ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [004.6:01]-021.4

В. М. Ефременкова, О. Б. Старцева, Н. Ф. Чумакова

Критерии качества библиографических баз данных

Сформулированы критерии качества библиографических БД. Рассмотрены особенности выполнения ряда критериев в ведущих политематических БД мира, представленных в сети STN International и БД ВИНИТИ.

Ключевые слова: *библиографические БД, критерии качества, БД ВИНИТИ, сеть STN International, Chemical Abstracts Service, COMPENDEX, PASCAL, INSPEC, Science Citation Index*

1. ВВЕДЕНИЕ

Понятие качества реферативных журналов (РЖ) и библиографических баз данных (БД) многоаспектно. Оно включает три основных критерия:

1) качество БД (содержательный аспект),
2) технологию генерации БД и
3) программное обеспечение.

Два последних критерия в эпоху информационных технологий тесно взаимосвязаны, поскольку и генерация БД, и выпуск печатного варианта БД — РЖ невозможны без использования программ и вычислительной техники.

Современные политематические или специализированные библиографические базы данных — одни из тех объектов в сфере информатизации, от которых требуется высокое качество и наличие возможности его оценки.

В настоящее время имеются лишь стандарты, характеризующие качество программного обеспечения БД. Это:

ISO 9126 (ГОСТ Р ИСО/МЭК 9126-93) — "Информационная технология. Оценка программного продукта. Характеристики качества и руководство по их применению".

ISO/IEC 12207 — "Информационные технологии. Процессы жизненного цикла программного обеспечения".

Проблема стандартизации характеристик качества по их применению становится все более актуальной в связи с быстрым изменением информационных технологий обработки документов: появлением электронных документов, введением новых полей для отражения таких характеристик, как индекс цитирования научных работ, импакт-фактор журнала, URL журнала и др. Не менее важной характеристикой качества информационных массивов является наличие критериев содержательной стороны, определяющее рейтинг РЖ и/или БД на мировом уровне.

2. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ КАЧЕСТВА БИБЛИОГРАФИЧЕСКИХ БД

Основными характеристиками реферативных БД, определяющими их качество на основе рейтингов на мировом уровне, являются следующие очевидные показатели:

1) предметная область (области знания);
2) глубина ретрофонда;
3) источники формирования (получение первичной информации и ее виды);
4) полнота отражения первоисточников;
5) актуальность данных;
6) достоверность информации (научный характер и рецензирование отражаемых первоисточников);
7) периодичность;
8) оперативность данных;
9) структурированность документов (наличие классификаторов, тезауруса, описания поисковых полей);
10) однократное отображение публикации с многоаспектными результатами аналитико-синтетической переработки документа, включая текст реферата;
11) целостность (генерация единой БД и возможность формирования отдельных фрагментов — электронных и/или печатных выпусков БД по заказам пользователей);
12) доступность описания БД (статистические характеристики документального информационного потока, списки журналов, конференций и т. д.) на сайте информационного центра — генератора БД;
13) наличие программных средств формирования РЖ, системы указателей к ним и других информационных продуктов;
14) присутствие программных средств, обеспечивающих возможность проведения наукометрии (режима анализа по различным поисковым полям);
15) доступность внешнему пользователю (возможность поиска в сети, что обеспечивает возможность использования режима кросс-поиска или обращение только к одной БД в зависимости от экономической выгоды для пользователя);
16) удобство пользовательского интерфейса;
17) визуализация результатов поиска;
18) многоязычный интерфейс.

2.1. Предметная область (области знания) и глубина ретрофондов

Развитие реферативных служб для обеспечения информацией научных исследований и практических работ по применению достижений науки, техники и технологий началось с выхода первого научного журнала "Le journal des scavans" 5 января

1665 г. в Париже (Франция). Это было первое реферативное издание, опубликовавшее краткие географические обзоры по региональной геологии. Но этот журнал так и остался научным журналом с гуманитарным профилем, к которому всегда относилась и география.

