ОГЛАВЛЕНИЕ

ОТ СОСТАВИТЕЛЕЙ

Настоящий тематический выпуск журнала «Электронные библиотеки» состоит из двух частей и включает статьи, подготовленные их авторами на основе материалов, представленных на научной конференции «Научный сервис в сети Интернет».

Эта конференция состоялась 23–28 сентября 2019 г. в окрестностях Новороссийска. Организатором конференции был Институт прикладной математики им. М.В. Келдыша Российской академии наук. Конференция собрала около 140 участников из разных городов России, в т. ч. Москвы, Санкт-Петербурга, Иркутска, Казани, Красноярска, Новосибирска, Ростова-на Дону, Томска и др.

Тематика конференции достаточно широка: от цифровых библиотек, библиографических баз и наукометрии до различных специальных областей использования возможностей интернета для научных исследований.

Первая часть тематического выпуска размещена в №3 журнала «Электронные библиотеки», вторая часть – в №4.

М. М. Горбунов-Посадов, А. М. Елизаров

УДК 004.432

РАЗВИТИЕ DVM-СИСТЕМЫ

В. Ф. Алексахин [1], В. А. Бахтин [1, 2], О. Ф. Жукова [1], Д. А. Захаров [1], В. А. Крюков [1, 2], Н. В. Поддерюгина [1], О. А. Савицкая [1]

[1] Институт прикладной математики им. М.В. Келдыша Российской академии наук, г. Москва;
[2] Московский государственный университет им. М.В. Ломоносова, г. Москва
valex@keldysh.ru, bakhtin@keldysh.ru, socol@keldysh.ru, s123-93@mail.ru, krukov@keldysh.ru, npodderyugina@gmail.com, savol@keldysh.ru

Аннотация

DVM-система предназначена для разработки параллельных программ научно-технических расчетов на языках C-DVMH и Fortran-DVMH. Эти языки используют единую DVMH-модель параллельного программирования и являются расширением стандартных языков Си и Фортран спецификациями параллелизма, оформленными в виде директив для компилятора. DVMH-модель позволяет создавать эффективные параллельные программы для гетерогенных вычислительных кластеров, в узлах которых в качестве вычислительных устройств наряду с универсальными многоядерными процессорами могут использоваться ускорители, графические процессоры или сопроцессоры Intel Xeon Phi. В статье представлены новые возможности DVM-системы, которые были разработаны в последнее время.

Ключевые слова*: автоматизация разработки параллельных программ, DVM-система, ускоритель, ГПУ, Фортран, Си, нерегулярная сетка, неструктурированная сетка.*

ВВЕДЕНИЕ

Модель программирования DVMH [1, 2] построена на парадигме параллелизма по данным. В основе этой модели лежит понятие распределенного многомерного массива. При этом у каждого процессора имеются не только локальная часть распределенного массива, но и так называемые теневые грани – копии элементов из локальных частей соседних процессоров, через которые осуществляется основное взаимодействие процессоров. Распределение вычислений про-

изводится посредством их отображения на распределенные массивы, при этом обращения происходят либо в свою локальную часть, либо в теневые грани, определяемые как продолжение локальной части по конкретному измерению распределенного массива на заранее известную ширину. Например, для шаблона типа «крест» с 4 соседями, элемент с индексами (i,j) рассчитывается по элементам с индексами (i-1,j), (i,j-1), (i+1,j), (i,j+1), что приводит к необходимости иметь теневые грани ширины 1 по обоим измерениям.

DVMH-компиляторы преобразуют обращения к распределенным многомерным массивам в форму, независимую от размеров и положения локальной части на каждом процессоре, при этом исходные индексные выражения остаются нетронутыми. В результате каждое обращение к распределенным данным ведется в глобальных (исходных) индексах, а при доступе к памяти применяются вычисляемые во время выполнения коэффициенты и смещения для каждого измерения. Такой подход (в отличие от изменения индексных выражений) позволяет абстрагироваться от содержания распараллеливаемых циклов, но и вводит серьезное ограничение на форму адресуемой каждым процессором части распределенного массива, называемой расширенной локальной частью, которая является объединением локальной части и теневых граней. В модели DVMH используются блочные распределения массивов с теневыми гранями. Таким образом, расширенная локальная часть представляет собой подмассив исходного массива вида (A1:B1, A2:B2, A3:B3, ..., An:Bn). Такое ограничение затрудняет процесс распараллеливания программ на неструктурированных сетках с помощью DVM-системы [3].

В статье описаны новые возможности DVM-системы, нацеленные на борьбу с этим ограничением. В первой главе представлены новые виды распределений массивов, новые конструкции для работы с теневыми гранями, новые средства для перехода от глобальных индексов к локальным, а также приведен пример программы на языке Fortran-DVMH, использующей названные возможности.

Во второй главе описаны средства, которые позволяют программисту вручную распределять данные, используя MPI или другие технологии параллельного программирования, оставляя при этом возможность использования DVM-языков внутри узла кластера для отображения вычислений по ядрам цен-

трального процессора или графического ускорителя. Новые спецификации параллелизма существенно упрощают процесс разработки программ, использующих неструктурированные сетки.

РАСШИРЕНИЕ DVM-СИСТЕМЫ ДЛЯ РАБОТЫ С НЕСТРУКТУРИРОВАННЫМИ СЕТКАМИ

Для работы с неструктурированными сетками в DVM-системе реализован новый вид распределения массивов и шаблонов – поэлементное распределение. Этот вид распределения не накладывает никаких ограничений на то, какие элементы массива должны располагаться на одном и том же процессоре или какие элементы массива должны располагаться на соседних процессорах. Напротив, он позволяет задать произвольную принадлежность каждого элемента массива независимо.

Введены два новых правила поэлементного распределения: косвенное (INDIRECT) и производное (DERIVED). Косвенное распределение задается массивом целых чисел, размер которого равен размеру косвенно распределяемого измерения, а значения задают номер домена. При этом доменов может быть как больше числа процессоров, так и меньше. DVM-система гарантирует принадлежность всех элементов домена одному и тому же процессору.

Производное распределение задается правилом, по форме похожим на правило выравнивания (ALIGN) модели DVMH. Однако у него появляется значительно большая гибкость. Синтаксис можно описать так, как показано на рис. 1.

```
      indirect-rule ::= INDIRECT ( var-name )
      derived-rule ::= DERIVED ( derived-elem-list WITH derived-templ )
      derived-elem ::= int-range-expr
      int-range-expr ::= прозвольное целочисленное выражение + в индексных выражениях допустимы диапазоны, использование align-dummy переменных.
      derived-templ ::= var-name [ derived-templ-axis-spec ]...
      derived-templ-axis-spec ::= [ ] | [ @ align-dummy [ + shadow-name ]... ] | [ int-expr ]
```

Рис. 1. Формула БНФ для новых правил распределения

Все ссылки на распределенные массивы в int-range-expr обязаны быть доступны (элемент входит в расширенную локальную часть) для соответствующего элемента шаблона (перебор элементов шаблона осуществляется по его локаль-

ной части и указанным теневым граням). Если производным правилом один и тот же элемент подлежит распределению сразу на несколько процессоров, то DVM-система решает, на какой из них фактически будет распределен такой элемент, а на остальных процессорах добавляет его в теневую грань с названием «overlay». Элементов, не распределенных ни на один процессор, быть не должно. Такие случаи являются ошибкой времени выполнения и приводят к останову. Вычисленные несуществующие индексы распределяемого массива игнорируются, не приводя к ошибке.

Наложение (overlay) вводится для возможности согласованного распределения сеточных элементов, например, ячейки, ребра, вершины. В таком случае появляется возможность построить одно распределение на основе другого, причем в любой последовательности.

В результате такого распределения у массива появляются два вида нумерации элементов: глобальная (она же исходная в последовательной программе) и локальная. Локальная нумерация непрерывна в рамках одного процессора, т. е. существует такой порядок локальных элементов, что их локальные индексы полностью заполнят некоторый целочисленный отрезок [Li, Hi].

Также вводятся поэлементые теневые грани. Теневая грань – это набор элементов, не принадлежащих текущему процессу (требование принадлежности соседнему процессу снимается), для которых, во-первых, возможен доступ без специальных указаний из любой точки программы, и, во-вторых, введены специальные средства работы с ними: обновление указанием SHADOW_RENEW, расширение параллельного цикла указанием SHADOW_COMPUTE и т. п.

В отличие от традиционных, поэлементные теневые грани добавляются к шаблонам во время работы программы и имеют имя для ссылки на них. Задаются они практически так же, как и производное распределение (рис. 2).

```
      shadow-add ::= SHADOW_ADD ( templ-name [ shadow-axis ]... = shadow-name ) [
INCLUDE_TO ( var-name-list ) ]
      shadow-axis ::= [ ] | [ derived-elem-list WITH derived-templ ]
```

Рис. 2. Формула БНФ для задания поэлементных теневых граней

Ровно один из shadow-axis должен быть непустыми скобочками. Все массивы из списка, указанного в INCLUDE_TO, должны быть выравнены на шаблон, к

измерению которого добавляется теневая грань. В результате выполнения такой директивы к шаблону добавляется теневая грань и включается в указанные распределенные массивы. После этой операции теневые элементы массивов доступны на чтение из программы, а также могут обновляться с помощью директивы SHADOW_RENEW.

Для реализации поэлементных теневых граней и производного распределения компилятор на основе указанных выражений генерирует специального вида функцию, в которую передаются системой поддержки параметры для обхода локальной части шаблона. Эта функция, обходя шаблон, заполняет буфер индексов элементов согласно выражениям в левой части правила отображения, а затем возвращает обратно в систему поддержки. Затем буфер анализируется средствами системы поддержки.

Для экспериментальной эксплуатации этих возможностей была введена вспомогательная директива локализации значений индексного массива, которая изменяет значения целочисленного массива, заменяя глобальные индексы указанного целевого массива на локальные (рис. 3).

```
      localize-spec ::= LOCALIZE ( ref-var-name => target-var-name [ axis-
specifier ]...
      axis-specifier ::= [ ] | [ : ]
```

Рис. 3. Формула БНФ для директивы локализации значений индексного массива

После проведения такой операции становится возможным использовать имеющийся способ компиляции параллельных циклов: они будут выполняться полностью в локальных индексах.

Вместе с модификацией директивы теневых обменов и реализацией обменов для поэлементных теневых граней (которые теперь происходят не обязательно с соседними процессорами, а с произвольным подмножеством процессоров), этот набор расширений позволяет распараллелить и запустить приложения, использующие нерегулярные сетки, на кластере с ускорителями.

