Российский индекс научного цитирования - утопия или реальность?

Г.О.Еременко
Научная электронная библиотека eLIBRARY.RU

Мой доклад называется "Российский индекс научного цитирования - утопия или реальность?". Я думаю, не в этой аудитории объяснять, что такое индекс цитирования и зачем он нужен. Многие наверняка работали с продуктом ISI Science Citation Index (SCI) или с его интернет версией Web of Science. Вы также уже наверное обратили внимание, что в последнее время не только ISI, но и другие ведущие поставщики научных информационных ресурсов стали дополнять свои продукты функциональностью, связанной с использованием пристатейных ссылок. Это и система SCOPUS Elsevier, которая сейчас рассматривается как вполне реальный конкурент для Web of Science. И как мы вчера видели, например, работает в этом направлении и компания EBSCO, и другие. Многие научные издательства уже производят метаописания статей вместе с пристатейной библиографией. То есть обработка пристатейных ссылок постепенно становится стандартом для информационных продуктов современного уровня.

В чем причины такого интереса к этим ссылкам? Дело тут не только в возможности расчета индекса цитирования и его наукометрическом значении, как показателя результативности научной деятельности ученых, научных коллективов и т.д. Возможность перейти по ссылке на соответствующую статью из пристатейной библиографии, или наоборот, перейти на статью, ссылающуюся на данную, сама по себе является очень удобным и полезным средством навигации в системе.

Индекс цитирования часто применяется и в качестве критерия объективной оценки научного уровня ученых, коллективов, организаций, стран или регионов. Во многих научных фондах он используется в качестве показателя при проведении экспертизы и выделении грантов. В некоторых странах он используется и на государственном уровне для оценки деятельности научных организаций и влияет на распределение финансирования. В последнее время все чаще обсуждается вопрос о возможности применения индекса цитирования для оценки уровня организаций и научных коллективов и в России. Особенно это стало актуальным в связи с проводимой реформой науки и образования в России и, соответственно, необходимостью статистического анализа состояния науки в России и потребности в объективных количественных критериях для оценки научной деятельности при распределении финансирования на конкурсной основе.

Казалось бы, в чем сложность? Берем базу данных SCI, охватывающую все страны, в том числе и Россию, и проводим необходимый анализ. Однако проблема в том, что

использование SCI для анализа российской науки дает далеко не полную картину и, соответственно, не может считаться вполне объективным критерием, поскольку лишь малая часть российских журналов обрабатывается ISI. Всего в SCI представлено около 70 российских журналов, в основном издаваемых на английском языке или имеющих английскую версию, тогда как американских, например, около 1500, т.е. почти 40% от общего числа индексируемых журналов. В то же время только список российских рецензируемых журналов, рекомендованных ВАК, составляет почти 1000 наименований, а общее число научных журналов, издаваемых в России, по крайней мере в два-три раза больше.

Можно, конечно, сказать, что каково развитие науки в России, таково и ее отражение в SCI. Да, конечно, в какой-то степени это отражает реальную картину, но в то же время существует целый ряд объективных и субъективных причин, из-за которых вклад российских ученых в мировую науку отражается неадекватно. Среди этих причин можно упомянуть следующие:

1. **Языковый барьер**. Не секрет, что SCI в основном ориентируется на англоязычные журналы или, по крайней мере, журналы, предоставляющие библиографию и аннотации статей на английском языке. Многие российские журналы, в том числе достаточно уважаемые в российском научном сообществе, не предоставляют такой возможности, что резко снижает их шанс попасть в список журналов, индексируемых службой SCI. Кроме того, даже при наличии аннотации, но отсутствии полного текста статьи на хорошем английском языке зарубежные ученые испытывают трудности при ознакомлении с результатами работы, что приводит к резкому падению цитирования этих статей. К сожалению, прошли времена, когда зарубежные ученые изучали русский язык только для того, чтобы иметь возможность читать статьи в российских журналах в оригинале, и когда в российских журналах публиковались ведущие зарубежные ученые.

2. **Особенности отбора журналов на основании индекса цитирования**. SCI отбирает журналы для включения в свой список на основании их импакт-факторов, отражающих суммарное цитирование статей, опубликованных в данном журнале. В то же время известно, что особенности национального менталитета могут приводить к существенному искажению объективной картины. Так, например, американский ученый при прочих равных условиях скорее всего сошлется в своей статье на работу его американского коллеги (около 70% ссылок), чем на работу китайского или российского ученого. В то же время, российский исследователь, наоборот, с большей вероятностью предпочтет сослаться на работы зарубежных ученых, опубликованные в ведущих зарубежных изданиях. В результате импакт-факторы российских журналов могут быть значительно занижены.

