решили выбрать всех российских учёных, информация о которых занесена в DBpedia, и построить схему распределения этих учёных по наукам. Для этого при помощи **SPARQL Explorer** (www.dbpedia.org/sparql) на языке SPARQL был составлен следующий запрос:

```
SELECT *WHERE {
?name <http://www.w3.org/1999/02/22-rdf-syntax-
ns#type><http://dbpedia.org/ontology/Scientist>.
?name <http://dbpedia.org/property/birthPlace>
<http://dbpedia.org/resource/Russia>.
?name <http://dbpedia.org/ontology/birthDate> ?date.
?name <http://dbpedia.org/ontology/field>?field}
```

В ответ мы получили набор триплетов следующего вида. Для экономии места предикат опущен, поскольку он везде одинаковый — «работает в области». Выдержка из списка приведена на рис. 12.

name	field
http://dbpedia.org/resource/Vladimir_Smirnov_(mathematician)	http://dbpedia.org/resource/Mathematics
http://dbpedia.org/resource/Ilya_Ivanovich_Ivanov	http://dbpedia.org/resource/Biology
http://dbpedia.org/resource/Oleg_Losev	http://dbpedia.org/resource/Electrical_engineering
http://dbpedia.org/resource/Oleg_Losev	http://dbpedia.org/resource/Physics
http://dbpedia.org/resource/Anatoly_Larkin	http://dbpedia.org/resource/Physics
http://dbpedia.org/resource/Nikolai_Semenovich_Kurnakov	http://dbpedia.org/resource/Chemist

Рис. 12. Набор триплетов

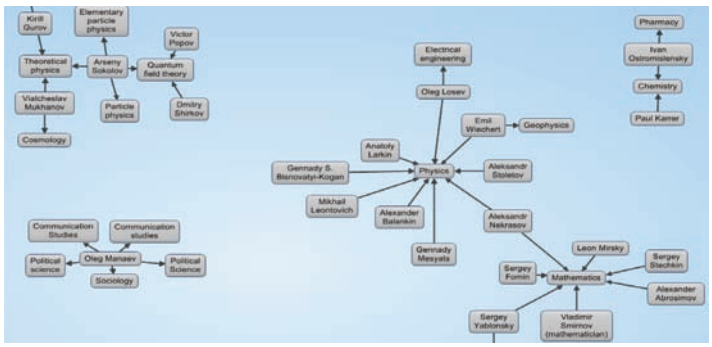


Рис. 13. Схема распределения учёных по наукам в Gruff

На основе этой таблицы мы можем при помощи сервиса **Gruff** (www.franz.com/agraph/gruff) построить нужный нам граф (рис. 13). Gruff — это средство для представления триплетов в различных видах. Он позволяет отображать запросы в виде графа, таблицы, SPARQL-запросов и в других формах. Приложение легко в использовании и доступно для всех платформ через API.

Таким образом, посредством запросов на языке SPARQL и дополнительных сервисов типа Gruff мы можем на портале нашей ВИИ объединять наборы данных из разных источников, добавлять к ним данные своих исследований и представлять в наиболее удобном для нас виде. Мы не хотели бы преувеличивать значение DBpedia как источника научной информации, однако как инструмент получения различных точек зрения и отладки технологии Linked Open Data, DBpedia работает замечательно.

Как показывает опыт, основные проблемы виртуальных исследовательских сообществ связаны не с техническими трудностями, а с неготовностью учёных к совместной работе в онлайне. Можно разрабатывать сколь угодно удобные сервисы, но если мы не сможем заинтересовать ими учёных, стоить они будут немного. Все успешные виртуальные научные сообщества отличаются наличием в них очень активных учёных и дружным коллективом специалистов разного профиля. Различные фонды поддержки науки готовы вкладывать деньги в такие коллективы, потому что они понимают, что ВИИ имеют высокий потенциал и могут предложить такие преимущества, как:

«Основные проблемы виртуальных исследовательских сообществ связаны не с техническими трудностями, а с неготовностью учёных к совместной работе в онлайне. Все успешные виртуальные научные сообщества отличаются наличием в них очень активных учёных и дружным коллективом специалистов разного профиля»

- поддержку географически разнесённых исследовательских групп;
- упрощение международной совместной работы;
- поддержку общей сети;
- поддержку в междисциплинарных исследованиях;
- повышение производительности исследователей;
- упрощение доступа к дорогой исследовательской инфраструктуре;
- увеличение скорости взаимодействия между исследователями;
- более быстрое распространение результатов исследований;
- и, возможно, самое важное — новое качество исследовательских результатов. ■

3. Сервлет является Java-интерфейсом, реализация которого расширяет функциональные возможности сервера. Сервлет взаимодействует с клиентами посредством принципа «запрос – ответ».