Для иллюстрации новых возможностей рассмотрим небольшой пример программы на языке Fortran, реализующий трехмерный алгоритм Якоби (рис. 4). В данной программе вместо трехмерных массивов используются одномерные

массивы. Из-за этого появляется косвенная адресация, инструментов для работы с которой ранее в DVM не было.

```
            program JAC_INDIRECT
            parameter (L=100, itmax=5000)
            real*8:: tmp,eps, maxeps=0.005
            integer x_t,y_t,z_t,cur
            real*8, allocatable :: A(:),B(:)
            integer, allocatable :: ibstart(:), ibend(:), ib(:)
            integer, allocatable :: indir_x(:), indir_y(:),indir_z(:)
            allocate(A(L*L*L),B(L*L*L), ibstart(L*L*L), ibend(L*L*L))
            allocate(indir_x(L*L*L), indir_y(L*L*L), indir_z(L*L*L))
! Здесь происходит создание одномерного массива, который "эмулирует"
! трехмерный массив в обычном трехмерном алгоритме Якоби
            cur = 1
            do i = 1,L*L*L
              x_t = (i-1) / (L*L)
              y_t = mod((i-1) / L, L)
              z_t = mod(i-1, L)
              indir_x(i) = x_t
              indir_y(i) = y_t
              indir_z(i) = z_t
              ibstart(i) = cur
              if (x_t.gt.0) cur = cur + 1
              if (x_t.lt.L-1  cur = cur + 1
              if (y_t.gt.0)   cur = cur + 1
              if (y_t.lt.L-1) cur = cur + 1
              if (z_t.gt.0)   cur = cur + 1
              if (z_t.lt.L-1) cur = cur + 1
              ibend(i) = cur - 1
            enddo
            allocate(ib(cur-1))
            cur = 1
            do i = 1,L*L*L
              x_t = (i-1) / (L*L)
              y_t = mod((i-1) / L, L)
              z_t = mod(i-1, L)
              if (x_t.gt.0) then
                  ib(cur) = i - (L*L)
                  cur = cur + 1
              endif
              if (x_t.lt.L-1) then
                  ib(cur) = i+(L*L)
                  cur = cur + 1
              endif
              if (y_t.gt.0) then
                  ib(cur) = i-L
                  cur = cur + 1
              endif
              if (y_t.lt.L-1) then
                  ib(cur) = i+L
                  cur = cur + 1
              endif
              if (z_t.gt.0) then
                  ib(cur) = i-1
                  cur = cur + 1
```

```
              endif
              if (z_t.lt.L-1) then
                  ib(cur) = i+1
                  cur = cur + 1
              endif
            enddo
! Для упаковки массива используется аналогичный CSR (Compressed Sparse Row)
! формат. У каждого элемента может быть до 6 соседних элементов – слева и
! справа по каждому из трех измерений. Для i-ого элемента массива A список
! его соседей содержится в массиве ib начиная с индекса ibstart(i) и заканчивая
! индексом ibend(i). Выше происходит создание этой похожей на CSR
! структуры. Также заполняются массивы indir_x/y/z, которые содержат
! индексы, которые были у элемента в трехмерном массиве.

! Перед итерационным циклом массивы заполняются. Так как все элементы
! теперь сложены в одномерный массив – требуется проверять трехмерные
! индексы, чтобы исключить обработку граничных элементов.
            do i = 1, L*L*L
              A(i) = 0
              if (indir_x(i) == 0 .or. indir_x(i) == L-1 .or.
           &      indir_y(i) == 0 .or. indir_y(i) == L-1 .or.
           &      indir_z(i) == 0 .or. indir_z(i) == L-1) then
                  B(i) = 0
              else
                  B(i) = 4 + indir_x(i) + indir_y(i) + indir_z(i)
              endif
            enddo
! После заполнения применяется видоизмененный алгоритм Якоби
            do it = 1, itmax
              eps = 0
              do i = 1,L*L*L
                 if (indir_x(i) /= 0 .and. indir_x(i) /= L-1 .and.
           &         indir_y(i) /= 0 .and. indir_y(i) /= L-1 .and.
           &         indir_z(i) /= 0 .and. indir_z(i) /= L-1) then
                    tmp  = ABS(B(i) - A(i))
                    eps = MAX(tmp, eps)
                    A(i) = B(i)
                 endif
              enddo
              do i = 1, L*L*L
                 if (indir_x(i) /= 0 .and. indir_x(i) /= L-1 .and.
           &         indir_y(i) /= 0 .and. indir_y(i) /= L-1 .and.
           &         indir_z(i) /= 0 .and. indir_z(i) /= L-1) then
! Косвенная адресация
                    B(i) = (A(ib(ibstart(i)))   + A(ib(ibstart(i)+1))
           &              + A(ib(ibstart(i)+2)) + A(ib(ibstart(i)+3))
           &              + A(ib(ibstart(i)+4)) + A(ib(ibstart(i)+5)))
           &              / 6.0
                 endif
              enddo
              print 200,  it, eps
     200      format(' it = ', i4, '   eps = ', e14.7)
              if ( eps .lt. maxeps )    exit
            enddo
            deallocate(ibstart,ibend)
            deallocate(ib)
            deallocate(A,B,indir_x,indir_y,indir_z)
            end program
```

Рис. 4. Последовательная версия программы, реализующая алгоритм Якоби

Начнем рассматривать параллельный вариант программы:

```
 program JAC_INDIRECT
parameter (L=100, itmax=5000)
real*8:: tmp,eps, maxeps=0.005
integer x_t,y_t,z_t,cur
real*8, allocatable :: A(:),B(:)
integer, allocatable :: ibstart(:), ibend(:), ib(:)
integer, allocatable :: indir_x(:), indir_y(:),indir_z(:)
integer MAP(L*L*L)
!DVM$   TEMPLATE E(L*L*L)
!DVM$   TEMPLATE :: E2(:)
!DVM$   DISTRIBUTE :: E
!DVM$   DISTRIBUTE :: E2
!DVM$   ALIGN :: A,B
!DVM$   ALIGN :: indir_x, indir_y,indir_z, ibstart, ibend
!DVM$   ALIGN :: ib
      call fillMap(map,L,1)
      allocate(A(L*L*L),B(L*L*L), ibstart(L*L*L), ibend(L*L*L))
      allocate(indir_x(L*L*L), indir_y(L*L*L), indir_z(L*L*L))
!DVM$ REDISTRIBUTE E(INDIRECT(map))
!DVM$ REALIGN (I)  WITH  E(I) :: A,B,indir_x, indir_y,indir_z
!DVM$ REALIGN (I)  WITH  E(I) :: ibstart, ibend
```

Первое изменение – добавление массива MAP. Этот массив будет служить «картой распределения», на основе которой мы будем распределять данные. Также объявляются два шаблона – статический шаблон E, который будет распределяться поэлементно, и динамический шаблон E2, о котором будет сказано чуть позднее. Для этих шаблонов указана директива DISTRIBUTE без параметров, которая означает, что эти шаблоны будут распределены позже. Также указана директива ALIGN без параметров для всех массивов, которая говорит о том, что эти массивы будут в дальнейшем выровнены на какой-либо шаблон или уже

распределенный массив. После этого добавляется функция заполнения карты – fillMap. Одна из возможных реализаций данной функции выглядит так:

```
      subroutine fillMap(map,L,axis)
      integer  numproc
      integer i,L,axis
      integer map(L*L*L)
! Эта строка нужна для совместимости программы с обычными
! компиляторами
      PROCESSORS_SIZE(axis) = 1
      numproc = PROCESSORS_SIZE(axis)
        do i = 1,L*L*L
           map(i) = ((i-1) * numproc) / (L*L*L)
        enddo
      end subroutine
```

PROCESSORS_SIZE(axis) – служебная функция, которая возвращает количество процессоров в оси axis решетки процессоров, на которой была запущена программа. Так как данная программа одномерная – axis равно 1, и в дальнейшем все будет описываться с учетом того, что решетка запуска одномерная. Конкретная реализация имитирует блочное распределение – карта делится на равные блоки, и все элементы из первого блока идут на процессор с индексом 0, все элементы из второго блока идут на процессор с индексом 1 и так далее.

После заполнения карты распределения она тут же используется в директиве REDISTRIBUTE. Здесь в качестве типа распределения указан INDIRECT – поэлементное распределение. При поэлементном распределении i-й элемент шаблона оказывается на том процессоре, индекс которого указан в карте на i-й позиции. Это позволяет распределять данные в любом формате – можно использовать блочное распределение, как здесь, можно распределять элементы поочередно, когда каждый следующий элемент распределяется на другой процессор, а можно и вовсе распределить их случайным образом. У программиста есть возможность задать любое отображение.

После этого все нужные массивы выравниваются на новосозданный шаблон через директиву REALIGN. После выполнения этой директивы элементы с

индексом i для всех указанных в ней массивов будут распределены на тот же процессор, на который был распределен i-й элемент шаблона E. Использование шаблонов для задания изначального поэлементного распределения в данном случае является необходимым, распределять поэлементно массив напрямую нельзя.

Следующее изменение в программе появляется после выделения памяти под массив ib:

```
      allocate(ib(cur-1))
!DVM$ TEMPLATE_CREATE(E2(cur-1))
!DVM$ REDISTRIBUTE E2(DERIVED((ibstart(i):ibend(i)) with E(@i)))
!DVM$ REALIGN (I)  WITH  E2(I) :: ib
```

Здесь появляется еще один новый тип распределения данных – DERIVED (производное распределение). Производное распределение – это вариант поэлементного распределения, идея которого состоит в том, что оно как раз является «производным» из какого-либо другого распределения. Стоит вспомнить, как выглядела косвенная адресация в последовательной программе:

```
       B(i) = (A(ib(ibstart(i)))   + A(ib(ibstart(i)+1)) + A(ib(ibstart(i)+2)) +
A(ib(ibstart(i)+3)) + A(ib(ibstart(i)+4)) + A(ib(ibstart(i)+5))) / 6.0
```

Отсюда мы можем заметить, что на одном процессоре вместе с B(i), который у нас уже распределен поэлементно, мы должны иметь элементы массива ib с индексами от ibstart(i) до ibstart(i)+5, то есть все элементы-соседи, учитывая формат хранения данных – конечный индекс на самом деле будет ibend(i), который для всех неграничных элементов как раз равен ibstart(i)+5. Обеспечить наличие всех нужных элементов мы сможем через производное распределение. Для распределения массива ib будет использоваться шаблон E2, который создается динамически, поскольку на старте программы мы не знаем размера массива ib, значит, и размер шаблона. Сразу после этого к шаблону применяется директива redistribute с производным типом распределения. Данная директива означает, что в новом шаблоне E2 элементы с индексами, начиная с ibstart(i) и

заканчивая ibend(i), должны находиться на том же процессоре, где находится i-й элемент шаблона E. Вместо указания диапазона ibstart(i):ibend(i) в директиве может указываться просто список индексов через запятую (или вовсе один индекс). После этого массив ib выравнивается на новосозданный шаблон, и тем самым гарантируется, что на одном процессоре вместе с элементом B(i) будут находиться все его соседи. Для всех неграничных элементов массива B это значит, что на один процессор вместе с элементом B(i) попадут все элементы от ib(ibstart(i)) до ib(ibstart(i)+5). Стоит отметить, что если при создании производного шаблона сразу несколько процессоров захотят получить один и тот же элемент, то этот элемент дается какому-то одному из процессоров, а для других он помещается в автоматически создаваемую теневую грань.

