Полидоменные модели электронных библиотек систем мониторинга сферы науки [1)]

© И.М. Зацман

Институт проблем информатики РАН
iz@a170.ipi.ac.ru

Аннотация

Доклад посвящен проблеме моделирования электронных библиотек систем мониторинга сферы науки, ресурсы которых используются одновременно для решения информационно-справочных и аналитических задач. В системах мониторинга решение аналитических задач, вытекающих из требований новой модели распределения бюджетных средств [1], связано с вычислениями широкого спектра индикаторов результативности и эффективности в сфере науки. Вычисление индикаторов предъявляет ряд новых требований к проектированию электронных библиотек. Для учета на этапе эскизного проектирования этих требований, включая требования к схемам информационных ресурсов электронных библиотек систем мониторинга, предлагается использовать полидоменные модели, которые являются развитием логико-лингвистических моделей управления [2].

1 Введение

В работе [3] введено понятие полидоменных моделей, включающих лексико-семантический, информационный, алгоритмический, математический и другие компоненты. Подобные сочетания перечисленных компонентов предназначены для моделирования электронных библиотек и других видов информационных систем, ресурсы которых используются одновременно для решения информационно-справочных и аналитических задач в слабоформализуемых и институционально сложных сферах применения, например, в сфере науки, для мониторинга правового пространства и правоприменительной практики, в сфере инноваций, в патентной сфере.

Полидоменные модели являются развитием логико-лингвистических моделей управления [2], ориентированным на моделирование сложных институциональных систем от наноуровня до макроуровня [4]. Например, для сферы науки, как институциональной системы, предметом исследования на наноуровне является деятельность конкретных ученых и специалистов, на микроуровне – научных лабораторий или неформальных научных коллективов, например, выполняющих проекты по грантам РФФИ или РГНФ, на мезоуровне – научных институтов и других учреждений сферы науки, а на макроуровне предметом исследования может быть РАН или вся сфера науки в целом.

Основная цель построения и использования полидоменных моделей электронных библиотек систем мониторинга заключается в декларативной и процедурной экспликации связей между векторными функциями, задающими весь спектр определяемых системой индикаторов, и информационными ресурсами, которые необходимы для вычисления векторных функций в некоторой заданной области определения. В более общем случае, полидоменные модели позволяют учитывать изменение и/или пополнение во времени информационных ресурсов электронных библиотек, а также эксплицировать изменение во времени области определения векторных функций.

Полидоменные модели электронных библиотек систем мониторинга включают следующие компоненты:

1) спецификация символьно-знаковой системы электронной библиотеки,
2) информационный компонент в форме набора схем информационных ресурсов электронных библиотек (в общем случае, полидоменные модели позволяют эксплицировать изменение схем во времени);
3) математический компонент в виде множества векторных функций, используемых для вычисления значений индикаторов;
4) алгоритмический компонент, необходимый для определения значений параметров векторных функций на основе информационных ресурсов электронных библиотек;

Труды 8[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

[1)] Работа выполнена при частичной поддержке РГНФ, грант № 05-03-12328в.

5) информационно-математический компонент, определяющий связи между параметрами векторных функций и теми информационными ресурсами, которые необходимы для определения значений параметров;
6) лексико-семантический компонент, включая используемые классификаторы, тезаурусы, системы онтологий (в общем случае, в модели может отслеживаться их изменение в зависимости от времени);
7) аналитический компонент в форме набора критериев, построенных на основе индикаторов, методики построения критериев и правил их применения пользователями систем мониторинга.

Предлагаемый подход к интеграции перечисленных компонентов в рамках единой модели, называемой полидоменной, позволяет описать сочетания конкретных знаковых и абстрактных формально-символьных объектов и структур, например, математические и химические формулы, корпуса текстов на естественных языках, вербально-образные тезаурусы и онтологии, используя явно определенную символьно-знаковую систему электронной библиотеки. В рамках полидоменной модели различаются три аспекта описания знаковых и формально-символьных объектов и структур: синтаксический, семантический и прагматический.

