Информационная инфраструктура науки и образования

Мой доклад посвящен вопросам повышения эффективности информационного обеспечения науки и образования в нашей стране. Слава Богу, сейчас уже не нужно никому объяснять, что доступ к информационным ресурсам необходим, и что электронные ресурсы являются самым эффективным способом организации такого доступа. Сейчас на повестку дня выходят вопросы, связанные с тем, как организовать процесс информационного обеспечения максимально эффективно, как придать этому процессу системный характер. Речь идет о создании и развитии информационной инфраструктуры науки и образования. Какие задачи для этого необходимо решить?

Сделать зарубежную научную информацию более доступной, обеспечить благоприятные условия для создания и распространения российских ресурсов - это важнейшие задачи, про которые много говорилось и будет говориться на нашей нынешней конференции. Я сейчас не буду на этом подробно останавливаться. Однако для создания современной информационной инфраструктуры науки и образования недостаточно обеспечить доступ к источникам информации. Доступ к информационным ресурсам - это ведь не самоцель. Наша стратегическая задача - повышение научной результативности и качества образования. Поэтому очень важным элементом информационной инфраструктуры является создание таких условий, при которых наличие этой информации приводит к созданию нового знания и нового уровня образования, условий, при которых эффективность финансовых средств, затраченных на информационное обеспечение, была бы максимальной.

Это, во-первых, активное продвижение ресурсов и обучение, о чем тоже еще будут доклады на нашей конференции, и, во-вторых, создание специального инструментария, облегчающего для пользователя работу с информацией и повышающего эффективность этой работы. Сюда входит также и создание систем, позволяющих оценить эффективность использования информационных ресурсов и эффективность развития науки и образования в целом. На этом последнем разделе я и хотел бы сегодня остановиться подробнее и рассказать, что делает Научная электронная библиотека в этом направлении.

Система унифицированного доступа

Целью данной проекта является создание системы унифицированного доступа к распределенным информационным ресурсам разных типов (научным статьям, книгам, диссертациям, отчетам, препринтам, материалам конференций, базам данных и т.д.). Система может обеспечивать доступ к различным источникам информации (российским и зарубежным, публикуемым и непубликуемым, электронным и печатным, коммерческим и находящимся в открытом доступе).

В чем суть проблемы, на решение которой направлен этот проект? Дело в том, что по мере роста количества доступных электронных информационных ресурсов в Интернет работать с ними становится все труднее. Ресурсы постоянно развиваются - разделяются, сливаются, меняют местоположение в сети, появляются новые. Соответственно, устаревают и становятся недействительными ссылки на эти ресурсы. Кроме того, основная часть значимых информационных ресурсов (научные электронные журналы, книги и т.д.) распространяются по подписке и, следовательно, недоступны для поисковых роботов Интернет. Впрочем, даже для открытых ресурсов использовать распространенные поисковые службы типа Google или Yandex не очень эффективно, поскольку они дают слишком много шумовой информации, отсечь которую практически невозможно.

Даже если поисковая система проиндексировала тот или иной ресурс, т.е. имеет информацию о том, что и где находится, она не знает о том, имеет ли пользователь, отправивший поисковый запрос, право доступа к данному ресурсу. В результате, перейдя по ссылке, например, на полный текст научной статьи, пользователь обнаруживает, что саму статью получить он не может. В лучшем случае, это приводит к потере времени, в

худшем - к тому, что пользователь просто пропустит важную для него информацию. При этом вполне вероятно, что данная статья на самом деле доступна для данного пользователя в составе другого информационного ресурса, по другому адресу или, например, в библиотеке его собственной организации.

В рамках данного проекта проведена работа по сбору информации о всех значимых зарубежных и российских информационных ресурсах, их составе, временном охвате, тематической направленности, стоимости, условиях доступа, адресе в Интернет и других характеристиках, а также о правах пользователей (российских организаций) по отношению к этим ресурсам. Эта информация систематизируется и актуализируется, отражая происходящие изменения. На данный момент времени собрана информация о более 400 информационных ресурсах, среди которых ресурсы, представляющие коллекции журналов, включают в совокупности более 40 тысяч журналов. Работа по регистрации и описанию ресурсов, а также по их тематической рубрикации производится лучшими специалистами по электронным информационным ресурсам непосредственно с их рабочих мест через Интернет с помощью специально разработанного интерфейса. (интерфейс для регистрации ресурсов)

Что это дает организациям - пользователям системы? Теперь ответственный от организации может настроить систему для своей организации, указав, на какие ресурсы его организация подписана (просто выбрав их из списка). В результате для этой организации формируется автоматически единый, охватывающий все подписанные ресурсы список журналов, к которым организация имеет доступ, вместе с адресами этих журналов в интернет. Эти журналы уже тематически прорубрицированы, и их описания приведены к единому стандарту. Для пользователей же из этой организации все выглядит очень естественно - он просто щелкает на названии журнала и автоматически попадает на сайт соответствующего издательства, где бы оно не находилось.