Следующее изменение появляется после заполнения массива ib:

```
! ……
        if (z_t.lt.L-1) then
            ib(cur) = i+1
            cur = cur + 1
        endif
      enddo
!DVM$ LOCALIZE(ibstart => ib(:))
!DVM$ LOCALIZE(ibend => ib(:))
!DVM$ SHADOW_ADD(E((ib(ibstart(i):ibend(i))) with E(@i)) = "nei1")
!DVM$& INCLUDE_TO A
!DVM$ LOCALIZE(ib => A(:))
```

Директива LOCALIZE – служебная директива, которая преобразует глобальные индексы в локальные, что необходимо для корректной адресации массивов. Данная директива должна быть применена ко всем массивам, которые используются для индексации поэлементно распределенных массивов. В директиве слева указывается массив, который нужно локализовать, а справа – массив, который будет индексироваться локализуемым массивом. Для массивов с 2 и более измерениями необходимо также указывать измерение, на которое производится локализация. Директива должна использоваться после того, как локализуемый массив был полностью заполнен и больше не будет изменяться, но до

его использования для индексации распределенного массива в параллельном цикле или в директиве SHADOW_ADD. В данном случае – массивы ibstart и ibend уже были заполнены и будут использованы для индексации тут же в директиве SHADOW_ADD.

Еще раз вспомним, как выглядела косвенная индексация в основном цикле:

```
        B(i) = (A(ib(ibstart(i)))   + A(ib(ibstart(i)+1)) + A(ib(ibstart(i)+2)) +
A(ib(ibstart(i)+3)) + A(ib(ibstart(i)+4)) + A(ib(ibstart(i)+5))) / 6.0
```

Мы позаботились о массиве ib, но у нас остался массив A, который индексируется посредством массива ib. Для того чтобы гарантировать наличие нужных элементов массива A на процессоре, где находится элемент B(i), нам необходимо добавить теневую грань к массиву A, что и делает директива SHADOW_ADD. Данный экземпляр директивы говорит о том, что на один процессор вместе с i-м элементом шаблона E (часть WITH E(@i)) мы должны добавить в теневую грань все элементы шаблона E (первое вхождение E в директиве), индексы которых находятся в массиве ib, начиная с индекса ibstart(i) и заканчивая индексом ibend(i). Далее эта теневая грань получает название «nei1», и указывается, что эту теневую грань нужно добавить для массива A. Тем самым мы создали теневую грань, которая для каждого элемента A(i) содержит всех его соседей. При этом директива SHADOW_ADD следит за тем, чтобы в теневой грани не было дублирующих элементов. Если элемент уже присутствует на процессоре, он не будет добавлен в теневую грань. Необходимо отметить, что массив ib локализуется после директивы SHADOW_ADD. Так как он используется для индексации массива A, локализуется он именно на него.

После этого остается лишь указать директивы PARALLEL и регионы:

```
!DVM$ REGION
!DVM$ PARALLEL (i) ON B(i)
      do i = 1, L*L*L
        A(i) = 0
        if (indir_x(i) == 0 .or. indir_x(i) == L-1 .or.
     &      indir_y(i) == 0 .or. indir_y(i) == L-1 .or.
```

```
     &      indir_z(i) == 0 .or. indir_z(i) == L-1) then


            B(i) = 0
          else
            B(i) = 4 + indir_x(i) + indir_y(i) + indir_z(i)
          endif
        enddo
!DVM$ END REGION
        do it = 1, itmax
!DVM$ REGION
          eps = 0
!DVM$   PARALLEL (i) ON B(i), REDUCTION(MAX(eps)), PRIVATE(tmp)
          do i = 1,L*L*L
             if (indir_x(i) /= 0 .and. indir_x(i) /= L-1 .and.
     &          indir_y(i) /= 0 .and. indir_y(i) /= L-1 .and.
     &          indir_z(i) /= 0 .and. indir_z(i) /= L-1) then
                tmp  = ABS(B(i) - A(i))
                eps = MAX(tmp, eps)
                A(i) = B(i)
             endif
          enddo
!DVM$   PARALLEL (i) ON B(i), SHADOW_RENEW(A)
          do i = 1, L*L*L
             if (indir_x(i) /= 0 .and. indir_x(i) /= L-1 .and.
     &          indir_y(i) /= 0 .and. indir_y(i) /= L-1 .and.
     &          indir_z(i) /= 0 .and. indir_z(i) /= L-1) then
                B(i) = (A(ib(ibstart(i)))   + A(ib(ibstart(i)+1))
     &               + A(ib(ibstart(i)+2)) + A(ib(ibstart(i)+3))
     &               + A(ib(ibstart(i)+4)) + A(ib(ibstart(i)+5)))
     &               / 6.0
             endif
          enddo
!DVM$ END REGION
```

```
!DVM$ GET_ACTUAL(eps)
        print 200,  it, eps
200     format(' it = ', i4, '   eps = ', e14.7)
        if ( eps .lt. maxeps )    exit
      enddo
```

Директива PARALLEL в данном случае распределяет витки цикла поэлементно на основе распределения массива B. i-й виток цикла выполняется на том процессоре, где находится элемент B(i), значит, на том процессоре, индекс которого был записан в map(i) на момент выполнения директивы распределения шаблона E.

Клауза SHADOW_RENEW для A в данном случае будет обновлять все теневые грани, привязанные к массиву A. В этом примере таковая одна – та самая nei1, которая была объявлена через SHADOW_ADD. Остальные директивы/клаузы ничем не отличаются от стандартного DVM без расширения. Клауза REDUCTION(max(eps)) обеспечивает то, что на каждом процессоре у нас будет максимальным значение eps по итогам всех витков цикла, а не только витков этого процессора. Клауза PRIVATE(tmp) говорит о том, что переменная tmp приватная, значит, ее значение на одном витке не влияет на другие витки. Директивы REGION и END REGION показывают участки кода, которые следует выполнять на графическом ускорителе, если таковой выделен программе, и директива GET_ACTUAL(eps) указывает на то, что конкретное значение переменной eps находится на графическом ускорителе, и его нужно скопировать в оперативную память.

Полученная программа может выполняться на гетерогенном вычислительном кластере с ускорителями.

ДОПОЛНИТЕЛЬНОЕ РАСПАРАЛЛЕЛИВАНИЕ СУЩЕСТВУЮЩИХ MPI-ПРОГРАММ

В настоящее время, когда параллельные машины уже не одно десятилетие эксплуатируются для проведения расчетов, имеется множество программ, которые уже распараллелены на кластер, однако не имеют распараллеливания по

ядрам центрального процессора, а также не используют графические ускорители.

Традиционно в DVM-подходе весь процесс программирования (или распараллеливания имеющихся последовательных программ) начинается с распределения массивов, а затем отображения на них параллельных вычислений. Это означает, что для использования средств DVM-системы распараллеленные, например, на MPI, программы приходится превращать обратно в последовательные и заменять распределенные вручную данные и вычисления на описанные на DVM-языке распределенные массивы и параллельные циклы.

Однако, во-первых, автору не всегда хочется отказываться от своей параллельной программы, во-вторых, не всегда удастся перевести исходную схему распределения данных и вычислений на DVM-язык. В частности, перевод задач на нерегулярных сетках в модель DVMH может потребовать применения нетривиальных решений, что не всегда возможно.

Одним из способов избавиться от обеих проблем является новый режим работы DVM-системы, в котором она не вовлечена в межпроцессорное взаимодействие, а работает локально в каждом процессе.

Данный режим включается заданием специально созданной MPI-библиотеки при сборке DVM-системы. Эта библиотека не производит никаких коммуникаций и не конфликтует с реальными MPI-реализациями. В результате для системы поддержки выполнения DVMH-программ создается иллюзия запуска программы на 1 процессоре.

Кроме такого режима, в языках Fortran-DVMH и C-DVMH введено понятие нераспределенного параллельного цикла, для которого нет необходимости задавать отображение на распределенный массив. Например, трехмерный параллельный цикл может выглядеть так (рис. 5):

```
#pragma dvm parallel(3) reduction (max(eps)) // Для языка C-DVMH
for (int i = L1; i <= H1; i++)
    for (int j = L2; j <= H2; j++)
        for (int k = L3; k <= H3; k++)
...
!DVM$ PARALLEL(I,J,K) REDUCTION (MAX(EPS))   ! Для языка Fortran-DVMH
```

```
DO I = L1,H1
   DO J = L2, H2
      DO K = L3, H3
...
```

Рис. 5. Нераспределенный параллельный цикл

По определению такой цикл выполняется всеми процессорами текущей многопроцессорной системы, но т. к. DVM-система в описанном новом режиме считает многопроцессорной системой ровно один процесс, такая конструкция не приводит к размножению вычислений, а только лишь позволяет использовать параллелизм внутри одного процесса – использовать ядра центрального процессора или графического ускорителя. Как следствие, появляется возможность не задавать ни одного распределенного в терминах модели DVMH массива и в то же время пользоваться возможностями DVM-системы:

- использовать параллелизм на общей памяти (задействовать ядра центрального процессора);
- задействовать графические ускорители: не только «наивное» портирование параллельного цикла на ускоритель, но и выполнение автоматической реорганизации данных, упрощенное управление перемещениями данных;
- подбирать оптимизационные параметры;
- использовать удобные средства отладки производительности.