Основная цель доклада заключается в том, чтобы продемонстрировать возможности полидоменных моделей в процессе эскизного проектирования новых или модернизации существующих электронных библиотек, информационные ресурсы которых используются для мониторинга сферы науки.

2 Методология построения моделей

Ранее отмечалось, что полидоменные модели являются развитием логико-лингвистических моделей управления, но ничего не было сказано о методологии их развития. В качестве методологической основы построения полидоменных моделей использовался подход, сформулированный С. Горном следующим образом: одно из главных интуитивных представлений специалистов в сфере информатики заключается в том, что любой процесс, который можно точно определить, может быть смоделирован в символьно-знаковой форме, поскольку точная спецификация процесса уже является некоторой разновидностью символьно-знакового моделирования этого процесса [5, с. 132].

Далее С. Горн пишет о том, что символьно-знаковые объекты и структуры не должны обязательно состоять только из чисел; в зависимости от предметной области они могут представлять собой:

- аналитические выражения в математике и физике,
- одно-, двух- и трехмерные структурные формулы в химии,
- структурно-графические представления полипептидных цепей и двойной спирали ДНК,
- партитуры симфоний,
- обозначения балетных движений по Р. Лабану с помощью системы хореографических символов,
- спецификации схем микропроцессоров.

После перечисления столь разных категорий символьно-знаковых объектов и структур С. Горн отмечает, что специалисты из большинства сфер деятельности согласны с тем, информатика обладает полезными для них возможностями, несмотря на возможные смысловые различия между их пониманием своих символьно-знаковых объектов и пониманием коллег из области информатики [5, с. 133].

Таким образом, естественным следствие обозначенного методологического подхода к построению полидоменных моделей является включение в их состав первого компонента - спецификации символьно-знаковой системы электронной библиотеки, позволяющей описывать конкретные знаковые и абстрактные формально-символьные объекты и структуры, а также процессы соответствующей предметной области или сферы деятельности.

С. Горн фиксирует три аспекта описания знаковых и формально-символьных объектов и структур следующим образом: «Информатика должна соотносить себя именно с **прагматическими вопросами** (выделено мной – *И.З.*), от которых, как мы уже убедились, не должны зависеть математика и физические науки в той части, которая касается если не методов, то результатов. В этом отношении информатика имеет большее сходство с лингвистикой, психологией, бихевиористическими науками, философией и различными профессиями, имеющими к ним отношение». Далее С. Горн пишет, что изучение отношений между символьно-знаковыми объектами и операций с ними, независимое от их смыслового содержания или прагматического контекста, называется синтактикой. В изучении вопросов синтактики и **синтаксических описаний** информатика является наиболее формальной дисциплиной, тесно связанной с математикой и ее методами. Изучение отношений между символьно-знаковыми объектами и их значениями (содержанием), независимое от целей и способов их использования, то есть с исключением из рассмотрения прагматических вопросов, называется **семантикой** [5, с. 132].

Таким образом, следствием и развитием обозначенного методологического подхода является включение в состав полидоменной модели электронной библиотеки системы мониторинга информационного, математического, алгоритмического и информационно-математического компонентов в качестве синтаксической составляющей модели, лексико-семантического компонента в

качестве семантической составляющей и аналитического компонента в качестве прагматической составляющей модели.

3 Индикаторы и критерии результативности и эффективности

При моделировании электронных библиотек, из трех аспектов описания знаковых и формально-символьных объектов и структур – синтаксического, семантического и прагматического – довольно редко учитывается последний, прагматический аспект описания. Основная задача этого аспекта моделирования электронных библиотек систем мониторинга заключается в символьно-знаковом описании целей и способов использования индикаторов, процессов смыслового понимания индикаторов пользователями систем мониторинга, а также понимания целей и способов их применения.

