

VII Международная научно-практическая конференция "SCIENCE ONLINE: электронные информационные ресурсы для науки и образования"

13-20 мая 2006 г., отель "Mediteran", Бечичи, Черногория

Инна В. Зибарева, Надежда В. Круковская, Вячеслав М. Бузник

«Chemical Abstracts» в сети STN – база данных для библиометрического исследования российской химической науки

На представительных примерах (отдельные ученые, исследовательские институты, академические научные центры, химические дисциплины, химические журналы) рассмотрена применимость базы данных Chemical Abstracts (БД СА), доступной через международную научно-техническую сеть STN International, для библиометрического изучения различных аспектов российской химической науки. Подчеркнуты достоинства (глубокая ретроспектива, широкий охват российских источников, быстрое обновление, наличие классификаторов и рубрикаторов, учет цитирования) и ограничения (неполная информация в некоторых полях) БД СА, которые следует иметь в виду при проведении подобных исследований, включая отбор и анализ исходных данных. Доступ к БД СА через сеть STN позволяет использовать командный язык STN Messenger, обеспечивающий эффективный многопараметровый online анализ больших объемов статистической информации.

Библиографическая информация, содержащаяся в научно-технических базах данных (БД), все чаще используется для объективной характеристики состояния и тенденций развития как фундаментальных, так и прикладных исследований. Изучению различных аспектов научных исследований библиометрическими (информетрическими) методами посвящено большое количество публикаций, в частности, в издаваемом с 1978 г. специализированном журнале Scientometrics. Конференции Science Online в последнее время также стали включать в свою программу обсуждение вопросов, относящихся к библиометрическим исследованиям научной деятельности.

Для библиометрических исследований чрезвычайно важны ретроспектива БД, тип реферируемых источников и широта их охвата, наличие контролируемой терминологии, тезаурусов, учет цитирования, возможность машинной обработки (анализа) найденных сведений, и многое другое. В частности, именно отсутствие встроенных компьютерных

программ обработки данных не позволяет пока широко использовать национальные реферативные БД ВИНТИ для изучения российской науки.

Традиционно в большинстве библиометрических исследований в области естественных наук используется мультидисциплинарная БД Science Citation Index (SCI), производимая Thomson Corporation. При изучении химических наук наряду с БД SCI часто применяется БД Chemical Abstracts (CA) производства Chemical Abstracts Service (CAS). Следует отметить, что в БД CA широко представлены не только химические, но и некоторые смежные дисциплины – биохимические, материаловедческие, физические, и др. Доступ к БД через международную научно-техническую сеть STN International позволяет использовать в работе командный язык STN Messenger, обеспечивающий эффективный многопараметровый статистический анализ больших объемов данных в режиме online.

С использованием БД CA и БД SCI и командного языка STN Messenger сотрудниками академических Центров STN в Москве и Новосибирске выполнены библиометрические исследования некоторых аспектов российской химической науки, включающие:

1) анализ научной деятельности отдельных ученых – создателей признанных научных школ, в том числе их публикаций и цитирования;

2) определение на больших периодах времени объективных наукометрических показателей некоторых академических институтов и научных центров с целью изучения продуктивности и динамики исследований, их воздействия на развитие науки, научного сотрудничества (национального и международного), и др.;

3) анализ отдельных химических дисциплин и особо актуальных направлений исследования (в частности, связанных с присуждением Нобелевских премий) с целью выявления ведущих отечественных авторов и учреждений в этих областях, наиболее цитируемых публикаций, и т.д.

4) анализ некоторых российских химических журналов для объективной характеристики их состояния и тенденций развития и выработки заключений и рекомендаций, направленных на повышение импакт-факторов (ИФ) и престижа этих изданий, в том числе вычисление ИФ тех журналов, которые не учитываются БД Journal Citation Report (JCR).

Используя полученный опыт, в настоящем докладе мы рассмотрим достоинства и недостатки БД CA в варианте, доступном через сеть STN, при библиометрическом изучении российской химической науки и смежных дисциплин.