3. **Уровень российских журналов**. На отбор журналов для SCI влияет также и качество самих журналов, их соответствие мировым стандартам, регулярный выход, наличие пристатейной библиографии. Среднее число ссылок пристатейной библиографии в российских журналах, по некоторым оценкам, не превышает 10, что почти в 2-3 раза меньше соответствующего показателя для журналов, входящих в SCI. Некоторые российские журналы вообще выходят без списков цитируемой литературы.

4. **Особенности научного развития в разных областях**. Существуют целые направления науки, которые развиваются относительно локализовано и в определенной степени замкнуты в рамках страны или региона. Особенно это ярко проявляется в некоторых областях общественных и гуманитарных наук. При этом внешнее цитирование на эти работы гораздо меньше, чем, например, в области более интернационализированных естественных наук.

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.

Какой выход из этой ситуации? Конечно, повлиять на политику отбора журналов в ISI мы не можем, поэтому нужно либо, чтобы все российские журналы выходили с англоязычной версией, что пока трудно себе представить, поскольку повлиять на них мы тоже не можем, либо создавать аналогичный SCI продукт, обрабатывающий российские журналы - Российский индекс научного цитирования. Нужно сказать, 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.

Тем не менее, Министерство образования и науки РФ посчитало необходимым предпринять еще одну попытку и объявило в этом году конкурс на создание Российского индекса цитирования с общим бюджетом в 20 миллионов рублей. Согласно условиям конкурса требовалось не только разработать технологие решения, программное обеспечение и т.д., но и полностью обработать годовой массив из 1000 ведущих российских журналов. И все это необходимо сделать за полтора года. Поскольку данное направление для нас очень интересно, и, более того, мы давно ведем работы, связанные с обработкой пристатейной библиографии для российских журналов, мы решили участвовать в этом конкурсе. Мы провели всесторонний анализ возможности создания такого продукта в заданные сроки и пришли к выводу, что несмотря на относительно небольшое для такого глобального проекта финансирование, перспектива создания Российского индекса научного цитирования вполне реальна. В результате мы написали заявку, где подробно изложили наш подход к решению данной проблемы, и буквально вчера мы получили подтверждение, что стали победителями в этом конкурсе.

Таким образом, сегодня я могу официально объявить о запуске в России нового масштабного информационного проекта, у которого пока даже нет названия. Поэтому мы объявляем конкурс на лучшее название для этого проекта. Пока же давайте условно называть это РИНЦ - Российский индекс научного цитирования.

Основная задача, которая стоит перед российским индексом научного цитирования - максимально полный охват российских журналов. В то же время, необходимо подчеркнуть, что главный принцип, который, на наш взгляд, должен быть заложен в основу создания РИНЦ и который определяется стратегической целью разработки системы - это прежде всего анализ научной деятельности российских ученых и организаций, а не анализ публикаций в российских научных журналах. Основная проблема, возникающая на пути решения этой задачи, заключается в том, что весь поток публикаций российских ученых в научных журналах (ориентировочно составляющий около 150 тысяч статей в год) разбивается на несколько частей:

1. Публикации в ведущих зарубежных журналах, входящих в список SCI (5%);
2. Публикации в зарубежных журналах, не индексируемых SCI (5%);
3. Публикации в ведущих российских журналах, входящих в список SCI (10%);
4. Публикации в российских журналах, не индексируемых SCI (80%).

Несмотря на то, что количество статей российских авторов (или с участием российских авторов), описанных в SCI, относительно невелико (поток публикаций п.1 вместе с п.3 ежегодно составляет примерно около 23000 статей, т.е. не более 15% от всего ежегодного объема публикаций российских ученых),

этот поток имеет не менее важное значение, чем публикации в российских журналах, поскольку в этих журналах обычно публикуются самые выдающиеся научные работы. Также необходимо отметить, что для объективной оценки результатов научной деятельности необходимо учитывать не только публикации в научной периодике, но также книги и монографии, написанные российскими авторами, материалы конференций, докторские диссертации.