В настоящее время существенно возрасла роль именно прагматического аспекта моделирования электронных библиотек систем мониторинга, предназначенных для анализа и оценки деятельности субъектов сферы науки, так как новая модель бюджетного процесса в стране предусматривает использование систем мониторинга во время распределения бюджетных средств. В рамках старой модели бюджетного процесса, системы мониторинга и значения индикаторов результативности и эффективности, определяемые на основе их информационных ресурсов, практически не влияли на бюджетный процесс [6, 7, 8, 9, 10].

Однако сейчас планируется, что через нескольких лет все 100% бюджета науки будут распределяться по целевым программам с использованием индикаторных оценок их результативности и эффективности [11]. Подобное распределение бюджетных средств существенно изменяет статус электронных библиотек, информационные ресурсы которых используются для мониторинга сферы науки.

Изменение статуса влечет изменение требований к проектированию новых или модернизации существующих электронных библиотек. Поэтому проектирование электронных библиотек, учитывающее в явном виде прагматические аспекты их использования, становится весьма актуальной задачей.

На рис. 1 приведены определения основных категорий индикаторов и их связи с другими показателями, которые планируется вычислять и использовать в процессе распределения бюджетных средств, включая вычисление критериев принятия решений [1]. На этом рисунке перечислены шесть основных целей использования критериев в рамках новой модели бюджетного процесса, а именно:

- отбор научно-исследовательских программ для бюджетного финансирования,
- преобразование программ,
- сокращение бюджетного финансирования или прекращение выполнения программ,
- распределение бюджетных средств по программам,
- оценка федеральных и ведомственных целевых программ сферы науки,
- оценка деятельности администраторов бюджетных средств, финансирующих программы в сфере науки.

Основное различие между индикаторами и критериями заключается в том, что индикаторы вычисляются только на основе информационных ресурсов электронных библиотек, а критерии принятия решений в рамках новой модели бюджетного процесса в общем случае зависят и от индикаторов, и от параметров, задаваемых извне, а не вычисляемых на основе информационных ресурсов. Таким образом, отражение в аналитическом компоненте полидоменной модели функциональных отношений между критериями и индикаторами, которые обозначены на рис. 1 двойной фигурной стрелкой, позволяет учитывать прагматический аспект использования электронных библиотек систем мониторинга в сфере науки.

Наиболее сложной задачей прагматического аспекта моделирования электронных библиотек является экспликация понимания индикаторов и критериев, а также целей и способов их применения теми пользователями систем мониторинга, которые принимают решения о финансировании программ в сфере науки.

Отчет семинара по оценке результативности федеральных научно-исследовательских программ США, состоявшемся 4-5 декабря 2003 года, содержит перечень актуальных вопросов, среди которых есть вопрос согласованного понимания (like-minded) индикаторов и критериев, а также целей и способов их применения пользователями [12].

В докладе эта задача не рассматривается. Здесь она упомянута, чтобы подчеркнуть сложность и актуальность учета прагматического аспекта моделирования электронных библиотек. Следствием сложности учета прагматического аспекта является необходимость применения сложного аппарата полидоменных моделей. Однако их использование позволяет получить на этапе эскизного проектирования описание электронных библиотек, инвариантное к используемым программным и аппаратным средствам.

Естественно, что также сохраняется актуальность синтаксического и семантического аспектов моделирования, включая моделирование информационного представления научных результатов в электронных библиотеках. В электронных библиотеках систем мониторинга научные результаты отражаются следующим образом:

1) как формы эксплицитного представления научных результатов (тексты статей, монографий, диссертаций и т.д.);