Потребность в информационно-справочной литературе наиболее остро стала ощущаться в XIX в. В области химии и химической технологии уже в 1817–1889 гг. существовало семь справочных изданий и каталогов.

Истоком создания американской реферативной службы Chemical Abstracts Service (CAS) явился один из ведущих журналов по химии — "Journal of American Chemical Society", в котором с 1893 г. публиковались обзоры изданных в США книг, а уже в 1885 г. был введен раздел "Обзор американских исследований по химии". С 1907 г. раздел стал самостоятельным реферативным изданием — Chemical Abstracts, ретрофонд которого к 2007 г. достиг 26 млн записей.

Для удовлетворения потребностей специалистов геологических специальностей Американским геологическим институтом в 1785 г. был создан РЖ Geological Reference (GeoRef).

В СССР в 1952 г. по заказу двух организаций — Академии наук СССР и Государственного комитета Совета министров по науке и технике — был образован Институт научной информации, впоследствии Институт научной и технической информации (ВИНИТИ). Поэтому в РЖ ВИНИТИ были представлены и фундаментальная наука, и технические дисциплины. В 1953 г. были выпущены первые номера РЖ по астрономии, математике, механике, химии, а с 1955 г. начали выходить в свет РЖ и по всем научно-техническим направлениям и информатике.

В XX в. с появлением вычислительной техники было создано несколько информационных центров, генерирующих БД различной тематической направленности и имеющих разные цели и задачи по информационному библиографическому обслуживанию ученых и специалистов: 1969 г. — политематическая БД INSPEC (физика, автоматика и вычислительная техника, электроника и электротехника, информатика); 1970 г. — политематическая БД COMPENDEX (технические дисциплины); 1974 г. — политематическая БД Science Citation Index; 1981 г. — политематическая БД ВИНИТИ РАН; 1984 г. — политематическая БД PASCAL (Франция); 1985 г. — политематическая БД JICST.

В последние годы большое внимание уделяется увеличению глубины ретроспективы БД, идет интенсивная оцифровка реферативных изданий. Например, массив реферативного журнала GeoRef оцифрован с 1785 г., INSPEC — с 1898 г., COMPENDEX — с 1884 г. Увеличение полноты отражения публикаций в БД идет не только путем добавления ретромассивов, но и включением не дублирующихся документов по интересующим информационную службу тематикам из других специализированных БД. Например, CAS провел успешные переговоры с FIZ *CHEMIE* (Германия) о включении документального потока, не представленного до сих пор в CAS.