Таким образом, для объективной оценки научной деятельности необходимо создать систему, позволяющую учитывать все эти потоки публикаций и определяющую суммарный индекс цитирования российских авторов и организаций по публикациям как в ведущих российских, так и зарубежных научных журналах. При этом для анализа публикаций российских ученых в зарубежных и переводных российских журналах можно использовать данные SCI, а для основной массы российских журналов такую информацию можно получить, только создав аналогичный механизм индексирования научных статей и пристатейной библиографии в России - российский индекс научного цитирования.

Таким же образом, кстати, поступили и в Китае. Китайский индекс цитирования охватывает только китайские журналы, а для комплексной оценки научной деятельности ученых и организаций разработана специальная система наукометрических показателей Chinese Scientometric Indicators, основанная на обработке информации как из SCI, так и из китайского индекса цитирования.

Следует отметить еще одну серьезную проблему, возникающую при попытках проанализировать научную активность отдельных ученых и организаций с помощью SCI. Это сложность идентификации авторов и, в особенности, организаций. Действительно, технология создания индекса цитирования SCI основана на выделении всей необходимой информации только из текста самих публикаций. При этом в разных журналах могут различаться стандарты библиографического описания самих статей и пристатейных ссылок. Например, в одном случае имя автора пишется полностью, а в другом указываются только инициалы. В результате, при совпадении фамилии и инициалов (а для больших баз данных это обычная практика) система не может четко разделить авторов, что приводит к ошибкам как в количествах статей данного автора, так и в индексах цитирования. Еще более сложная ситуация с организациями. Здесь, кроме проблемы идентификации, связанной с различными вариантами написания названия и адреса организации, есть и проблема привязки автора к конкретной организации. Автор может несколько раз сменить место работы или, например, опубликовать статью, работая за рубежом по научному обмену и т.д. Кроме того, иногда в статье с несколькими авторами в качестве адреса для контактов приводится только один адрес, несмотря на то, что авторы - из разных организаций.

Таким образом, для создания системы, позволяющей проводить количественный анализ и сравнение научной продуктивности ученых и организаций, необходимо из достоверных источников сформировать базу данных российских ученых и организаций с возможностью привязки публикаций к этой базе данных. Эта база данных должна содержать выверенную и нормализованную информацию и постоянно актуализироваться.

Теперь я хотел бы подробнее остановиться на том, что из себя будет представлять этот продукт, чем он будет отличаться от зарубежных аналогов, каким образом мы планируем организовать его производство.

Для решения поставленных задач необходимо осуществить целый комплекс взаимосвязанных работ, из которого можно выделить три основные направления:

1. Создание российского индекса научного цитирования (РИНЦ), охватывающего публикации в ведущих российских научных журналах.
2. Создание единого реестра научных публикаций российских ученых (ЕРНП).
3. Разработка поисковой системы и интерфейса для работы пользователей с данными РИНЦ и ЕРНП.

Рассмотрим более подробно каждое из этих направлений.

1. Создание российского индекса научного цитирования.

Российский индекс научного цитирования (РИНЦ) представляет собой базу данных, аналогичную SCI, сформированную в результате обработки ведущих российских научных журналов и содержащую библиографическую информацию, извлеченную из текста статей, а также пристатейных ссылок.

Вопрос отбора журналов для включения в РИНЦ имеет важное значение. На первом этапе предполагается использовать для обработки научные журналы из списка ВАК, включающего около 1000 наименований. В дальнейшем список индексируемых журналов может корректироваться уже на основании данных из самого РИНЦ, а также единого реестра научных публикаций российских ученых. 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.

Мы считаем необходимым включать в обработку РИНЦ все ведущие российские журналы, в том числе входящие в список обрабатываемых SCI, несмотря на то, что это в определенной степени дублирование работы, проводимой производителями SCI. Причины для этого следующие:

Во-первых, как уже отмечалось выше, для полного анализа необходимо использовать оба источника (SCI и РИНЦ). При этом нужно быть уверенным, что наукометрические показатели, полученные из разных систем, можно корректно сравнивать. Обработка и анализ данных из обеих систем по 70 российским журналам позволит определить, насколько эти системы совместимы. Кроме того, такое сравнение само по себе представляет наукометрический интерес.

Во-вторых, РИНЦ будет обрабатывать русскоязычные версии переводных журналов и, соответственно, собирать библиографическую информацию на русском языке, которая отсутствует в SCI.

И, наконец, относительное количество этих журналов не так велико (не более 10%), так что не сильно повлияет на общий объем работы.