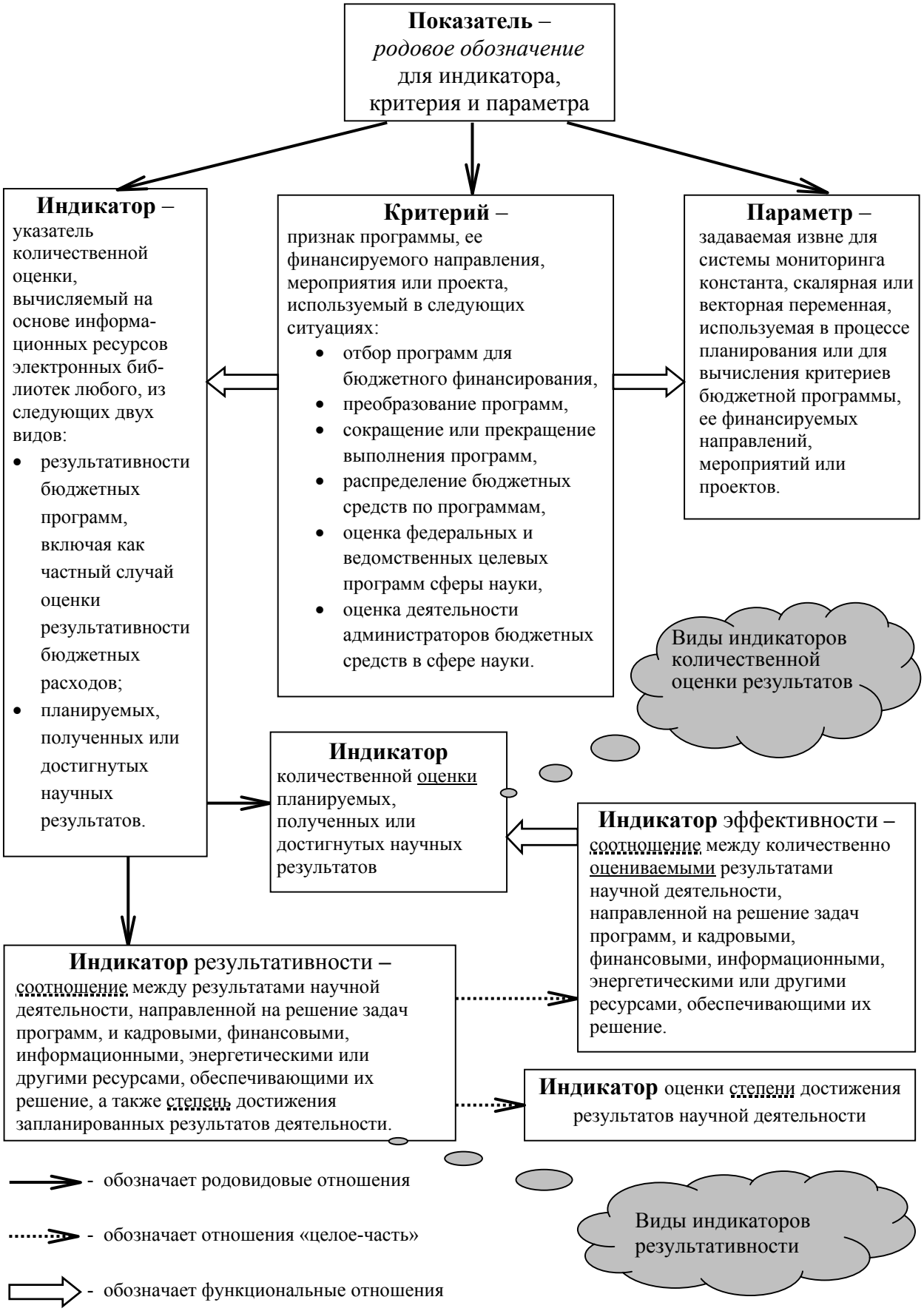

Рис. 1. Основные категории индикаторов и их связи с другими показателями сферы науки: критериями и параметрами

2) как оценка прогнозируемого влияния научных результатов на решение задач бюджетных научных программ или их прогнозируемое использование при получении новых научных результатов;
3) как оценка состоявшего влияния или использования научных результатов.