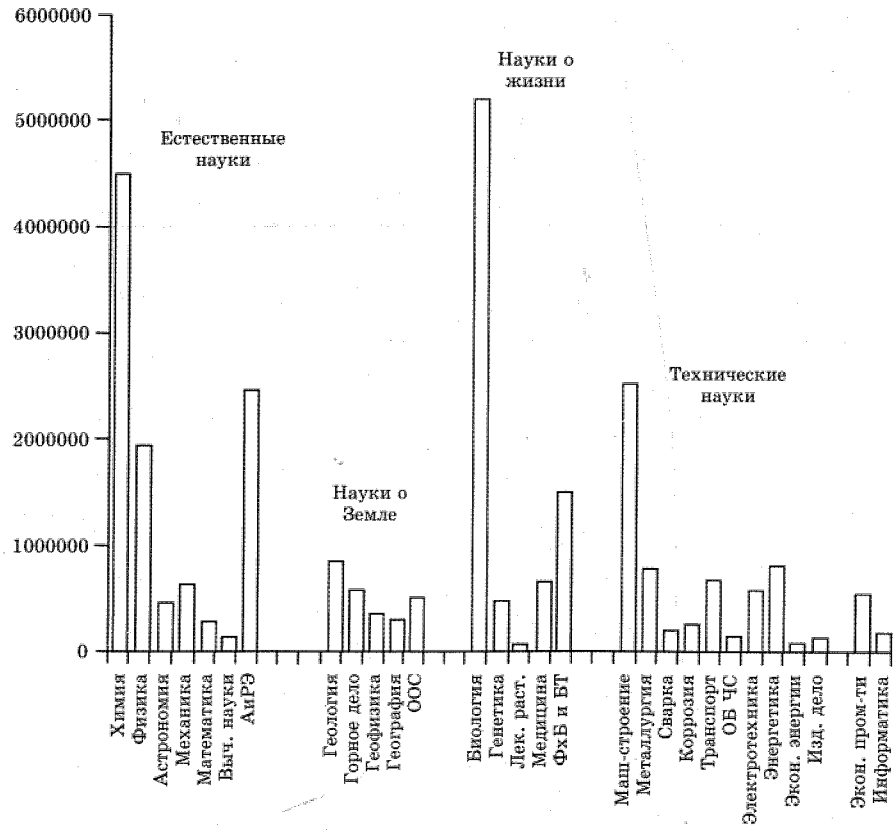

Рис. 1. Распределение суммарных потоков документов в соответствии с тематическим содержанием банка данных ВИНИТИ

Ни одно известное в мире крупное реферативное издание и/или БД не является полным аналогом РЖ/БД, формируемых другими информационными центрами. Это происходит, с одной стороны, из-за различий в выборе освещаемой тематики, определяемой различными классификационными схемами, с другой стороны, из-за различий в принципах формирования БД и/или РЖ. Каждый информационный центр разрабатывает свои подходы к отбору отражаемых первоисточников, языкам и странам-создателям (издателям) документов. Так как каждая информационная служба определяет основные характеристики документальных информационных потоков, то очевидно, что все вышеперечисленные БД ведущих стран мира имеют свои приоритеты в отражении, прежде всего, национальной литературы, в выборе тематики, видов первоисточников и полноты представления их в своем информационном массиве.

В качестве примера рассмотрим БД ВИНИТИ и Chemical Abstracts (рис. 1 и 2).

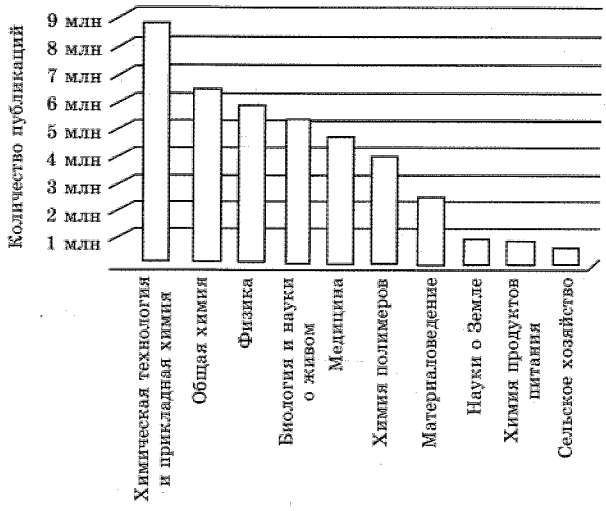

Рис. 2. Распределение потока публикаций по отраслям знания в БД Chemical Abstracts (2007 CAS Catalog.www.cas.org)

Тематически весь универсум знания, отраженного в БД ВИНИТИ, можно сгруппировать в ряд научных сфер:

— естественные науки;
— науки о Земле;
— науки о живом;
— технические науки;
— информация, организация, управление.

На рис. 1 представлено распределение суммарных потоков документов в соответствии с тематическим содержанием банка данных ВИНИТИ.

На рис. 2 показано распределение суммарных потоков публикаций по отраслям знания в БД Chemical Abstracts, представленное аналитиками службы CAS, из которого видно, что Chemical Abstracts является мультидисциплинарной БД.