Еще одной важной особенностью системы является возможность занесения в систему представителями библиотек информации о научных журналах, книгах и других информационных ресурсах, доступных в библиотеках в печатном виде с соответствующей переадресацией пользователя в библиотеку в случае отсутствия доступа к электронной версии необходимой статьи.

И, наконец, еще один пример. Предположим, вы нашли где-то ссылку на статью, например, в списке цитируемой литературы. Вы не знаете, какое издательство выпускает этот журнал, вы даже не всегда сможете правильно определить название этого журнала, поскольку оно дается в сокращенном варианте. Что вы можете сделать в данном случае, чтобы все-таки добраться до статьи? Можно попробовать поискать в Google, и потом долго листать страницы с результатами поиска. Он, кстати, может ничего и не найти, если журнал доступен только в закрытой базе данных. А можно отправить запрос в систему унифицированного доступа. Мы разработали алгоритм, позволяющий автоматически разбирать такую ссылку и делать запрос в базу знаний по информационным ресурсам. В результате обработки такого запроса пользователь будет автоматически переадресован либо на полный текст статьи, либо на страницу с аннотацией к статье, либо на страницу журнала.

Таким образом, система унифицированного доступа решает сразу три важные для повышения эффективности использования информационных ресурсов задачи:

1) Дает пользователю возможность узнать о существующих информационных ресурсах в интересующей его области науки, т.е. позволяет ориентироваться в мире научной информации;

2) Упрощает доступ к этим информационным ресурсам;

3) Создает основу базы знаний, необходимой в том числе и для других информационных проектов.

Я уже говорил в начале, что конечной целью развития информационной инфраструктуры является повышение результативности и качества научных исследований. Однако, возникает естественный вопрос - как оценить эту результативность и качественный уровень научной работы? Здесь я хотел бы перейти к наиболее интересному, на наш взгляд, и амбициозному проекту, над которым мы сейчас работаем - это создание российского индекса научного цитирования. Почему амбициозному? Дело в том, что попытки создать Российский индекс цитирования уже предпринимались несколько раз, но даже в советское время в условиях мобилизационной экономики это оказалось на под силу, поскольку на осуществление этого проекта требовалось колоссальное вложение средств. Действительно, несложно оценить требуемый объем работы. Всего в России сейчас издается около 3000 научных журналов. При среднем числе статей в год около 127 это составляет почти 400 тысяч статей в год. У каждой статьи в среднем 15 ссылок в списке цитируемой литературы. Таким образом, для обработки годового массива необходимо выделить, разобрать по отдельным полям, провести рубрикацию и ввести в БД более 6 миллионов ссылок. Если оператор потратит на одну ссылку хотя бы пять минут, это займет в общей сложности 500 тысяч часов рабочего времени. Таким образом, для обработки годового массива потребуется не менее 250 человек, т.е. целый институт. Кстати, среднее количество статей в год и среднее число ссылок в статье - эти цифры интересны сами по себе, поскольку получены впервые в результате статистической обработки российских журналов. Они показывают, что по числу статей на один журнал мы практически не отличаемся от мировой статистики, а вот число ссылок меньше примерно в два раза.

Глядя на эти цифры, кажется, что реализация этого проекта в нынешних условиях - это утопия. Тем не менее, мы утверждаем, что создание такой системы вполне реально, и более того, уже можем, продемонстрировать первые результаты, полученные в рамках выполнения этого проекта в течение полугода. За счет чего мы можем добиться более высокой производительности и эффективности расходования вложенных средств?

1) Широкое использование электронных версий научных журналов. Это позволяет отказаться от стадии сканирования и последующего редактирования текстов и приводит к уменьшению ошибок. По данным, полученным нами в результате анкетирования издательств, почти 90% журналов готовы предоставлять электронные версии журналов, по крайней мере, новые выпуски. Это же исследование показало, что из ответивших на анкету 60% издательств готовы предоставлять также и полные тексты статей в электронном виде, что придает качественно новый уровень создаваемой системе. Необходимо отметить, что стремление участвовать в проекте по созданию РИНЦ стимулирует издательства к переходу на электронные версии, и это само по себе является важным результатом выполнения проекта.

2) Широкое использование современных компьютерных технологий для автоматизации всей технологической цепочки. Мы разрабатываем программное обеспечение, которое позволяет заменить человека везде, где машина сама может принять решение. Оператору остается только функции контроля и коррекции ошибок автоматической обработки всего потока информации.