Примером индикатора, для определения которого используется данные по опубликованным статьям, может служить таблица 1, содержащая число научных статей, опубликованных в течение года учеными перечисленных в таблице стран в журналах, обрабатываемых Институтом научной информации (США), в расчете на 1 млн жителей этой страны [9, vol. 1, p. 5-40]. Примером оценки прогнозируемого влияния результатов фундаментальных исследований может служить таблица со средневзвешенными импакт-факторами всех статей, опубликованных исполнителями проектов РФФИ, приведенная в работе [6, 13].

Примером оценки состоявшего использования научных результатов может служить таблица 2, содержащая значения двух индикаторов цитируемости ученых разных стран (регионов), характеризующих влияние на развитие научной сферы: число ссылок на ученых страны (региона) и доля этих ссылок в процентах от их общего числа в 1992, 1996, 2001 и 2003 гг. [9, 14].

4 Примеры использования полидоменных моделей

В докладе приведены два примера, иллюстрирующие возможные области применения полидоменных моделей. Примеры ограничены четырьмя компонентами: информационный компонент в форме набора схем информационных ресурсов, алгоритмический, математический и информационно-математический компоненты.

Пример вычисления индикаторов иллюстрирует возможность моделирования электронной библиотеки «Указатель РФФИ», экспериментальную эксплуатацию которой планируется начать в конце 2006 года. В настоящее время ее информационные ресурсы позволяют определять индикаторы возрастной структуры заявителей РФФИ и ряд других индикаторов, включая импакт-факторы статей, опубликованных по проектам РФФИ [6, 7, 13]. В докладе приводятся значения индикатора возрастной структуры заявителей из РАН и остальных заявителей. Такой индикатор можно записать как векторную функцию $\mathbf{f}=(f_1, f_2)$, где f_1 обозначает индикатор для заявителей РАН, а f_2 – индикатор для всех остальных заявителей.

Векторная функция **f** зависит от дискретной переменной возраста заявителей t_i, где $t_0=16$, $t_1=17$ и так далее до 101 года и дискретной переменной, фиксирующей год определения возрастной структуры заявителей РФФИ T_j, где $T_0=1997$, $T_1=2002$, то есть в пример включены только два года определения возрастной структуры. В краткой математической форме этот индикатор возрастной структуры исследователей имеет вид:

$\mathbf{f}(t_i, T_j)$, *где* t_i и T_j *являются дискретными переменными*, i, j=0,1, … ; $\mathbf{f}=(f_1, f_2)$.

Например, в 1997 году общее число исследователей РАН, подавших заявки в РФФИ, было равно 21804 человека. Из них 59 человек достигли 20-летнего возраста. Следовательно, f_1 (20, 1997)=59:21804*100=0,27%, т.е. доля 20-летних исследователей РАН, подавших заявки в РФФИ в 1997 году составила 0,27% от общего числа исследователей РАН, подавших заявки в РФФИ в этом году. Однако подобное математическое описание не раскрывает алгоритма, с помощью которого было получено число 59, которое является необходимой величиной для вычисления значения функции f_1 (20, 1997) в математическом компоненте полидоменной модели.

Это число определяется в результате поиска в электронной библиотеке записей персональных данных, имеющих атрибут Year, который содержит последовательность четырех цифр «1977» при условии, что в соответствующей записи реквизитов организаций, являющихся местом работы исследователя, атрибут Ministry содержит последовательность литер «РАН». Число найденных в электронной библиотеке записей равно 59. Это число является результатом поиска, а не результатом вычислений.

Приведенный пример иллюстрирует совместное использование информационного, математического и алгоритмического компонентов. Последний должен включать формальную запись всех тех алгоритмов, которые используются для вычисления индикатора возрастной структуры исследователей. Отметим, что информационно-математический компонент должен связывать системой явных ссылок описания векторной функции, схем информационных ресурсов и алгоритмов.