По тематическому содержанию БД ВИНИТИ и Chemical Abstracts, как показали работы по сопоставлению классификаторов этих БД, достаточно близки. Главное отличие: в БД Chemical Abstracts не отражаются проблемы "чистой математики" и слабее представлены такие дисциплины, как общая биология (в основном представлены разделы биохимии), машиностроение (кроме химического), отдельные вопросы транспорта, издательского дела, информатики и вычислительной техники (но не вычислительной математики). Однако в Chemical Abstracts больше полнота отражения первоисточников, особенно это относится к патентным документам.

Ретрофонды БД Chemical Abstracts значительно больше, чем БД ВИНИТИ, из-за временной разницы возникновения информационных центров и оцифровки ретромассивов с 1907 г.

2.2. Источники формирования (получение первичной информации и ее виды)

Пользователям и брокерам, осуществляющим поиск в БД, очень важны сведения об источниках отражаемой литературы, ее видах. Следует принимать во внимание, что наибольшее количество публикаций во всех политематических и специализированных БД представлено статьями из сериальных изданий (журналов). При этом во всех БД, близких по тематике, отражается один и тот же блок наиболее важных журналов (так называемые ядерные журналы — key journal), имеющих по данным службы SCI большие значения импакт-фактора, т. е. имеет место значительное дублирование отражаемых документов. Полнота отражения трудов конференций невелика в силу специфики их подготовки к печати: перед конференцией — в специальных сборниках трудов, после проведения конференции — в выбранном по соответствию тематике конференции журнале. В этом случае оперативность значительно ниже. Наиболее "полно" труды конференций представлены в БД INSPEC и COMPENDEX. Патентные документы отражаются лишь в БД ВИНИТИ и Chemical Abstracts, что связано с широтой и спецификой тематической направленности обеих БД (см. таблицу). Электронные документы, не имеющие печатного аналога (on-line файлы, мультимедиа, оптические диски), представлены только в БД Chemical Abstracts.

Распределение документального информационного потока по основным типам документов в ведущих политематических БД мира в 2007 г. (%)

Типы документов / Базы данных	Статьи из журналов	Препринты	Труды конференций, сборники	Книги	Патенты	Диссертации	Стандарты	Отчеты	Обзоры	Карты, атласы	Электронные ресурсы (on-line файлы, мультимедиа, оптические диски)
ВИНИТИ	72,6		12,1	2,0	11,2	1,6	0,01			0,01	
PASCAL	99,2		12,0	0,5		0,3		0,03	0,3	0,005	
COMPENDEX	65,5		34,2						1,9		
CAS	56,5	2,5	3,8	0,3	38,4	1,4	2,5	0,001	7,8		2,4
JICST	87,5	0,1	12,5				0,01				
INSPEC	70,9		32,5	0,04			0,01	0,001	0,1		
SCI	89,7		8,3						2,2		

Еще одна особенность представления документов в БД — библиография с ключевыми словами, но без рефератов. Доля таких документов во всех политематических БД ведущих стран мира колеблется от 0,5% до 3%. Исключением является БД CAPlus, в которой, помимо ежегодного массива реферативной БД Chemical Abstracts, содержится в последние годы до 40% документов, представленных только библиографией. Одним из видов такого типа документов являются выпуски сообщений ежегодных собраний Американского химического общества.

2.3. Полнота отражения первоисточников, актуальность и достоверность данных

Оценка полноты отражения "мирового потока" документов по области знания, конкретной дисциплине или узкотематическому направлению ранее проводилась экспертами путем количественного сравнения выдачи документов по запросам в БД с количеством документов в печатном аналоге (РЖ), адекватных потребностям пользователя. Введение новых компьютерных технологий позволяет программно оценить полноту выдачи после поиска в нескольких БД сети STN International, используя команду "Duplicate" для удаления из выдачи дублирующихся в разных БД документов (команда работает в массиве не более 5 тыс. записей). Таким образом, можно оценить полный поток публикаций по выбранной пользователем тематике. В настоящее время в области материаловедения, нанотехнологий и биотехнологий лидирующей по полноте отражения документов является БД CAPlus.