Такой режим может быть использован, в том числе, для получения промежуточных результатов в процессе проведения полноценного распараллеливания программы в модели DVMH. Он позволяет быстро и заметно проще получить программу для многоядерного центрального процессора и графического ускорителя, а также оценить перспективы ускорения целевой программы на кластере с многоядерными процессорами и ускорителями.

ИСПОЛЬЗОВАНИЕ НОВЫХ ВОЗМОЖНОСТЕЙ

Для демонстрации использования новых возможностей при работе с неструктурированными сетками рассмотрим программу для решения задачи газовой динамики методом Галеркина, который широко применяется на практике и характеризуется высоким порядком точности получаемого решения. В качестве

тестовой задачи рассмотрим простую волну, в которой энтропия и инвариант Римана являются постоянными [4, 5], используется классический лимитер Кокбурна, который легко реализуется в многомерном случае на сетках произвольной структуры. Основными вычислительными элементами в программе являются массивы ячеек сетки, содержащие различные величины и характеристики. Сетка является неструктурной, а ячейки в этой сетке – треугольные, поэтому обращения к соседним элементам по всей программе происходят с использованием косвенной адресации.

Для данной программы были разработаны 2 параллельных версии – с использованием «старых» и «новых» возможностей DVM-системы. «Старая» версия программы использует блочное распределение данных, «новая» версия программы использует расширение для работы с неструктурированными сетками, описанное в данной статье. На рис. 6 показано сравнение времен выполнения 2-х версий программы для сетки из 200000 ячеек на различном числе ядер вычислительного кластера К-100 (ИПМ им. М.В. Келдыша РАН).

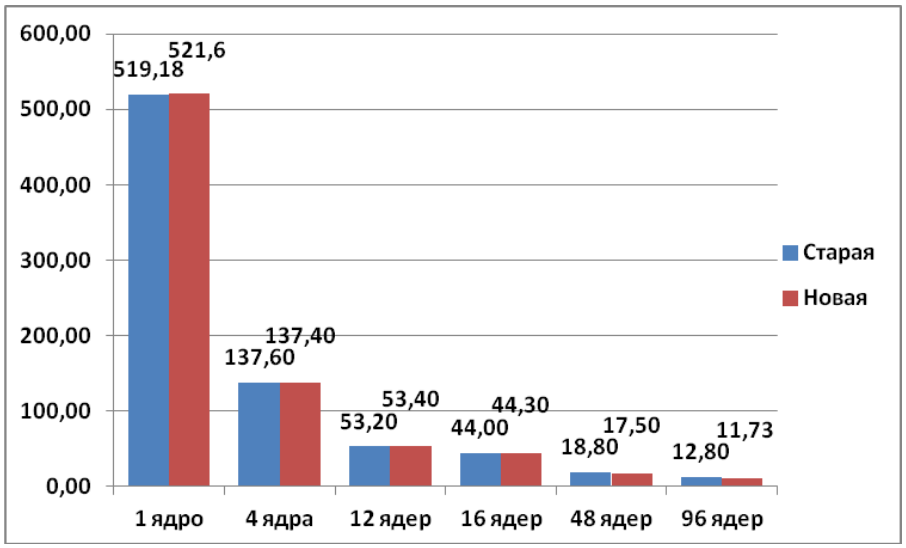


Рис. 6. Времена выполнения 2-х версий программы на кластере К-100 (в секундах)

Времена выполнения «старой» и «новой» версий программы практически не отличаются. При использовании большого числа ядер «новая» версия программы начинает выполняться быстрее, чем «старая». При этом следует отметить, что процесс распараллеливания программы без использования расширения был значительно сложнее. Потребовалось существенное изменение кода

программы, а самое главное – такое распараллеливание стало возможным лишь после переупорядочивания элементов сетки по принципу локальности. В результате выполнения данного переупорядочивания сетки удалось найти небольшое значение M, при котором все соседи ячейки с номером N имели номера, отличающиеся от N не более, чем на M, что позволило использовать блочное распределение для массивов и старый механизм теневых граней для обновления значений соседних ячеек (максимальный размер теневых граней был задан равным M).

Для демонстрации возможности дополнительного распараллеливания существующих MPI-программ рассмотрим программу, являющуюся частью большого развитого комплекса вычислительных программ (В.А. Гасилов, А.С. Болдарев, ИПМ им. М.В. Келдыша РАН). Будучи ориентированным на решение по явной схеме систем гиперболических уравнений (в основном, газовой динамики) в двумерных областях сложной формы с использованием неструктурированных сеток, этот код был написан на C++ с очень широким использованием объектно-ориентированного подхода для обеспечения максимальной универсальности и простоты дальнейшего развития.

Так как эта программа является частью целого комплекса, код ее основан на богатой платформе базовых понятий и структур данных. Это приводит к значительным размерам (39 тыс. строк) и сложности всей программы, если ее рассматривать целиком.

Полноценное распараллеливание программы в модели DVMH вряд ли возможно без рассмотрения и модификации всей программы. Новые возможности позволили выполнить «локальное» распараллеливание вычислительноемких частей программы. Модификации подверглись лишь 3 из 39 тыс. строк программы.

В результате такого распараллеливания на 12 ядрах ЦПУ с использованием OpenMP-нитей было получено ускорение в 9,83 раза относительно последовательной версии, а на ГПУ NVIDIA GTX Titan — в 18 раз относительно последовательной версии. Данные результаты подтверждают эффективность отображения рассматриваемой программы DVM-системой на ускорители и многоядерные процессоры и дают основания продолжить распараллеливание программы уже с использованием распределенных массивов в модели DVMH.

ЗАКЛЮЧЕНИЕ

DVM-система автоматизирует процесс разработки параллельных программ.

Получаемые DVMH-программы без каких-либо изменений могут эффективно выполняться на кластерах различной архитектуры, использующих многоядерные универсальные процессоры, графические ускорители и сопроцессоры Intel Xeon Phi. Это достигается за счет различных оптимизаций, которые выполняются как статически, при компиляции DVMH-программ, так и динамически.

Выше были представлены новые возможности DVM-системы, которые позволяют расширить область ее применимости и позволяют рапараллеливать не только задачи на структурированных сетках, для которых DVM-система была предназначена изначально [6], но и задачи на неструктурированных сетках.

В последнее время для численного решения задач математической физики стали активно использоваться адаптивные сетки – метод, который позволяет локально перестраивать сетку. Адаптация требуется, чтобы сгустить сеточные элементы в областях, где они наиболее необходимы, оставив сетку грубой в остальных местах. Такие сетки позволяют максимально точно передать ударные волны, фазовые переходы и другие области больших градиентов функций. Авторы проекта работают над расширением возможностей DVM-системы для поддержки адаптивных сеток.

СПИСОК ЛИТЕРАТУРЫ

1. Язык C-DVMH. C-DVMH компилятор. Компиляция, выполнение и отладка CDVMH-программ. URL: http://dvm-system.org/static_data/docs/CDVMH-reference-ru.pdf

2. Язык Fortran-DVMH. Fortran-DVMH компилятор. Компиляция, выполнение и отладка DVMH-программ. URL: http://dvm-system.org/static_data/docs/FDVMH-user-guide-ru.pdf

3. Система автоматизации разработки параллельных программ (DVM-система). URL: http://dvm-system.org

4. *Ладонкина М.Е., Неклюдова О.А., Тишкин В.Ф.* Лимитер повышенного порядка точности для разрывного метода Галеркина на треугольных сетках // Препринты ИПМ им. М.В.Келдыша. 2013. № 53. 26 с.

5. *Ладонкина М.Е., Неклюдова О.А., Тишкин В.Ф.* Исследование влияния лимитера на порядок точности решения разрывным методом Галеркина // Препринты ИПМ им. М.В. Келдыша. 2012. № 34. 31 с.

6. *Бахтин В.А., Захаров Д.А., Колганов А.С., Крюков В.А., Поддерюгина Н.В., Притула М.Н.* Решение прикладных задач с использованием DVM-системы // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2019. Т. 8. № 1. С. 89–106. DOI: 10.14529/cmse190106

PROGRESS IN DVM-SYSTEM

***V.F. Aleksahin*[1], *V.A. Bakhtin*[1,2], *O.F. Zhukova*[1], *D.A. Zakharov*[1], *V.A. Krukov*[1,2], *N.V. Podderyugina*[1], *O.A. Savitskaya*[1]**

[1] *Keldysh Institute of Applied Mathematics,*

[2] *Lomonosov Moscow State University*

valex@keldysh.ru, bakhtin@keldysh.ru, socol@keldysh.ru, s123-93@mail.ru, krukov@keldysh.ru, npodderyugina@gmail.com, savol@keldysh.ru

Abstract

DVM-system is designed for the development of parallel programs of scientific and technical calculations in the C-DVMH and Fortran-DVMH languages. These languages use a single DVMH-model of parallel programming model and are an extension of the standard C and Fortran languages with parallelism specifications in the form of compiler directives. The DVMH model makes it possible to create efficient parallel programs for heterogeneous computing clusters, in the nodes of which accelerators, graphic processors or Intel Xeon Phi coprocessors can be used as computing devices along with universal multi-core processors. The article presents new features of DVM-system that have been developed recently.

Keywords*: automation of development of parallel programs, DVM-system, accelerator, GPU, Fortran, C, irregular grid, unstructured grid*

REFERENCES

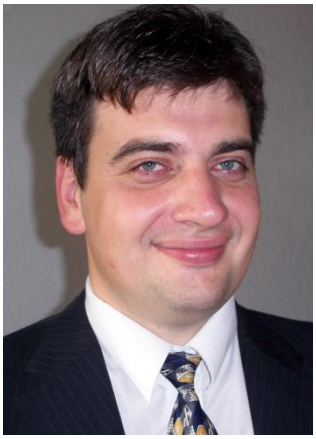
1. C-DVMH language, C-DVMH compiler, compilation, execution and debugging of DVMH programs. URL: http://dvm-system.org/static_data/docs/CDVMH-reference-en.pdf
2. Fortran DVMH language, Fortran DVMH compiler, compilation, execution and debugging of DVMH programs. URL: http://dvm-system.org/static_data/docs/FDVMH-user-guide-en.pdf
3. System for automating the development of parallel programs (DVM-system). URL: http://dvm-system.org
4. *Ladonkina M.E., Neklyudova O.A., Tishkin V.F.* Limiter povyshennogo poryadka tochnosti dlya razryvnogo metoda Galerkina na treugol'nyh setkah // Preprinty IPM im. M.V. Keldysha. 2013. No 53. 26 s.
5. *Ladonkina M.E., Neklyudova O.A., Tishkin V.F.* Issledovanie vliyaniya limitera na poryadok tochnosti resheniya razryvnym metodom Galerkina // Preprinty IPM im. M.V. Keldysha. 2012. No 34. 31 s.
6. *Bakhtin V.A., Zaharov D.A., Kolganov A.S., Krukov V.A., Podderyugina N.V., Pritula M.N.* Development of Parallel Applications Using DVM-system. Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering. 2019. V. 8. No 1. P. 89-106. (in Russian) DOI: 10.14529/cmse190106

СВЕДЕНИЯ ОБ АВТОРАХ

АЛЕКСАХИН Валерий Федорович – ведущий инженер ИПМ им. М.В. Келдыша РАН. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; параллельные алгоритмы.