3) Бесплатное предоставление издательствами своих журналов для обработки в РИНЦ. Это позволяет экономить средства, затрачиваемые на подписку. Мы проводим систематическую работу со всеми российскими научными издательствами, объясняя им те преимущества, которые они получат от участия в этом проекте. Сейчас мы имеем уже соглашения с более 250 журналами, половина из которых входит в перечень ВАК. Серьезную помощь здесь оказал и проект создания электронных версий российских научных журналов, выполняемый Казанским государственным университетом, в результате которого добавилось более 40 журналов. Работа в этом направлении будет продолжаться, проект набирает обороты, уже сейчас в среднем в систему добавляется по одному новому наименованию российских журналов каждый день.

4) Широкое использование базы знаний, интеграция с системой унифицированного доступа. Мощности современных компьютерных систем вполне достаточно, чтобы в процессе обработки данных в реальном времени делать обращения в базы данных, содержащие десятки миллионов записей. Это позволяет использовать новую интеллектуальную технологию обработки данных, когда решение принимается на основании анализа уже накопленных системой знаний. При этом каждый раз, когда система сама не в состоянии принять правильное решение, к работе подключается оператор. Оператор не просто исправляет ошибку или разрешает неоднозначность, но и добавляет соответствующую информацию в базу знаний системы, что позволяет системе таким образом самосовершенствоваться. Создана общая база знаний, включающая в себя миллионы записей об авторах, организациях, журналах, вариантах их написаний и сокращений, и другая нормативная информация, используемая как для системы унифицированного доступа, так и для системы автоматического разбора пристатейной библиографии в РИНЦ.

5) Использование технологии распределенной работы. Для этого создаются виртуальные рабочие места, с помощью которых операторы могут могут работать с системой в любом месте и влюбое время через Интернет. Это снимает необходимость выделения дополнительных площадей для реализации этого проекта и уменьшает тем самым себестоимость проекта.

Итак , это были основные принципы организации работы по проекту, позволяющие оптимизировать затраты на его выполнение. Теперь я хотел бы остановиться на **принципах отбора журналов для РИНЦ**.

На первом этапе предполагается использовать для обработки научные журналы из списка ВАК, включающего более 1100 наименований. В дальнейшем список индексируемых журналов может корректироваться уже на основании данных из самого РИНЦ. Более того, импакт-факторы журналов или их аналоги, рассчитанные в РИНЦ, дадут возможность количественной оценки уровня журналов при принятии решения о их включении в список ВАК или исключении из него.

Надо сказать, что вопрос существования списка ВАК в его нынешнем виде сейчас широко обсуждается в научном сообществе. Я напомню, что этот список представляет из себя список российских научных журналов, в которых рекомендовано публиковать результаты работы для защиты докторских диссертаций. При этом даже публикации в ведущих зарубежных высокоимпактных журналах не засчитываются. Обосновывается это необходимостью широкого ознакомления российских ученых с результатами работы. Наверно, это действительно имело какой-то смысл в советские времена, когда и сформировался такой подход. Однако, в наше время несравнимо большее количество ученых во всем мире может ознакомиться с результатами работы, опубликованной в Интернет - в одном из электронных журналов или открытом архиве. Сейчас список ВАК имеет смысл только для определенной поддержки российских журналов. Если же мы ставим цель максимально широкого ознакомления с результатами работы, то необходимо требовать от журналов в качестве необходимого условия для включения в список ВАК обязательного наличия электронной версии, списков пристатейной библиографии и других параметров, отличающих современные научные периодические издания.

Также как и база данных SCI, которая служит не только для расчета индексов цитирования, но и является **многоцелевой поисковой системой** по научным публикациям для огромного количества пользователей во всем мире, РИНЦ должен стать важным источником достоверной библиографической информации для российских и

зарубежных пользователей, причем источником уникальным, поскольку до сих пор единой базы данных, охватывающей публикации российских ученых, не существует.