Параметр актуальности определяется востребованностью тематики в рассматриваемый период времени, прежде всего это приоритетные направления развития науки, техники и технологий. Наиболее важными параметрами при этом являются оперативность, полнота и достоверность (определяемая научным характером первоисточников, включая и рецензирование публикаций в них).

2.4. Оперативность данных

Оперативность отражения информации определяется периодичностью ее поступления в информационный центр и периодичностью обновления БД. В информационной практике встречается ежедневное обновление массива — как в БД Chemical Abstracts, еженедельное — как в БД COMPENDEX, PASCAL, INSPEC, JICST и ежемесячное — как в БД ВИНИТИ, METADEX.

Кроме того, наиболее оперативно отражаются публикации, освещающие приоритетные направления и наукоемкие технологии в соответствии с формируемыми списками первоисточников.

В ряде областей знания, таких, как математика, науки о Земле, информатика и т. д., актуальность информации сохраняется независимо от сроков ее появления в БД.

Наиболее оперативно информация поступает в те БД, генераторами которых являются крупнейшие научные издательства, например Elsevier (БД SCOPUS). В этом случае информация попадает в БД сразу после рецензирования публикации, но до выхода журнала в свет (раньше на 1–2 месяца).

2.5. Структурированность БД

Структурированность БД определяется наличием классификатора, тезауруса и набором поисковых полей. Каждая БД имеет свой линейный или иерархический (с различной глубиной уровней) классификатор.

Линейная классификация применяется в политематической БД SCI и специализированных БД GEOREF, METADEX.

Двухуровневый классификатор, имеющий 80 кодов первого уровня, используется в мультидисциплинарной БД Chemical Abstracts; при этом он обеспечивает максимальную точность поиска. В этой БД создан тезаурус по классификатору, отражающий временны́е характеристики изменения и связь с терминами контролируемой лексики. Не менее удобен и двухуровневый классификатор БД технической направленности COMPENDEX.

5–6-уровневые классификаторы имеют политематические БД PASCAL, INSPEC, JICST.

4–9-уровневый классификатор используется в БД ВИНИТИ.

Все зарубежные БД ведущих стран мира имеют тезаурусы.

Помимо основных поисковых полей (вид документа, поля библиографического описания, языка и страны) в БД INSPEC, COMPENDEX и Chemical Abstracts существуют указатели роли документа (теория, эксперимент, применение, экономические или правовые аспекты, менеджмент и т. д.)

Возможность представления многоаспектного содержания публикаций при однократном их отражении осуществляется либо введением второго поля, в которое помещается информация о смежных направлениях (в БД Chemical Abstracts), либо в одном поле "код классификатора" проставляется несколько кодов в соответствии с отражаемыми в документе тематическими направлениями.

2.6. Доступность и полнота описания характеристик БД

Для пользователей очень важно содержание сайта информационной службы. В настоящее время наиболее информативным является сайт CAS, состоящий из двух частей и отвечающий на следующие вопросы: 1) что и как можно найти в пяти БД CAS и 2) какую информацию по интересующей потребителя теме можно еще получить из 200 БД, входящих в сеть STN International. На сайте можно увидеть статистические характеристики отражаемого документального информационного потока с 1907 г. по настоящее время, динамику и ретрофонды суммарного массива и потоков основных видов документов, их распределения по основным странам и языкам. Приведен список профильных журналов. Освещаются различные режимы поиска, позволяющие получать наиболее полную и точную выдачу документов по запросам пользователей.

2.7. Визуализация результатов поиска информации

Проводимые наукометрические исследования с использованием программных средств информационных служб Science Citation Index и CAS дают

возможность предварительной оценки состояния работ в рассматриваемой области, получения списков журналов, в которых публикуются преимущественно результаты интересующих специалистов исследований, списков организаций, в которых проводятся подобные исследования.