БД СА, представленная в сети STN, реферировует > 9500 научных журналов и патенты > 57 патентных ведомств с ретроспективой до 1907 г. Сверх этого в ней учтено более 7600 статей, опубликованных в некоторых журналах American Chemical Society (ACS) до 1907 г. Кроме патентов и журналов БД СА учитывает некоторые другие типы источников – книги, труды конференций, научно-технические отчеты, диссертации. Для расширенного варианта БД CAPlus из 1500 т. н. ведущих журналов (core journals) реферировуются такие типы журнальных публикаций, как биографии, письма редактору, анонсы новостей и др. БД СА обновляется еженедельно, БД CAPlus – ежедневно.

С 1995 г. CAS перешел к реферированию англоязычных версий российских журналов. До этого практиковался охват всех изданий на оригинальном языке, а переводные версии игнорировались из-за неполноты и запоздания. В последние годы для некоторых изданий оригинал и перевод публикуются одновременно, и их содержание полностью совпадает. По этой причине, а также из-за того, что англоязычные издания распространены и архивируются более широко, чем издания на других языках, CAS предпринял селективный переход к англоязычным версиям в случаях, когда есть уверенность, что не будет потерь в полноте и своевременности материалов.

В БД СА для наиболее точной характеристики содержания реферированной публикации к имеющейся в оригинале библиографической информации (авторы, название, источник публикации и пр.) добавляется т. н. контролируемая (индексируемая) терминология (предметные заголовки, ключевые слова и т.п.). Использование контролируемой терминологии существенно облегчает поиск релевантной информации и дает возможность проводить анализ по направлениям, объектам и методам исследований. Имеющийся в БД СА тезаурус CA Lexicon позволяет использовать контролируемую терминологию для составления исчерпывающих поисковых стратегий и выявления главных (основных) концептов документов.

С 1997 г. БД СА учитывает цитируемые ссылки для ~1500 ведущих и ~5300 частично индексируемых журналов, использующих латинский алфавит, а также трудов конференций и патентов 4 патентных ведомств – США, Германии, Европейского патентного офиса (EPO) и Всемирной организации интеллектуальной собственности (WIPO). С 2003 г. в ней учтены ссылки в британских и французских патентах.

Командный язык STN Messenger (команды Select Cit и Transfer Cit) позволяет перевести все найденные в БД СА документы в формат ссылок, по которым затем возможен поиск цитирующих публикаций.

Для библиометрического анализа научной деятельности отдельных ученых были выбраны создатели признанных научных школ. Во всех случаях ретроспектива и

реферируемые БД СА типы источников позволили провести исчерпывающий отбор публикаций изученных авторов. Примечательно, что кроме журнальных статей найдены тезисы докладов на российских конференциях, отчеты, монографии, и даже такие издания, как, например, Труды комиссии по аналитической химии АН СССР, депонированные в ВИНТИ рукописи, сборники трудов молодых ученых университетов и пр. Анализ соавторов публикаций позволил охарактеризовать масштаб соответствующих научных школ.

При анализе институтов основной недостаток БД СА состоит в том, что для журнальных публикаций она приводит адрес лишь первого автора (всех авторов – только для патентов) и не унифицирует названия учреждений, что сказывается на полноте отбора публикаций. Для 6 химических институтов Сибирского Отделения РАН число найденных в БД СА публикаций составляло 54-66 % от данных самих институтов (в БД SCI – 67-80 %). Ситуация была существенно улучшена совместным применением обеих БД, позволившим довести число найденных публикаций до 78-92 %. Следует однако отметить, что в БД СА были найдены публикации в некоторых российских журналах и зарубежных продолжающихся изданиях, не учитываемых БД SCI.

При анализе отдельных научных дисциплин очень полезны рубрикатор и классификатор БД СА. Например, при изучении одного из методов электрохимического анализа (создание которого было в свое время отмечено Нобелевской премией), полученный набор документов был ограничен лишь теми, в которых название метода присутствовало в контролируемой терминологии – т. е. он был главным концептом публикации, а не, например, одним из использованных в работе рутинных методов исследования. Использование классификатора СА Lexicon позволило провести поиск по данному контролируемому термину и всем более узким терминам, входящим в его иерархию. Полученный массив публикаций (~ 32140) рассматривался как набор концептуальных документов по дисциплине. Данную процедуру можно соотнести с поиском в печатном издании Chemical Abstracts с использованием General Subject Index. При этом ретроспектива БД СА и типы реферируемых источников позволили изучить рассматриваемый метод с 1920-х гг. прошлого века, проанализировав полученный массив по типам документов, годам публикаций, языкам, странам, учреждениям, авторам, изобретателям, патентовладельцам, названиям журналов. Последнее обстоятельство позволило идентифицировать журналы, в которых наиболее часто публикуются работы по развитию метода и выявить наиболее цитируемые на рубеже XX-XXI вв. (1997-2002 гг.) публикации. Проведенный анализ позволил также идентифицировать наиболее активных в данной области авторов.