В перспективе, в состав журналов, обрабатываемых РИНЦ, могут быть включены и некоторые зарубежные журналы (например, журналы, издаваемые в странах СНГ на русском языке или журналы на иностранных языках, которые не охвачены SCI), в которых регулярно публикуются статьи российских авторов.

Немаловажное значение имеет и глубина охвата обрабатываемого архива журналов. Для того, чтобы создаваемый информационный продукт приобрел действительную ценность для потребителя, необходимо обработать массив журналов хотя бы за 3 года. В этом случае уже можно будет посчитать импакт-факторы журналов. В перспективе необходимо довести глубину архива не менее

чем до 10 лет - по некоторым оценкам, научные публикации за этот период покрывают не менее 85% всех текущих поисковых запросов пользователей.

Успех работы по созданию РИНЦ во многом зависит от правильно выбранной технологии обработки научных изданий и формирования библиографической базы данных. Традиционный подход, используемый в SCI, основан на подписке и полном сканировании и распознавании всех выпусков обрабатываемых журналов. Затем оператор в ручном режиме выделяет в тексте название статьи, авторов, адрес, номер тома, страницы и т.д. и копирует эти данные в поля базы данных. Аналогичным образом обрабатываются библиографические ссылки к статье. При этом одновременно проводится автоматическая проверка на наличие в базе данных библиографического описания для каждой ссылки, и если она уже имеется, это засчитывается в ее счетчике цитирования. В целом, эта работа очень трудоемка и требует предельного внимания со стороны оператора, в результате чего его производительность не может быть высокой, иначе велика вероятность появления ошибок в базе данных (при том, что каждая ошибка - это потерянная ссылка на работу конкретного ученого или же вообще потеря этой публикации в привязке к данному автору). Именно этим объясняется и высокая цена подписки на базы данных SCI, и его практически полный монополизм в данной области в течение многих лет. В то же время, тщательные исследования информации, содержащейся в SCI, проведенные некоторыми исследователями-наукометристами, показали, что, несмотря на усилия производителей SCI по поддержанию высокого качества обработки данных, ошибок в данных SCI остается немало.

Однако в последнее время в результате развития как программных алгоритмов обработки и распознавания информации, так и все большего распространения электронных версий научных журналов появились новые возможности для упрощения и удешевления производства информационных продуктов, связанных с обработкой пристатейной библиографии. В результате появилось несколько проектов, в какой-то степени конкурирующих с SCI, т.е. предлагающих аналогичный сервис.

Так, например, поисковая система SCOPUS компании Elsevier, запущенная в прошлом году уже сейчас обрабатывает больше журналов, чем SCI (хотя и с меньшим временным охватом). Технология производства этого продукта больше ориентирована на обработку электронных версий журналов, поступающих от издательств. Учитывая то, что подавляющая часть научных журналов, издаваемая зарубежными издательствами, имеет сейчас официально продаваемые электронные версии, и, более того, многие издательства уже самостоятельно производят разбор и структуризацию библиографических ссылок к статьям в своих журналах, такой подход является безусловно перспективным. Он позволяет по крайней мере отказаться от стадии сканирования и распознавания, приводящей к значительному количеству ошибок, особенно в пристатейных ссылках, набранных, как правило, мелким шрифтом. Правда, справедливости ради, необходимо отметить, что цена подписки на SCOPUS ненамного дешевле, чем на SCI.

Есть и примеры успешной реализации технологии, основанной на автоматическом распознавании, извлечении и разборе пристатейной библиографии из текста статьи. В качестве примера можно привести систему ResearchIndex (NEC Research Institute). Разработанное в рамках данного проекта программное обеспечение позволило создать базу данных из более 300000 научных статей в области компьютерных технологий, собранных из различных

источников в Интернет, и автоматически обработать пристатейную библиографию к этим статьям. Конечно, в полностью автоматическом режиме система не обеспечивает стопроцентное отсутствие ошибок - по проведенным тестам, при разборе ссылок она корректно отрабатывает 95% ссылок. Это и не удивительно, если учесть, что довольно много ошибок содержится изначально в самих описаниях ссылок, сделанных авторами статей, причем часто эти ошибки кочуют из одной статьи в другую, когда авторы просто копируют ссылку из чужой статьи в свою. Без участия человека компьютеру решить такую задачу сложно, хотя и в этом направлении есть варианты решений.