Наиболее интересную и подробную информацию можно получить с помощью разработанного в CAS нового модуля AnaVist, позволяющего получать в процессе поиска карты выбранного научного направления, списки организаций, проводить анализ патентной литературы (для тематик, связанных с прикладными работами) и т. д.

ЗАКЛЮЧЕНИЕ

Предложенный в работе перечень критериев качества библиографических БД, отражающих их содержательный аспект, позволит пользователям различного уровня (от студентов до ученых и специалистов-практиков) выбирать те информационные ресурсы, которые наиболее полно отражают их потребности. Анализ результатов поиска может в дальнейшем выявить новые направления практического применения и "подсказать" пути создания конкурентоспособной инновационной продукции. Рассмотренные в данной статье содержательные критерии качества БД могут быть применимы и при работе с библиотечными массивами.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. www.cas.org
2. http://www.georef.org/
3. http://www.viniti.ru
4. http://scientific.thomson.com/products/sci/
5. www.iee.org/inspec/
6. www.inist.fr
7. www.engineeringvillage.com
8. Хуторецкий В. М., Ефременкова В. М. Русскоязычная химическая информация в Chemical Abstracts Service и ВИНИТИ // Изв. РАН. Сер. Химия.— 2000.— № 1.— С. 183–187; Russian Chem. Bull.— 2000.— Vol. 49, № 1.— P. 185–190.
9. Ефременкова В. М., Каменская М. А., Хуторецкий В. М. Соотношение классификационных схем базы данных Chemical Abstracts и соответствующих ей частей системы баз данных ВИНИТИ. Ч. 1. Биологическая химия // НТИ. Сер. 1.— 1999.— № 12.— С. 20–34.
10. Хуторецкий В. М., Ефременкова В. М., Тартаковский В. А. Соотношение классификационных схем базы данных Chemical Abstracts и соответствующих ей частей системы баз данных ВИНИТИ. Ч. 2. Органическая химия // НТИ. Сер. 1.— 2000.— № 2.— С. 26–30.
11. Ефременкова В. М., Лялюшко Н. С., Хуторецкий В. М. Соотношение классификационных схем базы данных Chemical Abstracts и соответствующих ей частей системы баз данных ВИНИТИ. Ч. 3. Макромолекулярная химия // НТИ. Сер. 1.— 2000.— № 5.— С. 27–38.
12. Ефременкова В. М., Нестерова Е. Н., Хуторецкий В. М. Соотношение классификационных схем базы данных Chemical Abstracts и соответствующих ей частей системы баз данных ВИНИТИ. Ч. 4. Прикладная химия и химическая технология // НТИ. Сер. 1.— 2000.— № 7.— С. 29–45.
13. Ефременкова В. М., Хуторецкий В. М. Соотношение классификационных схем базы данных Chemical Abstracts и соответствующих ей частей системы баз данных ВИНИТИ. Ч. 5. Физическая, неорганическая и аналитическая химия // НТИ. Сер. 1.— 1999.— № 12.— С. 20–34.
14. Ефременкова В. М., Круковская Н. В., Якимов В. И. Публикации по фуллеренам в зеркале баз данных мира // НТИ. Сер. 1.— 2005.— № 8.— С. 20–38.
15. Sirovsky F. S., Krukovskaya N. V., Efremenkova V. M. Proc. 6th World Multiconference on Systemics, Cybernetics and Informatics (July 14–18, 2002, Orlando, Florida, USA).— 2002.— Vol. 17.— P. 106–110.
16. Ефременкова В. М., Севастьянов В. Г. Информационное сопровождение научных исследований по карбиду кремния. Ч. 1. История. Проблемы оптимизации поиска информации в системе баз данных STN International. Информационные ресурсы // НТИ. Сер. 1.— 2004.— № 9.— С. 16–27.

Материал поступил в редакцию 25.12.08.