Для 3 выбранных российских химических журналов (все т.н. центральные, один с формальным статусом международного) ретроспектива БД СА и охват ею отечественных источников позволили провести полное библиометрическое исследование, начиная с первого года издания каждого журнала (1960 г., 1993 г., 2000 г.). Для журнала, издаваемого с 2000 г., использовалась также БД CAPlus, в которой на момент поиска в ноябре 2005 г. был учтен уже сентябрьский выпуск журнала. В национальной БД «РЖ Химия» (РЖХ), доступной авторам через ГПНТБ СО РАН, на то же время было зарегистрировано меньше публикаций этого журнала, относящихся к 2004-2005 гг., чем в БД CAPlus. Вместе с тем, в БД РЖХ найдены публикации, отсутствующие в БД CAPlus, такие как информация о конференциях, юбилеях, памятных датах и пр. В БД CAPlus подобные публикации учитываются только для core journals (т.н. реферирование cover-to-cover), но рассматриваемый журнал в их число не входит (в отличие от 2 других изученных).

На основе цитирования публикаций журналов проведена оценка их импакт-факторов, причем для двух впервые, так как эти издания не реферированы БД SCI и не присутствуют в БД JCR. Для журнала, издаваемого с 1960 г., полученный ИФ-2002 (0.430) сравним со стандартным из БД JCR (0.415).

В представленных выше исследованиях для исчерпывающего поиска цитирования все публикации, отобранные по тому или иному признаку (авторы, институты, журналы, дисциплины), с помощью специальных команд STN Messenger автоматически превращались в формат ссылок (например, Abakumov G A, 1972, V13, P922, или Ainshtein V G, 2000, P39). Полученные ссылки затем сопоставлялись со ссылками, хранимыми в БД, посредством процедуры Citation Matching. При этом каждая часть ссылки помещается в дополнительные поля: первый цитируемый автор – поле /RAU, журнал – поле /RWK, патент – поле /RPN, год публикации – /RPY, и др.

При поиске цитирования конкретных авторов с использованием процедуры Citation Matching отбираются т. н. корректные ссылки. Можно также, проведя поиск по реферируемому первому автору ссылки, отобрать некорректные ссылки (содержащие ошибки в цифровой части – неправильно указаны или отсутствуют номер тома или страницы – или в названии источника). Этот прием, однако, следует использовать с осторожностью и не всегда, так как в БД СА поле цитируемого автора содержит только фамилию и первый инициал. Таким образом, для авторов с распространенными фамилиями и первыми именами (инициалами) велика вероятность ошибок.

Для исследованных журналов была использована процедура Journal Title Matching – поиск всех вариантов названий цитируемых журналов в БД. В этом случае находятся

ссылки на публикации в исследуемых журналах, содержащие ошибки в авторской и цифровой частях (годе, томе, первой стр.). Для всех изученных журналов результаты цитирования их публикаций последних лет в БД СА практически совпали с аналогичными результатами БД SCI, поэтому БД СА можно использовать как для изучения цитирования журналов, так и для оценки импакт-факторов – начиная с 1997 г. (см. выше).

Сравнение – там, где было возможно – показало, что в изученных случаях количественные результаты поиска по цитированию в БД СА и БД SCI совпадали, подтверждая достоверность полученных величин. Однако, поскольку номенклатура цитирующих источников различна для БД SCI и БД СА, в общем случае целесообразно проводить поиск цитирования одновременно в обеих БД.

Примененная методология может быть легко распространена на наукометрические исследования любой естественнонаучной направленности. Для исследования физических институтов (факультетов) можно использовать БД Inspec (Physical Abstracts), биологических и биомедицинских – БД Biosis (Biological Abstracts) и БД Medline, геологических – БД Georef (Geological Abstracts), компьютерных – БД Compendex, и т.д. Возможно также совместное использование БД SCI и/или БД СА и тематических кластеров БД, доступных в сети STN International.

Работа выполнена при финансовой поддержке РГНФ (проект № 04-03-00547а).