И еще один важный принцип формирования РИНЦ, отличающий его от аналогичных зарубежных систем. Мы будем отталкиваться не от научных журналов, отбирая из них лучшие, как это делает, например, ISI. У нас другая задача - **статистический анализ российской науки**. Поэтому нас интересуют работы российских ученых, независимо от того, где они были опубликованы. То есть мы отталкиваемся от авторов. Но как узнать, где и как часто публикуются российские ученые. Для решения этой задачи мы в рамках выполнения проекта создали специальный интерфейс, позволяющий ученым и представителям организаций самостоятельно регистрировать свои публикации в системе. Назвали мы эту подсистему Единый реестр публикаций российских ученых. (интерфейс регистрации публикации)

С помощью этого интерфейса авторы не только могут добавить новую публикацию, но и вносить исправления в библиографические описания уже имеющихся в системе статей, а также, например, удалить из списка своих публикаций статью, попавшую туда по ошибке (например, своего полного однофамильца). Естественно, любые изменения и дополнения, сделанные авторами, попадают в базу данных только после прохождения стадии контроля администраторами системы. Проблема в том, что ошибки в таких системах неизбежны, и их масса как в SCI, так и в SCOPUS. Мы же в данном случае предлагаем механизм, позволяющий поддерживать качество информации в РИНЦ на высоком уровне.

Зачем нужен российский индекс научного цитирования

Иногда приходится сталкиваться с мнением, что РИНЦ не нужен или даже вообще вреден, что бессмысленно поддерживать российские научные журналы и т.д., что все это приведет только к обособлению российской науки и ее окончательной деградации. Действительно, зачем создавать свой индекс цитирования, если есть широкораспространенная американская система SCI, зачем развивать российские научные журналы, если есть достаточное количество высокоимпактных зарубежных журналов, где могут опубликовать свои работы российские ученые? Наше общее мнение по этому вопросу можно коротко сформулировать таким образом.

Мы считаем, что научные журналы, поисковые системы, информационные сервисы, базы данных - все это элементы общей информационной инфраструктуры науки и образования в любой развитой стране. Невозможно развивать науку и образование и выводить ее на современный качественно новый уровень, не развивая информационную составляющую, роль которой в повышении эффективности научных исследований на самом деле только увеличивается, ведь новое знание рождается только в результате осмысления уже накопленного человечеством опыта. Что касается важности развития национальных проектов, я бы выделил несколько важных моментов:

1. Неадекватное представление российской науки в ISI. Из 3000 российских научных журналов лишь около 150 представлены в зарубежных базах, таких как WOS (ISI) или SCOPUS (Elsevier). В основном это переводные журналы. Здесь есть как объективные причины, так и субъективные:

- языковый барьер;
- уровень журналов, их доступность;
- национальные особенности цитирования;
- национальная обособленность некоторых направлений науки (пример "Известия ТИНРО").

Некоторые из этих причин в принципе решаемы. Например, перевод журнала или хотя бы библиографических описаний на английский язык, выпуск журнала в электронном виде значительно повышают шанс журнала быть включенным в список индексируемых в базе данных WOS. Это, безусловно, важно, и мы всячески движение в этом направлении поддерживаем, но какое отношение это имеет к научному уровню журнала, к качеству публикуемых в нем статей?

Следует отметить, что с аналогичными проблемами сталкиваются не только в России, но и в других неанглоязычных странах. Так, например, из более 4000 китайских научных журналов в SCI представлено только 30, т.е. менее 1 процента. Для решения проблемы объективной количественной оценки научных результатов в Китае еще в 1989 году был создан собственный индекс цитирования Chinese Science Citation Index, охватывающий сейчас более 1000 ведущих китайских журналов. Аналогичные проекты имеются и в Европе.

2. Сложность использования данных из SCI для статистического анализа. А это является основной задачей данного проекта. При попытках использовать данные из ISI для целей статистического анализа мы сталкиваемся с целым рядом проблем. Начиная с того, что многие вещи приходится делать вручную, поскольку интерфейс для этого не приспособлен, и заканчивая серьезными проблемами идентификации организаций и авторов. Для примера скажу лишь, что в аналитических отчетах ISI российская академия наук обычно учитывается как отдельная организация, хотя в нее входят несколько сотен самостоятельных научных институтов.

3. Отсутствие полноценной поисковой системы по российским научным журналам, включающей хотя бы оглавления журналов, не говоря уже о полных текстах. Этого действительно не существует, и надеяться на то, что какая-то зарубежная компания это сделает, не приходится.

4. Как уже отмечалось, важную роль этот проект играет для стимулирования российских издательств, для повышения уровня журналов, их конкурентоспособности. Мы не пытаемся выдвинуть вперед посредственность, сравнивая откровенно слабые журналы или научные статьи между собой. Наоборот, мы предоставляем возможность ОБЪЕКТИВНОГО сравнения этих журналов с лучшими мировыми журналами. Кроме то, включение журнала в РИНЦ будет способствовать его распространению в мире и, соответственно, повышению цитируемости публикуемых в нем статей.

5. Наконец, немаловажное значение имеет вопрос цены и доступности таких систем. К сожалению, стоимость зарубежных систем даже при подписке в составе консорциума составляет не менее 10-20 тысяч долларов в год, что для большинства организаций просто неприемлемо.