Valery Fedorovich ALEKSAHIN – leading engineer of Keldysh Institute of Applied Mathematics. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms.

email: valex@keldysh.ru



БАХТИН Владимир Александрович – ведущий научный сотрудник ИПМ им. М.В. Келдыша РАН, доцент кафедры системного программирования факультета ВМК МГУ им. М.В. Ломоносова. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; параллельные алгоритмы; методы, средства и системы обработки данных большого объема.

Vladimir Aleksandrovich BAKHTIN – leading researcher of Keldysh Institute of Applied Mathematics, docent of the faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms; methods, tools and systems of large data processing.

email: bakhtin@keldysh.ru

ЖУКОВА Ольга Федоровна – научный сотрудник ИПМ им. М.В. Келдыша РАН. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; параллельные алгоритмы.

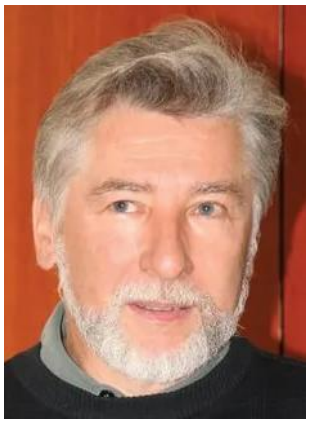
Olga Fedorovna ZHUKOVA – researcher of Keldysh Institute of Applied Mathematics. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms.

email: socol@keldysh.ru

ЗАХАРОВ Дмитрий Александрович – программист ИПМ им. М.В. Келдыша РАН. Сфера научных интересов – программные средства и системы для распределенных вычислений; параллельные алгоритмы; автоматизация параллельного программирования; распараллеливание программ, использующих неструктурные сетки.

Dmitry Aleksandrovich ZAKHAROV – programmer of Keldysh Institute of Applied Mathematics. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms; automatization of parallel programming; parallelization of unstructured grid applications.

email: s123-93@mail.ru



КРЮКОВ Виктор Алексеевич – главный научный сотрудник ИПМ им. М.В. Келдыша РАН, профессор кафедры системного программирования факультета ВМК МГУ им. М.В. Ломоносова. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; параллельные алгоритмы; методы, средства и системы обработки данных большого объема.

Victor Alekseevich KRUKOV – chief researcher of Keldysh Institute of Applied Mathematics, professor of the faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms; methods, tools and systems of large data processing.

email: krukov@keldysh.ru

ПОДДЕРЮГИНА Наталия Викторовна – старший научный сотрудник ИПМ им. М.В. Келдыша РАН. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; параллельные алгоритмы.

Nataliya Victorovna PODDERYUGINA – senior researcher of Keldysh Institute of Applied Mathematics. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms.

email: npodderyugina@gmail.com

САВИЦКАЯ Ольга Антониевна – научный сотрудник ИПМ им. М.В. Келдыша РАН. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; параллельные алгоритмы.

Olga Antonievna SAVITSKAYA – researcher of Keldysh Institute of Applied Mathematics. Research interests include mathematical software, software and systems for distributed computing; parallel algorithms.

email: savol@keldysh.ru

Материал поступил в редакцию 13 ноября 2019 года

УДК 004.65 + 005 + 001.5

ФОРМИРОВАНИЕ РАСШИРЕННЫХ ПОИСКОВЫХ ЗАПРОСОВ НА ОСНОВЕ ТЕЗАУРУСА ПРЕДМЕТНОЙ ОБЛАСТИ В ОНТОЛОГИИ ЗНАНИЙ СЕМАНТИЧЕСКОЙ БИБЛИОТЕКИ

О. М. Атаева[1], В. А. Серебряков[2], Н. П. Тучкова[3]

[1,2, 3]*Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Аннотация

Обсуждены возможности расширения поискового запроса при наличии тезауруса предметной области. Роль контекста, задаваемого связями терминов тезауруса, заключается как в уточнении запроса, так и в увеличении масштабов выборки по запросу. Особое значение процесс расширения запроса имеет для научных предметных областей, где поиск опирается на специальную терминологию. В этом случае необходимо использовать тезаурусы предметных областей, чтобы минимизировать появление информационного шума. Предлагаемый подход позволяет учитывать особенности применения аналогичной терминологии в различных предметных областях. Примеры использования тезауруса отдельных разделов уравнений математической физики и смежных областей демонстрируют эффективность выбранного подхода исследований. Благодаря связям с понятиями информационных ресурсов других областей знаний, расширение информационного запроса захватывает поисковые поля отдаленных предметных областей и различных типов данных, текстов, символьных, звуковых и видеоархивов. Исследования показали, что расширение запроса на основе семантики контекста улучшает качество поиска научных публикаций в цифровой информации и повышает эффективность научных междисциплинарных исследований.

Ключевые слова*: сравнение научных текстов, семантический поиск, тезаурус для онтологии знаний*, *информационный запрос с помощью тезауруса, семантические библиотеки.*

ВВЕДЕНИЕ

Исследования в области использования терминов тезауруса в поисковом запросе в контексте повышения эффективности информационных систем ведутся в различных коллективах. Известны многочисленные разработки для приложений в автоматическом реферировании, системах перевода специализированных текстов, обучающих системах и др. Важность и актуальность этих исследований определяются возрастающим потоком цифровых данных и разнообразием их типов, необходимостью работы с научной информацией, особенностями предметных областей (ПО) в междисциплинарных исследованиях. Особый круг задач рассматривается в проблеме *расширения поискового запроса*. Средства расширения запроса позволяют *уточнять запрос* с помощью подсказок пользователю, сужая поле поиска с помощью дескрипторов тезаурусов, и использовать имеющиеся связи терминов (синонимов, аббревиатур и т. д.), *увеличивая поле поиска* и получая тем самым дополнительный информационный шум. Эти два процесса находятся в противоречии, но в итоге приводят к получению пертинентного результата, то есть удовлетворяющего информационный запрос пользователя. Разработки в этом направлении ведутся довольно давно, и многие информационные системы допускают расширение запроса. В работе [1] приведены результаты, свидетельствующие, что *привлечение синонимов* из базы WordNet, *не связанных с контекстом*, не улучшают качество информационного запроса. И только привлечение технологии *прописывания «вручную» семантических связей* позволяет расширить запрос до полезного информационного поля, но, естественно, что таким образом не удастся охватить сколько-нибудь значительное количество связей. В итоге возникает необходимость сформулировать *задачу автоматического учета семантических связей*, что возможно *при наличии тезауруса, соответствующего тематике*. Особенную трудность уточнения и расширения информационного запроса представляет *процесс поиска научной информации*, поскольку основу для поиска составляет использование специальной терминологии и связей, задаваемых логикой ПО. Сложность составляет также *иерархическая система представления научных данных*, когда появляется *проблема установления горизонтальных связей между понятиями* [2]. На примере ПО задач математической физики и смежных областей предла-

гается показать, как расширение запроса на основе тезауруса LibMeta может улучшать результаты поиска.

1. ОСОБЕННОСТИ МЕТОДА РАСШИРЕНИЯ ИНФОРМАЦИОННОГО ЗАПРОСА

Расширение информационного запроса (Query expansion[1]) предполагает *переформулирование исходного запроса* с целью улучшения результата поиска. Этот процесс непосредственно связан с *пониманием* предмета поиска как со стороны пользователя (уровень компетентности в некоторой ПО), так и со стороны информационно-поисковой системы (наличие информационных и функциональных средств расширения и уточнения запроса).

Расширение запроса включает такие *методы*, как:

- поиск и использование синонимов для слов из запроса, а также поиск новых синонимов;
- поиск и использование семантических связей с другими словами; это могут быть, например, антонимы (противоположные по смыслу), меронимы (части слов), гипонимы (видовые понятия), гиперонимы (родовые понятия);
- поиск и использование всех различных морфологических форм слов из поискового запроса;
- фиксация ошибок правописания и автоматический поиск исправленной или предложенной словоформы;
- переназначение смысловой нагрузки слов в оригинальном запросе.

Последнее, а именно, «переназначение смысловой нагрузки», может оказать негативное влияние на результат поиска, если новая смысловая нагрузка не связана семантически с исходной ПО поискового запроса, что может привести к увеличению поискового шума.

В связи с расширением поискового запроса обсуждались также «термины расширения» (term expansion) и вопросы «улучшения запроса» (query enhancement). В работе [3] отмечено, что историю исследований расширения информационного запроса можно отследить, начиная с 1965 года, когда в работе [4] было дано формализованное описание релевантности результатов поискового запроса на основе векторной модели обратной связи (известное, как ал-

[1] https://en.wikipedia.org/wiki/Query_expansion

горитм Роккио — Rocchio algorithm). Более ранние исследования в области оценки веса связанных и не связанных терминов при расширении запроса принадлежат Спарку Джонсу (Spärck Jones) [6] и Ван Ризербергу (van Rijsbergen) [7]. Идея обратной связи по релевантности (Relevance Feedback — RF) заключается в привлечении пользователя к процессу поиска, чтобы улучшить итоговый список результатов. В частности, пользователь сообщает системе о релевантности документов в первоначальном списке результатов. Алгоритм Роккио — классический алгоритм для реализации метода RF. Он добавляет модель обратной связи по релевантности в модель векторного пространства [5]. Автоматическая генерация тезаурусов обсуждалась в работах Кью и Фрая (Qui and Frei) [8] и Шютце (Schütze) [9]. Использование локальных и глобальных методов расширения запросов исследовано в работах Крофта (Croft) и соавторов [10].