С нашей точки зрения, оптимальным вариантом на нынешний момент времени можно считать технологические решения, основанные на совмещении автоматического разбора пристатейной библиографии и последующего ручного контроля информации перед ее занесением в базу данных. Такой подход использует, например, компания Parity Computing, которой в результате удалось достичь точности обработки ссылок 99%. Этот уровень уже вполне сопоставим с качеством информации в SCI.

Такой же подход уже несколько лет разрабатывается в Научной электронной библиотеке eLibrary.Ru для автоматической обработки пристатейной библиографии в процессе разметки текста статьи и формирования библиографического описания в формате XML, позволяющего осуществлять автоматическую загрузку таких описаний в базу данных НЭБ. Модуль разбора пристатейной библиографии входит в состав программы разметки, разработанной в НЭБ и бесплатно предоставляемой российским издательствам, размещающим свои издания в Научной электронной библиотеке.

Из-за ограниченности времени моего выступления я не буду останавливаться сейчас подробно на деталях технологии автоматической обработки журналов и путях повышения качества этой обработки. Эта тема для отдельного разговора. Отмечу лишь, что это будет интеллектуальная самообучаемая система с возможностью эвристического анализа, опирающаяся при принятии решений на мощную, постоянно совершенствующуюся собственную базу знаний.

Для проведения анализа тенденций развития отдельных научных направлений необходимо, чтобы библиографические записи, попадающие в базу данных РИНЦ, были прорубрицированы, причем желательно, чтобы такая тематическая рубрикация была проведена на уровне отдельных статей. Следует отметить, что это представляет собой достаточно сложную и трудоемкую задачу, поскольку процесс назначения рубрик практически не поддается автоматизации (хотя такие разработки существуют, в том числе и в России). Причем для проведения этой работы с высоким качеством необходимо привлекать большое число очень квалифицированных специалистов. Не случайно даже в таких авторитетных и дорогостоящих продуктах, как SCI или SCOPUS, не осуществляется постатейной рубрикации входного потока публикаций. Назначение рубрик в этих системах производится на уровне журналов.

Какой рубрикатор выбрать для использования в РИНЦ? С точки зрения выдерживания стандартов оптимальным вариантом является использование Государственного рубрикатора научно-технической информации (ГРНТИ). Так, например, Научная электронная библиотека для рубрикации более 6000 зарубежных и российских журналов в электронном виде использует рубрикатор ГРНТИ. В то же время в печатных российских научно-технических изданиях используется, как правило, Универсальная десятичная классификация (УДК). В

базе данных SCI используется свой рубрикатор, что необходимо учитывать, поскольку при совместной обработке данных о российских публикациях из РИНЦ и SCI возникнет задача их сведения к единой системе рубрикации.

Рассмотрим возможные технологические варианты проведения тематической рубрикации:

1. **Ручная рубрикация.** Каждая статья просматривается специалистом-библиографом с проставлением соответствующей рубрики (или нескольких рубрик). Поскольку физически собрать в одном месте большое количество таких специалистов невозможно, для проведения этой работы создается специальный интерфейс, позволяющий одновременно работать многим специалистам с центральной базой данных через Интернет.

2. **Автоматическая рубрикация с использованием кодов УДК**, проставляемых в журналах. Довольно значительная часть российских журналов использует УДК для постатейной рубрикации. Такие журналы можно попробовать прорубрицировать автоматически, с помощью таблицы соответствия между рубриками УДК и ГРНТИ.

3. **Самостоятельная рубрикация авторами** при регистрации публикаций в Едином реестре научных публикаций (подробнее о ЕРНП см. ниже). Этот вариант кажется весьма привлекательным, поскольку, во-первых, позволит прорубрицировать также статьи, не входящие еще в РИНЦ, во-вторых, должен обеспечить вполне приемлемое качество (в конце концов, кто как не автор лучше других разбирается в тематике своей работы) и, в-третьих, не требует привлечения внешних специалистов и соответственно увеличения затрат на производство РИНЦ. Кроме того, довольно распространена практика, когда редакции российских научных журналов требуют от авторов самостоятельно проставлять рубрики УДК при отправке своей статьи для публикации.