В докладе, кроме рассмотрения этого примера, решается задача добавления нового индикатора публикационной активности исполнителей проектов РФФИ каждого возраста. Показано, что при существующей схеме информационных ресурсов этот индикатор можно определить только в результате выполнения многошаговой процедуры сопоставления данных о каждом авторе публикаций в форме 509 «Публикации по результатам года» и персональных данные формы 512 «Данные о руководителе и основных исполнителях» по отчетам РФФИ. При изменении схемы информационных ресурсов отпадает необходимость в выполнении многошаговой процедуры сопоставления данных этих форм. Для вычисления нового индикатора, кроме изменения схемы, то есть изменения информационного компонента, необходимо также внести изменения в алгоритмический, математический и информационно-математический компоненты полидоменной модели, модификация которых также рассматривается в докладе.

Таблица 1. Число статей, опубликованных в течение одного года, на 1 млн жителей в период 1999-2001 гг.

Страна	Число статей/1 млн	Страна	Число статей/1 млн
Швейцария	1165,0	Португалия	191,3
Швеция	1139,3	Польша	139,9
Израиль	1055,2	Россия	116,4
Финляндия	960,5	Все страны	108,8
Дания	932,2	Болгария	103,7
Великобритания	821,9	Аргентина	77,8
Голландия	800,5	Чили	75,7
Австралия	794,2	Южная Африка	55,8
США	722,2	Бразилия	38,8
Норвегия	720,0	Ливан	37,3
Сингапур	590,3	Мексика	31,8
Франция	538,6	Египет	23,2
Германия	530,5	Коста Рика	22,8
Япония	445,6	Малайзия	21,9
Ирландия	429,9	Китай	14,8
Испания	382,7	Иран	13,6
Италия	371,4	Таиланд	10,8
Тайвань	330,3	Индия	10,8
Чешская Республика	241,4	Кения	8,6
Южная Корея	206,8	Гватемала	1,5

Таблица 2. Индикаторы цитируемости ученых разных стран и регионов в 1992, 1996, 2001 и 2003 годах

	1992		1996		2001		2003	
Регион/страна	Число	%	Число	%	Число	%	Число	%
Все страны	2 684 777	100,00	3 325 455	100,00	3 846 519	100,00	4339511	100,00
США	1 389 314	51,75	1 624 607	48,85	1 678 293	43,63	1839481	42,39
Германия	157 285	5,86	207 673	6,24	274 520	7,14	305555	7,04
Китай	9 910	0,37	16 539	0,50	33 245	0,86	65326	1,51
Индия	14 421	0,54	19 250	0,58	24 442	0,64	31534	0,73
Южная Корея	2 260	0,08	6 563	0,20	23 551	0,61	40726	0,94
Венгрия	4 755	0,18	5 988	0,18	8 345	0,22	9714	0,22
Польша	8 671	0,32	10 692	0,32	14 909	0,39	18672	0,43
Россия	0	0,00	19 047	0,57	31 602	0,82	32176	0,74
СССР	31 036	1,16	9 852	0,30	0	0,00	0	0,00
Украина	0	0,00	2 078	0,06	3 606	0,09	3921	0,09
Северная Африка	25 502	0,95	31 975	0,96	43 463	1,13	50629	1,17
Центральная/Южная Африка	12 657	0,47	19 714	0,59	34 879	0,91	55870	1,29
Регион Сахары	10 001	0,37	10 491	0,32	12 104	0,31	13732	0,32

Примечание: Индикаторы цитируемости - число ссылок на работы ученых страны и процент от их общего числа в строке "Все страны" - определяется в пределах 3-х лет с двухлетним лагом. Например, число ссылок в 2001г. берется из всех статей, опубликованных в журналах 2001г., обрабатываемых Институтом научной информации (США), при условии, что цитируемая работа была опубликована в период 1997–1999 гг.