Эти работы заложили основу для дальнейших исследований в области расширения информационного запроса для текстовых документов в эпоху, когда еще не было достаточно инструментов для обработки символьной информации. В настоящее время получили развитие программные средства, позволяющие учитывать в базах данных формулы [11], стало возможно использовать формульную запись для расширения поискового запроса.

Предлагается подход, основанный на *учете смежных областей,* благодаря ассоциативным связям терминов тезауруса. Ранее была предложена *технология пополнения тезауруса адресата*, тезауруса по обыкновенным дифференциальным уравнениям и уравнениям смешанного типа [12, 13]. Это технология — для осведомленного пользователя. Если пользователь недостаточно знаком с ПО, то любая информация может быть (или не быть) *пополнением тезауруса адресата*. Расширение поискового запроса для такого пользователя может служить полезной подсказкой в процессе поиска. Рассматриваются варианты онтологии *одной ПО* и онтологий *различных ПО*.

1.1. Расширение запроса для одной предметной области

В качестве примеров продемонстрируем процесс поиска математических публикаций. Эти примеры характерны тем, что в тезаурусе математических ПО термины довольно часто сопровождаются формулами, символьным представлением терминов. На примере рис. 1 показано увеличение полей поиска науч-

ных публикаций за счет *связей терминологии смежных областей и переформулирования запроса*.

Известно, что одни и те же явления, встречающиеся в естественных науках, поддаются моделированию в различных областях знаний, при этом могут использоваться идентичные (аналогичные) символьные выражения. Например, «волновое уравнение» используется при моделировании различных технических процессов. Запись волнового уравнения практически везде одна и та же.

Рис. 1. Схема расширения запроса для смежных предметных областей

По цепочке связей тезауруса уравнений математической физики можно легко перейти к поиску из массивов литературы по одной ПО к другой. В примере на рис. 1 показано, как от формулировки запроса по теме «волновое уравнение» осуществляется переход к формулировке по теме «околозвуковая газовая

динамика». Другой пример: «уравнение Трикоми» из раздела «уравнений смешанного типа» также имеет многочисленные приложения — от описания задач «магнитогидродинамических течений» до задач «околозвуковой газовой динамики» из одной более общей ПО «уравнений математической физики» (УрМатФиз). Этих примеров можно найти неограниченное количество, поскольку УрМатФиз, как предметная область, появилась для моделирования физических и технических процессов, т. е. имеет множество приложений и смежных областей. Их информационные образы в поисковых системах могут быть охвачены, благодаря возможностям расширенных запросов.

1.2. Расширение запроса для различных предметных областей

Особое значение имеет процесс расширения запроса при интеграции большого объема информации из различных ПО и распределенных источников. Имея онтологии ПО, можно организовать поиск с расширением запроса в различных направлениях, задаваемых цепочками семантических связей.

В рамках разрабатываемой технологии на основе LibMeta [14] проводится интеграция данных из различных областей знаний, представленных в виде предметных онтологий. В частности, энциклопедические данные, интегрированные в систему, позволяют использовать ассоциативные связи терминов для расширения информационного запроса вплоть до обращения не только к смежным областям знаний. Схематично история и технология расширения информационно-поискового запроса при наличии онтологий различных ПО отражена на рис. 2.

Создание предметных тезаурусов и внедрение этих знаний в виде онтологий ПО позволяет предоставлять пользователям информационных систем *расширять поисковые запросы*. Таким образом, используя связи смежных областей, можно обеспечить переход к *поиску в различных типах* цифровых ресурсов из *различных ПО*. Этот подход реализуется для поиска среди объектов информационной системы, но и за ее пределами, в доступных для интеграции базах данных, где встречаются термины из расширенного запроса.

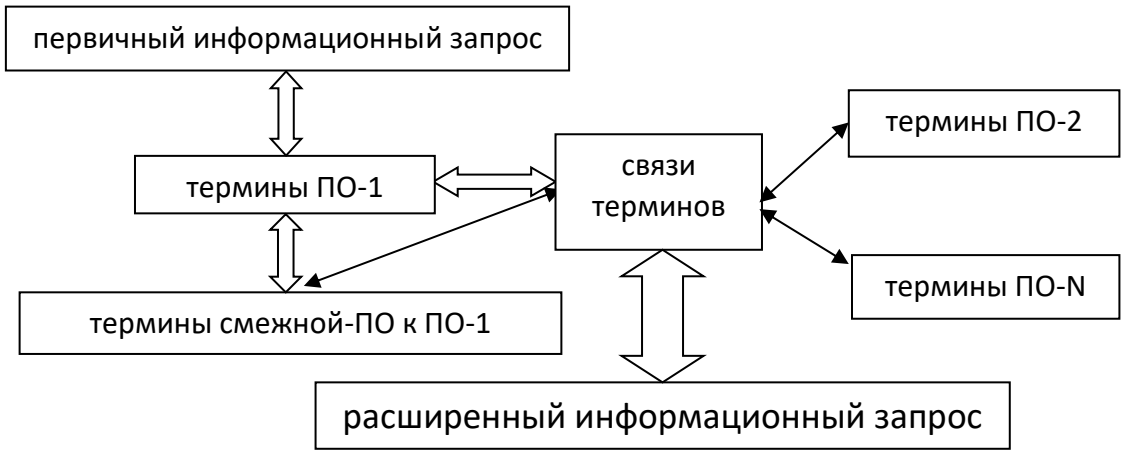


Рис. 2. Схема формирования расширенного информационного запроса для поиска в различных предметных областях

2. ПРЕИМУЩЕСТВА ИНТЕГРАЦИИ ДАННЫХ В КОНТЕКСТЕ РАСШИРЕНИЯ ЗАПРОСА

На первый взгляд, преимущества очевидны: чем больше охват запроса, тем больше информации в качестве результата получит пользователь интегрированной информационной системы. Тем не менее, известно также, что расширение запроса приводит к увеличению информационного шума, что никак нельзя отнести к преимуществам при поиске. Сочетание этих двух особенностей должно принимать некое «оптимальное» значений для того, чтобы услуга расширения поискового запроса составляла полезное свойство информационной системы.

Оптимальные свойства интегрированной системы, обеспечивающие *эффективность расширения информационного запроса*, реализуются, благодаря следующим особенностям:

- структуре данных;
- функциональным свойствам;
- возможности "настройки" на ПО пользователя.

2.1. Структура данных LibMeta

Онтология описывает ресурсы ПО и их взаимосвязи. Для каждой ПО LibMeta набор ресурсов может отличаться как по формату, так и по набору самих ресурсов.

Для описания библиотеки в LibMeta используются смысловой контент *конкретной* ПО и понятия, общие для *любой* из них, то есть предлагается набор по-

нятий, формирующих описание контента библиотеки, достаточно универсальный для включения в систему конкретной ПО. Такой подход позволяет реализовать средства интеграции данных в рамках библиотеки, адаптируемые под условия любой ПО с учетом ее специфики. Это позволяет решать одну из основных проблем интеграции данных из различных источников, а именно, согласование разнородной цифровой информации.

Понятия онтологии в системе LibMeta можно условно разделить по функциональному предназначению для следующих целей:

- описание контента ПО;
- формирование тезауруса любой ПО,
- описание тематических коллекций,

описание задачи интеграции контента библиотеки с данными из внешних источников.

Между этими группами понятий определены семантически значимые связи.

2.2. Функциональные особенности системы LibMeta

Семантическая библиотека LibMeta представляет собой информационную систему, в рамках которой задается описание ПО с терминологической поддержкой и возможностью интеграции данных из разных источников данных, удовлетворяющим требованиям, предъявляемым к источникам данных в LOD (Linked open data[2] [15]). Соответствие требованиям может быть *неполным,* и это означает, что, возможно, требование, касающееся *связанности данных* с другими источниками, может не выполняться, но с помощью LibMeta появляется возможность достаточно просто выполнить его. Для этого от пользователя – эксперта в ПО – не требуется специальных технических знаний об используемом для этого стеке технологий LOD.

Перечислим основную функциональность системы:

- создание/просмотр/редактирование информационных ресурсов и их структуры;
- создание/просмотр/редактирование информационных объектов и их структуры;

[2] https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

- подключение источников данных;
- загрузка данных из подключенных источников данных, в дальнейшем становящихся частью контента библиотеки;
- создание/просмотр/редактирование структуры тезауруса поддерживаемой ПО;
- создание/просмотр/редактирование понятий тезауруса
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным информационным объектам системы;
- атрибутный/семантический/полнотекстовый поиск по источникам данных;

создание/просмотр/редактирование коллекций информационных объектов.

2.3. Настройка на предметную область пользователя в LibMeta

Адаптация данных ПО трактуется как «настройка» источников, в которой можно выделить несколько основных этапов:

- *Подключение источника данных* S_i. Каждый источник данных характеризуется соответствующим уникальным *URL*-адресом и некоторым набором параметров, необходимых для доступа к данным. Проводится предварительный анализ доступной из источника информации, в частности, определяются *типы его ресурсов и их свойства*, участвующие в интеграции. Результатом этого первого этапа становится определение той части схемы источника S_i, по которой будут извлекаться данные.
- Определение *типов ресурсов библиотеки LibMeta*, соответствующих *типам ресурсов источников*. Для каждого ресурса источника, определенного его схемой, извлеченной на этом этапе, ставится в соответствие ресурс библиотеки LibMeta. Результатом этого этапа становится *установление связи* между ресурсом библиотеки и ресурсом источника с помощью соответствующей операции, которая декларирует, что существуют экземпляры этих ресурсов, соответствующие одному и тому же объекту *реального* мира. На базе определенных (выявленных) связей на следующем этапе проходит отображение атрибутов.
- Для каждого ресурса LibMeta определяется отображение атрибутов на соответствующие им свойства ресурса источника данных. В первую очередь строится отображение для идентифицирующих атрибутов, являющихся обяза-

тельными, затем для остальных. *Для каждой такой пары определяются тип связи и набор операций*.

Благодаря такому построению отображения получаем *набор правил*, по которым можно представить каждый *найденный объект в источнике* в рамках понятий библиотеки LibMeta и, соответственно, позволить сохранить его *метаданные в локальном хранилище* по требованию пользователя либо просто *сохранить связь* между найденным объектом в источнике и объектом в библиотеке.

2.4. Интеграция данных различных предметных областей

Формальная модель процесса интеграции данных из различных ПО может быть представлена следующим образом.

Исходя из основных понятий LibMeta, модель контента библиотеки G представляет собой:

- множество ресурсов $R = \{r_j\}$,
- множество атрибутов $A = \{a_i\}$,
- набор атрибутов $N(r) \subset A$, то есть $r_j(a_1,.., a_n)$, $a_n \in N(r)$, определенный для каждого ресурса.