4. **Рубрикация на уровне журналов с использованием данных цитирования**. Этот метод применяется, например, при подготовке информационного аналитического продукта, построенного на использовании индексов научного цитирования SCI - ISI Essential Science Indicators. По этой методике все журналы разбиваются на две группы. Узкоспециализированным журналам присваиваются соответствующие рубрики, и все статьи, опубликованные в этих журналах, автоматически получают эти же рубрики. Для мультидисциплинарных журналов рубрикация на уровне журнала смысла не имеет, поэтому для рубрикации статей в них используется другой метод. Автоматически анализируются на предмет тематической направленности статьи, на которые ссылается данная статья, и статьи, которые ссылаются на данную статью. Основная тематика по всей совокупности этих ссылок и проставляется для обрабатываемой статьи из мультидисциплинарного журнала.

Все перечисленные выше варианты проведения тематической рубрикации могут быть использованы в процессе обработки российских журналов для РИНЦ. В течение первого года реализации проекта предполагается провести тестирование всех вариантов с оценкой качества рубрикации. По результатам испытаний будут подготовлены методические рекомендации, в соответствии с которыми и будет проведена рубрикация основного массива журналов в течение 2006 года.

Теперь рассмотрим более подробно предлагаемую технологию подготовки выпусков журналов с точки зрения получения исходных текстов статей для последующего автоматического разбора. Прежде всего, отметим, что технология

обработки зависит от того, в каком виде доступны исходные материалы. Здесь может быть три основных варианта:

1. **Журнал доступен в электронном виде**. Если при этом статьи предоставляются в виде текстового форматированного файла (в формате PDF или RTF), то это самый оптимальный вариант, поскольку исключается стадия сканирования и распознавания текста. Однако в случае, если электронная версия сама по себе представляет уже отсканированные страницы журнала, то это может даже усложнить задачу, поскольку качество сканирования может оказаться недостаточным для корректного распознавания, а оригинал для повторного сканирования отсутствует.

Использование текстовых электронных версий журналов для обработки при создании РИНЦ является, безусловно, самым приемлемым вариантом, дающим на выходе самое высокое качество и имеющим при этом минимальную трудоемкость. Особенно перспективным в этом плане является вариант, когда данные из издательства поступают в уже структурированном виде (например, в формате XML), и пристатейные ссылки при этом уже разобраны или, по крайней мере, разделены на отдельные записи.

Все электронные версии журналов, размещаемые в НЭБ, будут обрабатываться по этой технологии. Также весьма перспективным является сотрудничество в этом направлении между НЭБ и Информрегистром, который является соисполнителем по данному проекту. Обязательные копии электронных изданий, предоставляемые на хранение в Информрегистр, могут обрабатываться для включения их библиографических описаний в состав базы данных РИНЦ.

В то же время необходимо констатировать, что на данный момент времени не очень большая часть российских издательств в состоянии предоставить электронные версии своих журналов, хотя прогресс в этом направлении наметился. По нашим оценкам, получить полноценные электронные версии непосредственно от издательств в 2006 году удастся не более чем для 200 журналов. Основную часть новых выпусков российских журналов пока можно получить, только подписавшись на их печатные версии.

2. **На журнал можно оформить подписку**. Этот подход в ближайшие несколько лет, скорее всего, будет основным при обработке новых выпусков российских печатных журналов. Его преимущество заключается в том, что подписавшись на журнал и имея его печатную версию, специально предназначенную для обработки, можно разрезать выпуски журнала на страницы и автоматически сканировать в поточном сканере. Скорость и качество такого сканирования гораздо выше, а стоимость - гораздо ниже, чем при ручном сканировании, когда оператору приходится вручную переворачивать страницы выпуска и заправлять его в сканер. По мере перехода все большего количества российских журналов на распространение в электронном виде этот способ обработки будет постепенно замещаться на первый вариант.

3. **Журнал можно найти только в архивах библиотек**. В этом случае журнал сканируется вручную, поскольку специального экземпляра для автоматического сканирования найти невозможно. При этом достаточно провести сканирование и распознавание только страниц, содержащих необходимую для ввода в систему информацию (как правило, это первая и последняя страницы каждой статьи выпуска). Этот вариант придется применять для обработки архивных выпусков с привлечением библиотек, в которые поступают обязательные копии российских научных журналов. Для проведения этой работы в качестве соисполнителей в данном проекте привлечены Государственная публичная научно-техническая библиотека СО РАН, которая имеет в своих фондах и может обработать около 600 российских журналов из списка ВАК в

области естественных наук, и Институт научной информации по общественным наукам РАН, который может обработать около 200 журналов в области общественных и гуманитарных наук.