5 Заключение

Необходимость формирования и использования электронных библиотек в рамках новой модели распределения бюджетных средств не отменяет использования традиционных статистических баз данных в бюджетном процессе. Отметим основную причину необходимости создания электронных библиотек в дополнение к уже существующим статистическим базам данных. Статистические базы данных не позволяют решать задачи верификации из-за отсутствия исходной научной информации, так как содержат число статей, а не сами статьи, число специалистов, а не списки персоналий, число конференций, журналов и т.д., а не списки их названий. Электронные библиотеки, хранящие первичную информацию, позволяют лицам, принимающим решения, проверить значения любого индикатора в процессе распределения бюджетных средств в сфере науки.

Литература

[1] Зацман И.М. Терминологический анализ нормативно-правового обеспечения создания систем мониторинга и оценки результативности в сфере науки // Экономическая наука современной России, № 4, 2005.- С. 114-129.

[2] Поспелов Д.А. Логико-лингвистические модели в системах управления.- М.: Энергоиздат, 1981.

[3] Зацман И.М. Полидоменные модели в системах оценки инновационного потенциала и результативности научных исследований // Труды международной конференции Диалог-2006 "Компьютерная лингвистика и интеллектуальные технологии".- М.: Изд-во РГГУ, 2006.- С. 178-183.

[4] Клейнер Г.Б. Эволюция институциональных систем.- М.: Наука, 2004.

[5] Gorn S. Informatics (computer and information science): its ideology, methodology, and sociology. In: The studies of information: Interdisciplinary messages / Ed. By F. Machlup and U. Mansfield. – New York: Wiley, 1983. – P. 121-140.

[6] Алфимов М.В., Минин В.А., Либкинд А.Н. Страна науки - РФФИ. В кн.: Гранты РФФИ: результаты и анализ.- М.: Янус-К, 2001г.- С. 11-57.

[7] Минин В.А. Мониторинг научных исследований российских ученых. В кн.: Российский фонд фундаментальных исследований: десять лет служения российской науке.- М.: Научный мир, 2003.- С. 295-314.

[8] National Science Board, Science and Engineering Indicators – 2002. Two volumes. Arlington, VA: National Science Foundation, 2002.

[9] National Science Board, Science and Engineering Indicators – 2004. Two volumes. Arlington, VA: National Science Foundation, 2004.

[10] Маркусова В.А. Информационные ресурсы для мониторинга российской науки // Вестник РАН, № 7, 2005.- С. 607-612.

[11] Стенограмма выступления Заместителя Председателя Правительства РФ А.Д. Жукова на VI Международной научной конференции "Модернизация экономики и выращивание институтов" на сайте ГУ-ВШЭ по адресу: http: //www.hse.ru/temp/2005/files/04_06_2005_jukov.doc.

[12] Planning for Performance and Evaluating Results of Public R&D Programs: Meeting the OMB PART Challenge (Workshop Report).- Washington: The Washington Research Evaluation Network, 2004.

[13] Зацман И.М. Информационные ресурсы для систем мониторинга в сфере науки // Системы и средства информатики. Вып. 15.- М.: Наука, 2005.- С. 288-318.

[14] National Science Board, Science and Engineering Indicators – 2006. Two volumes. Arlington, VA: National Science Foundation, 2006.

Polydomain models for digital libraries of monitoring systems in scientific sphere

I. Zatsman

The paper is devoted to a modelling problem of digital libraries for monitoring systems in scientific sphere whose resources are used simultaneously for the solution of information and analytical problems. In monitoring systems the solution need of analytical problems follows from the new model of budgeting process and is connected to computation of a wide spectrum of performance and efficiency indicators in scientific sphere. Computation need of indicators implies a number of new requirements to designing digital libraries. For draft stage specifications, including requirements to information resource scheme for digital libraries of monitoring systems, it is offered to use polydomain models. They are development of logical-linguistic models of management.