В каждый набор атрибутов входят *идентифицирующие атрибуты*, $I(r) \subset N(r) \subset A$, используемые для однозначной идентификации информационных объектов этого ресурса.

Формально подсистема интеграции I_T представляется тройкой $<G,\{S_i\},\{M_i\}>$, где G – предварительно определенная модель контента, состоящая из множества ресурсов R и их описаний в виде набора атрибутов $N(r)$, S_i – схема i-го источника, подключенного к системе, M_i – отображение i-го источника, $1 \le i \le n$, где n – количество источников данных.

Использование источника данных может происходить по двум сценариям:

1. *в режиме проставления связей* с объектами, имеющимися в библиотеке,
2. *в режиме атрибутного поиска* по источнику данных в рамках заданного отображения.

При этом сохранение данных об объектах из источников может быть выполнено двумя способами:

1. *связывание* – этот способ идентичен по смыслу проставлению связи «смотри также» и означает, что на одном конце содержится более полная и обширная информация по ресурсу;
2. *идентификация* – этот способ идентичен по смыслу проставлению связи «такой же как» и означает, что на одном конце содержится точно такой же по качеству информации объект, как и с другой.

В связи с гибкостью модели контента библиотеки предполагается возможным сценарий создания дополнительных типов ресурсов для подключаемых источников, информацию из которых можно использовать как значения некоторых атрибутов основных ресурсов.

Схему ресурса библиотеки *G,* как источника данных *S,* так и контента, можно представить в виде графа (рис. 3), который включает *объекты* и *отношения*.

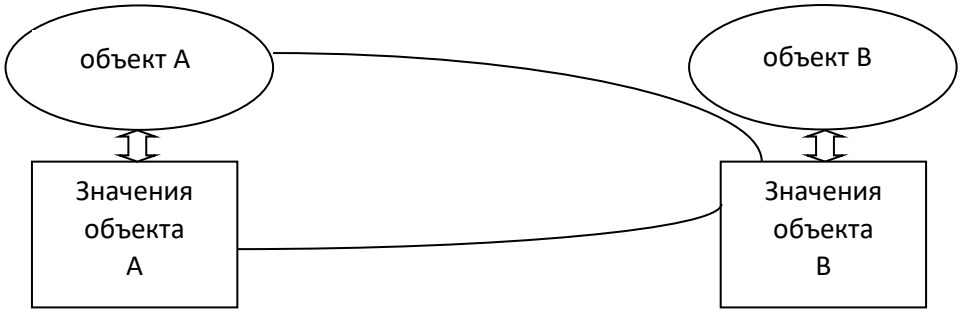


Рис. 3. На схеме линии представляют примеры связей «значение – значение» и «объект – значение» при отображении ресурсов из источников

Каждый объект может быть связан отношением с другим объектом, значения которого представлены простыми типами данных (*строки, числа, даты*) или отношениями с другими объектами, значения которых соответствуют некоторым ресурсам. При этом для *отображения ресурса* мы можем использовать *его представление* Z_s, то есть выбрать не полный набор его атрибутов и отношений с другими объектами для отображения на схему *G*. При этом *представление* Z_s должно обязательно включать в себя набор атрибутов, значения которых позволяют однозначно идентифицировать объект в системе.

Благодаря описанному подходу, структура тезауруса может гибко настраиваться для произвольных ПО. В этом смысле понятия тезауруса могут иметь *мультидисциплинарный характер и*, в силу указанных связей, содержать *указание на смежную область науки* или *явную ссылку на понятия тезауруса смежной ПО*. Также для каждого понятия могут указываться, например, соответствующий код УДК и/или любого другого рубрикатора науки и использоваться наряду с другими как *средство расширения запроса*. Это позволяет *уточнять семантику связанных ресурсов* и использовать ключевые слова экземпляров ресурсов и термины тезауруса как ключевые слова соответствующих рубрик используемых рубрикаторов. Помимо *основных понятий в тезаурусе можно ввести дополнительные категории*, поддерживающие возможность сохранения дополнительной информации. Для этого в тезаурусе могут вводиться связи с ресурсом библиотеки, а именно, в структуре понятия тезауруса могут быть предусмотрены дополнительные соответствующие атрибуты.

2.5. Пример эффективного расширения информационного запроса в LibMeta

Для поддержки поиска по формулам в системе было введено понятие *Формула,* которое позволяет хранить оригинальную строку формулы из того источника, откуда она получена. Строка может быть в формате Content MathML, Presentation MathML, LaTeX. При необходимости количество типов представления формулы в различных нотациях легко расширяется. Понятие *Формулы* связано отношениями с *информационными объектами,* составляющими контент семантической библиотеки и *понятиями* тезауруса. Таким образом, мы всегда можем построить сеть связей формулы как с понятиями тезауруса, так и с различными информационными объектами системы. Каждая формула может быть дополнена ключевыми словами. Ключевые слова могут проставляться как экспертом системы, так и добавляться автоматически, поступая вместе с формулой из ее источника, а также дополняясь ключевыми словами связанных объектов.

Рассмотрим механизм использования парадигматических связей на примере ПО «задачи математической физики для уравнений смешанного типа. Для понятия тезауруса «уравнения Трикоми» покажем преимущества уточнения запроса при использовании формул. Наиболее распространенная запись для уравнения Трикоми – это $ u_{xx} + u_{yy} = 0 $, остальные составляют, «с точки зре-

ния тезауруса», *формулы-синонимы* (так же, как и все записи, аналогичные приведенной с точностью до обозначений). Этой формулой индексирована, в частности, работа [16].

Попытаемся найти публикации, которые также посвящены этой тематике. Делаем поисковый запрос, содержащий выражение: u_{xx}+u_{yy}, которое является частью описания понятия тезауруса «уравнения Трикоми», и получаем список публикаций, связанных с задачей Трикоми, хотя сам термин в поисковом запросе не использовался. Поиск производится по данным, извлеченным из присоединенных источников. При этом при формировании запроса учитываются структура понятия тезауруса предметной области, ключевые слова, привязанные к этим понятиям, или осуществляется навигация по связям тезауруса для дальнейшего уточнения запроса.

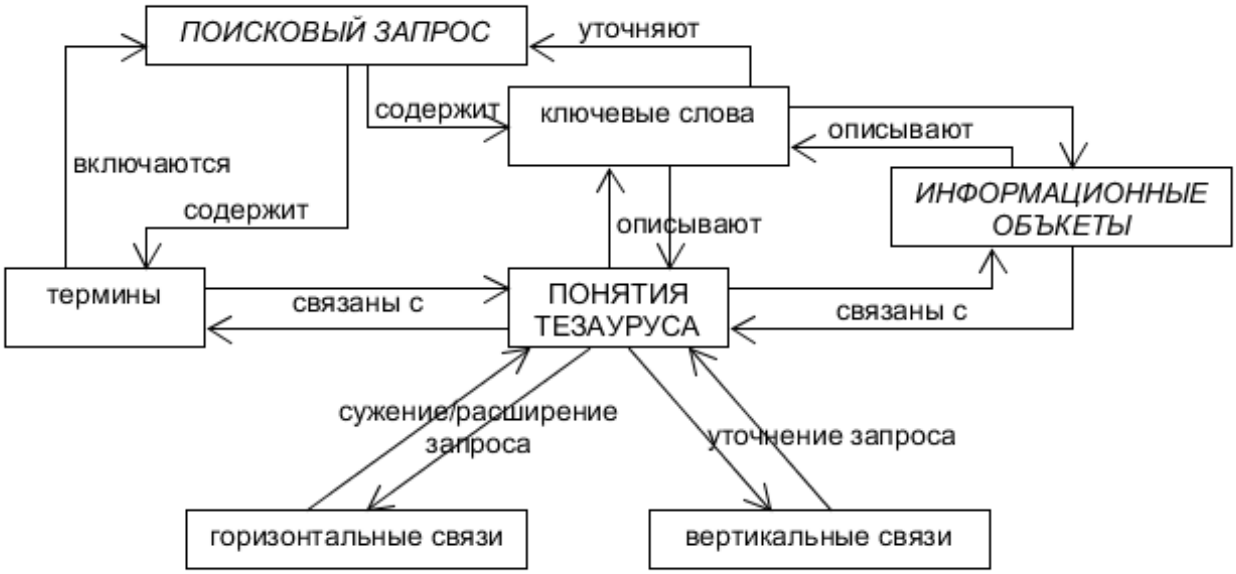


Рис. 4. На схеме представлены примеры формирования уточняющих запросов на основе понятий тезауруса и их связей

На рис. 4 схематически приведены пути формирования уточняющих запросов с использованием понятий тезауруса и его основных связей в системе LibMeta. При этом любой путь от «Поискового запроса» до «Информационного объекта» может оказаться достаточным для получения необходимого ответа на запрос. Ключевые слова могут использоваться не только в общепринятом

понимании, но, как было описано выше, в качестве них могут выступать, например, формулы.

В частности, ещё одно возможное расширение запроса основано на использовании соответствия между разделами MSC и других рубрикаторов, например, УДК. В случае успеха достаточно связать класс, соответствующий разделам MSC, с новым классом для разделов УДК, чтобы категоризовать понятия тезауруса и связанные с ними ресурсы по-новому. Подобного рода работа была проделана на основе использования понятий Математической энциклопедии в качестве тезауруса системы. Это позволило использовать внушительный объём знаний, содержащийся в Математической энциклопедии [17], и связи между ними для представления широкому кругу пользователей-любителей и экспертов в области математики, что особенно ценно при отсутствии в открытом доступе аналогичных ресурсов на русском языке.

В силу специфики данных нам не удалось найти источник данных в LOD для осуществления демонстрации возможностей LibMeta в области интеграции данных с такими источниками. Поэтому мы смоделировали эту ситуацию и выбрали в качестве гипотетического источника из LOD свой локальный источник данных. Наполнением этого источника является массив данных об авторах и публикациях из MathNet[3], который накопился у нас в рамках совместной работы. Этот массив хранится в виде RDF-троек[4].

В результате поиска, в том числе получаем список авторов, которые работают в этой области (например, Ю.М. Крикунов, Г.Л. Алфимов, Richard H., Cushman, Larry M. Bates и др.), а также возможность отследить семантическую сеть связей этой формулы в рамках тезауруса, а также в публикациях, представленных в библиотеке, и их авторов.