Для обработки архивных версий может использоваться также и технология, традиционно применявшаяся для подготовки реферативных изданий еще во времена, когда сканирование и автоматическое распознавание текстов были недоступны. Речь идет о ручном вводе необходимой информации из текста статьи в базу данных. Однако на нынешнем уровне технологического развития вряд ли стоит рассматривать этот вариант для решения задач обработки больших массивов информации ввиду его высокой трудоемкости и стоимости.

Все три варианта технологии подготовки журналов могут быть и, скорее всего, будут, хотя и в разной степени, использованы при создании РИНЦ. В процессе реализации проекта будут отлажены технологии, позволяющие в дальнейшей работе по наполнению РИНЦ работать как электронными журналами, так и с печатными, как с новыми поступлениями, так и с архивными выпусками.

2. ЕРНП

Единый реестр научных публикаций представляет собой базу данных, содержащую следующую информацию:

- **Полный выверенный список российских научно-образовательных организаций** в привязке к ведомствам, городам, регионам и т.д. (всего около 3000 организаций). Кроме идентифицирующей и контактной информации может содержать перечень количественных показателей, отражающих научную деятельность в организации (число ученых, студентов, аспирантов, кандидатов, докторов, общий бюджет, информацию о структурных подразделениях и тематических направлениях и т.д.).
- **Выверенный список российских ученых - авторов научных публикаций** в привязке к организациям (всего около 300000 человек). Включает фамилию, имя и отчество ученого на русском и английском языках (в том числе возможные варианты английского написания), год рождения, пол, контактную информацию, ученую степень, ученое звание, должность, ключевые слова и коды научного классификатора. Также может содержать другую информацию, отражающую и характеризующую научную деятельность ученого (наличие патентов, грантов, научных наград, участие в редколлегиях, членство в научных обществах, число аспирантов, студентов и т.д.)
- **Полный список российских научных периодических изданий** (всего около 3000 записей) с подробной информацией о каждом из них (ISSN, тематическая направленность, состав редколлегии, годы выпусков, контактная информация, условия подписки, адрес в Интернет и т.д.).
- **Библиографическая база данных публикаций российских ученых** (за период 10 лет - 1,5 млн. записей). Эта библиографическая база данных должна охватывать научные публикации российских ученых в российских и зарубежных журналах, книги и монографии российских авторов, материалы конференций, докторские диссертации.

3. Разработка поисковой системы и интерфейса для работы пользователей с данными РИНЦ.

Появление Российского индекса научного цитирования создаст основу для развития мощного аналитического инструментария, позволяющего не только получать количественные данные о научной продуктивности отдельных ученых, организаций, регионов и т.д., но и проводить самые разнообразные наукометрические исследования. Для этого необходимо разработать удобный интерфейс, дающий возможность не только осуществлять поиск и просмотр нужных данных, но и позволяющий решать определенные аналитические задачи.

Рассмотрим вначале основные принципы реализации пользовательского интерфейса РИНЦ:

1. 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. Соответственно, интерфейс РИНЦ должен иметь удобные средства для поиска публикаций по различным параметрам, характерные для современных поисковых информационных систем.

2. Мы считаем целесообразным интегрировать интерфейс РИНЦ с возможностями, предоставляемыми Научной электронной библиотекой. В этом случае, пользователь, найдя нужную статью с помощью поисковой системы РИНЦ, может тут же получить и полный текст этой статьи, размещенный в НЭБ (при условии, конечно, что он имеет на это соответствующие права). Кроме того, возможна и более тесная интеграция этих двух информационных систем, рассчитанных в первую очередь на российских пользователей - с единой системой регистрации, статистики, базой данных пользователей и организаций и т.д. В результате такого подхода к организации интерфейса системы появится и еще одна уникальная возможность для объективной оценки уровня востребованности научных публикаций - на основании учета количества запросов пользователей, запрашивающих полный текст данной статьи. Этот параметр отражает интерес широкой читательской аудитории к самой статье. Совместный анализ статистики цитирования статьи и статистики обращений к ней пользователей позволит получить уникальные возможности для наукометрических исследований.

3. Библиографическая информация, доступная в РИНЦ, может быть дополнена информацией о наличии полных текстов статей в различных внешних базах данных. В этом случае, даже если полный текст статьи отсутствует в НЭБ, пользователь получит ссылку на эту статью из другого источника в Интернет (сервера издательства или агрегатора информационных ресурсов). Такая тенденция развития, направленная на предоставление пользователю не только поисковых возможностей, но и доступа к полным текстам, характерна сейчас для всех библиографических баз данных. Использование полнотекстовой базы данных НЭБ, содержащей уже более 8 миллионов статей, вместе с информацией о доступных внешних полнотекстовых информационных ресурсах, позволит превратить РИНЦ в мощную многофункциональную поисково-аналитическую систему, в равной степени полезную как для ученых, студентов и аспирантов, так и для специалистов в области наукометрии.

4. Включение в РИНЦ данных, полученных через Единый реестр научных публикаций (ЕРНП) создает дополнительные аналитические возможности,

которые отсутствуют в SCI. Речь идет о точной привязке публикаций к российским авторам и организациям. Используя дополнительную информацию из регистрационных данных авторов и организаций, можно проводить интереснейший для наукометрии анализ научной продуктивности и цитируемости авторов и научных коллективов в зависимости от многих факторов. Соответственно, аналитическая часть интерфейса системы должна предусматривать возможность проведения такого анализа.

5. Очевидно, что все возможные статистические выборки не могут быть заложены в интерфейс разрабатываемой в рамках данного проекта системы. Поэтому для того, чтобы при необходимости можно было провести какой-то дополнительный анализ с использованием внешних программных средств или для сведения вместе и сравнения данных из разных информационных систем, необходимо предусмотреть возможность экспорта исходных данных из РИНЦ в нескольких распространенных форматах.

При создании интерфейса РИНЦ должны быть реализованы следующие функциональные возможности:

1. Возможность поиска и отбора публикаций, авторов, организаций, журналов

2. Персональные профили авторов и организаций

5. Возможность формирования и работы с подборками статей, журналов, авторов и организаций.

6. Представление для каждого автора, организации, журнала количества публикаций и количества ссылок с возможностью разбивки по годам.

7. Возможность вывода данных в виде зависимости одного параметра от другого с табличным или графическим представлением полученного распределения.

8. Возможность экспорта библиографических данных в системы персональной работы с библиографией (типа Reference Manager, EndNote и т.д.)

9. Возможность перемещения назад по времени, используя цитируемые ссылки, для выявления исследований, которые оказали влияние на работу автора.

10. Возможность перемещения вперед по времени, используя ссылки на данную статью, для выявления влияния данной работы на текущие исследования.

11. Возможность вывода релевантных (родственных) работ, которые имеют одну или более одинаковых цитируемых ссылок.

12. Возможность сохранения поисковых запросов и их последующего повторного использования.

13. Возможность настройки почтовой рассылки новых поступлений в соответствии с запросом пользователя.

14. Интерфейс системы должен быть реализован как на русском, так и на английском языках.

Кроме этого, должны быть предусмотрены дополнительные административные возможности для представителей организаций и администраторов системы.

Очень важно понимать, что успех такого глобального проекта как создание Российского индекса научного цитирования, определяется также и тем, каким образом проект сможет продолжать свое существование после окончания действия государственного контракта. В оптимальном варианте проект должен перейти на самоокупаемость и продолжать самостоятельное развитие за счет платной подписки на свои услуги для российских и зарубежных организаций.

В то же время, как уже отмечалось выше, затраты на обработку годового массива российских журналов для РИНЦ достаточно велики, а ведь нужно еще постепенно обрабатывать архивы за прошлые годы.

Именно поэтому очень важно на этапе создания РИНЦ отработать технологические решения, позволяющие максимально автоматизировать и оптимизировать работу по обработке журналов, чтобы довести себестоимость производства данного информационного продукта до приемлемой величины. В конце 2006 года необходимо также провести маркетинговые исследования с тем, чтобы определить потенциальное количество подписчиков и оптимальную ценовую политику распространения продукта. По нашим очень предварительным оценкам, стоимость доступа к РИНЦ для российской организации, позволяющая проекту развиваться самостоятельно, может составить не более 20 тысяч рублей в год (для сравнения стоимость подписки на Web of Science или SCOPUS даже для организаций, входящих в крупные консорциумы, составляет не менее 10 тысяч долларов). Минимальное количество подписчиков, необходимое для поддержания проекта и обработки массива журналов текущего года, составляет 250 организаций. При большем количестве подписчиков можно параллельно обрабатывать архивные выпуски журналов.