При необходимости можно расширять раздел ссылок соответствующей статьи-формулы указателя и расширить описание тезауруса. Ссылки на найденные публикации и авторов можно включить в статьи указателя, где встречаются формулы-синонимы, поиск по которым в данном случае не проводился. Так, через связи реализуется процесс пополнения тезауруса для ПО. В результате в

[3] http://www.mathnet.ru

[4] https://www.w3.org/RDF/

библиотеке LibMeta появятся новые данные о публикациях, и пользователь библиотеки при запросе получит новый список публикаций по теме «уравнения Трикоми». Делая запрос по этой теме, пользователь также получит полную информацию о семантических связях формулы, которая будет включать ссылки на формулы-синонимы, что особенно важно для специалиста.

ЗАКЛЮЧЕНИЕ

Развитие онтологического представления научных предметных областей способствует повышению эффективности поисковых запросов и научных исследований в целом. Учет ассоциативных связей терминов тезауруса позволяет делать выборку не только по смежным областям, но по цифровым массивам различных областей знаний, не увеличивая при этом поисковый шум. Эти выводы вполне ожидаемы, а проблемы, обсуждаемые в работе, актуальны с точки зрения объединения онтологий отдельных областей знаний. Сама эта проблема слияния онтологий представляет собой нетривиальную задачу, как с технологической, так и методологической точек зрения. Этот процесс может привести к качественному возрастанию времени обработки запроса и методологическим противоречиям, характерным для различных научных школ и направлений науки. В приведенных примерах в основном используются данные математических предметных областей, как характерные для расширения запроса за счет использования формул в смежных областях, что, естественно, не ограничивает расширение запроса на другие предметные области, интегрированные в LibMeta. В проекте реализованы связи с любыми источниками, удовлетворяющими требованиям LOD, и идут информационное наполнение и тестирование связей с лингвистической базой данных и математической энциклопедией. Исследования в данном направлении составляют предмет дальнейшей работы.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты № 17-07-00217а, 18-00-00297комфи, 17-07-00214.

СПИСОК ЛИТЕРАТУРЫ

1. *Voorhees E.M.* Query expansion using lexical-semantic relations. In SIGIR 94. ACM 1994. P. 61–69.

2. *Golden P., Shaw R., Buckland M.* Decentralized coordination of controlled vocabularies // Proceedings of the American Society for Information Science and Technology. Annual Meeting, October 31 – November 4, 2014, Seattle, WA, USA. 2014 DOI: 10.1002/meet.2014.14505101146 77th ASIS&T

3. *Vechtomova O.* Query Expansion for Information Retrieval. In: LIU L., ÖZSU M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston, MA. 2009 DOI: 10.1007/978-0-387-39940-9_947

4. *Salton G.* The SMART retrieval system (Chapter 14). Prentice-Hall, Englewood Cliffs NJ. (Reprinted from Rocchio J.J. (1965). Relevance feedback in information retrieval. In Scientific Report ISR-9, Harvard University), 1971.

5. *Маннинг К.Д., Рагбхаван П., Шютце Г.* Введение в информационный поиск. Издательский дом Вильямс. 528 с. ISBN 978-5-8459-1623-5.

6. *Spärck Jones K.* Automatic keyword classification for information retrieval. Butterworths, London, 1971.

7. *van Rijsbergen C.J.* A theoretical basis for the use of co-occurrence data in information retrieval // J. Doc. 1977. V. 33. No 2. P. 106–119.

8. *Qui Y., Frei H.* Concept based query expansion. SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval Pittsburgh, Pennsylvania, USA June 27 – July 01, 1993. ACM New York, NY, USA. P. 160–169. ISBN 0-89791-605-0. DOI:10.1145/160688.160713.

9. *Schütze H.* Automatic Word Sense Discrimination // Computational Linguistics, March 1998 – Special Issue on Word Sense Disambiguation. 1998. V. 24. No 1. P. 97–123. https://www.aclweb.org/anthology/J98-1004.pdf

10. *Larkey L.S., Croft W.B.* Combining classifiers in text categorization // SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland. August 18–22. 1996. 1996. P. 289–297. ISBN:0-89791-792-8 DOI: 10.1145/243199.243276.

11. Zentralblatt MATH https://zbmath.org

12. *Муромский А.А., Тучкова Н.П.* Об онтологии адресата в математической предметной области // Электронные библиотеки. 2018. Т. 21. № 6. С. 506–533.

13. *Моисеев Е.И., Муромский А.А., Тучкова Н.П.* О тезаурусе предметной области смешанные уравнения математической физики // CEUR Workshop Proceedings. 2018. V. 2260. P. 395–405. DOI: 10.20948/abrau-2018-43

14. *Атаева О.М., Серебряков В.А., Тучкова Н.П.* Подходы к организации математических знаний при формировании предметных тезаурусов различных разделов математики // CEUR Workshop Proceedings. 2018. V. 2260. P. 42–54. ISSN:1613-0073. DOI: 10.20948/abrau-2018-66.

15. *Bizer C., Heath T., Berners-Lee T.* Linked Data – The Story So Far // International Journal on Semantic Web and Information Systems. 2009. V. 5. No 3. URL: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. DOI:10.4018/jswis.2009081901.

16. *Моисеев Е.И., Лихоманенко Т.Н.* Собственные функции задачи Трикоми с наклонной линией изменения типа // Дифференциальные уравнения. 2016. Т. 52, № 10, С. 1375–1382.

17. *Виноградов И.М. (ред.).* Математическая энциклопедия: В 5-ти т. Сов. энцикл., 1979.

CREATION OF QUERY EXPANSION BASED ON THE SUBJECT DOMAIN THESAURUS IN THE ONTOLOGY OF KNOWLEDGE OF THE SEMANTIC LIBRARY

O.M. Ataeva[1], V.A. Serebriakov[2], N.P. Tuchkova[3]

[1,2,3]*Dorodnicyn Computing Centre FRC CSC RAS, Moscow*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract

Possibilities of query expansion with subject area thesaurus are discussed**.** The role of the context defined by thesaurus term links is both to refine the query and to increase the size of the sample on the query. Of particular importance is the process of expanding the query for scientific subject areas where the search based on special terminology. In this case, thesauruses of subject areas must be used to minimize the occurrence of information noise. The proposed approach takes into account the application of similar terminology in various subject areas. Examples of the use of thesaurus of separate sections of equations of mathematical physics and related fields demonstrate the effectiveness of the chosen approach of research. By linking to concepts of information resources of other areas of knowledge, the extension of the information query captures search fields of remote subject areas and various types of data, texts, symbolic, audio and video archives. Research shows that expanding the query based on context semantics improves the search quality of scientific publications in digital information and increases the effectiveness of scientific interdisciplinary research.

Keywords*: comparison of scientific texts, semantic search, thesaurus for the ontology of knowledge, information query using the thesaurus, LibMeta*

REFERENCES

1. *Voorhees E.M.* Query expansion using lexical-semantic relations. In SIGIR 94. ACM 1994. P. 61–69.

2. *Golden P., Shaw R., Buckland M.* Decentralized coordination of controlled vocabularies // Proceedings of the American Society for Information Science and

Technology. Annual Meeting, October 31 – November 4, 2014, Seattle, WA, USA. 2014 DOI: 10.1002/meet.2014.14505101146 77th ASIS&T

3. *Vechtomova O.* Query Expansion for Information Retrieval. In: LIU L., ÖZSU M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston, MA. 2009 DOI: 10.1007/978-0-387-39940-9_947

4. *Salton G.* The SMART retrieval system (Chapter 14). Prentice-Hall, Englewood Cliffs NJ. (Reprinted from Rocchio J.J. (1965). Relevance feedback in information retrieval. In Scientific Report ISR-9, Harvard University), 1971.

5. *Manning C.D., Raghavan P., Schütze H.* Introduction to Information Retrieval, Cambridge University Press. 2008. 544 p. ISBN: 0521865719. Online edition (c) 2009 Cambridge UP. URL: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

6. *Spärck Jones K.* Automatic keyword classification for information retrieval. Butterworths, London, 1971.

7. *van Rijsbergen C.J.* A theoretical basis for the use of co-occurrence data in information retrieval // J. Doc. 1977. V. 33. No 2. P. 106–119.

8. *Qui Y., Frei H.* Concept based query expansion. SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval Pittsburgh, Pennsylvania, USA June 27 – July 01, 1993. ACM New York, NY, USA. P. 160–169. ISBN 0-89791-605-0. DOI:10.1145/160688.160713.

9. *Schütze H.* Automatic Word Sense Discrimination // Computational Linguistics, March 1998 – Special Issue on Word Sense Disambiguation. 1998. V. 24. No 1. P. 97–123. https://www.aclweb.org/anthology/J98-1004.pdf

10. *Larkey L.S., Croft W.B.* Combining classifiers in text categorization // SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland. August 18–22. 1996. 1996. P. 289–297. ISBN:0-89791-792-8 DOI: 10.1145/243199.243276.

11. Zentralblatt MATH https://zbmath.org

12. *Muromskij A.A., Tuchkova N.P.* Ob ontologii adresata v matematicheskoj predmetnoj oblasti // Elektronnye biblioteki. 2018. T. 21. No 6. S. 506–533.

13. *Moiseev E.I., Muromskij A.A., Tuchkova N.P.* O tezauruse predmetnoj oblasti smeshannye uravneniya matematicheskoj fiziki // CEUR Workshop Proceedings. 2018. Vol. 2260. P. 395–405. DOI: 10.20948/abrau-2018-43.

14. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* Podhody k organizacii matematicheskih znanij pri formirovanii predmetnyh tezaurusov razlichnyh razdelov matematiki // CEUR Workshop Proceedings. 2018. Vol. 2260. P. 42–54. ISSN:1613-0073. DOI: 10.20948/abrau-2018-66.

15. *Bizer C., Heath T., Berners-Lee T.* Linked Data – The Story So Far // International Journal on Semantic Web and Information Systems. 2009. V. 5. No 3. URL: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. DOI:10.4018/jswis.2009081901.

16. *Moiseev E.I., Lihomanenko T.N.* Sobstvennye funkcii zadachi Trikomi s naklonnoj liniej izmeneniya tipa // Differencial'nye uravneniya. 2016. T. 52, № 10, S. 1375–1382.

17. *Vinogradov I.M.* Matematicheskaya entsiklopediya [Mathematical Encyclopedia] //Moscow, Sovetskaya entsiklopediya Publ. 1979.

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – researcher of the of Dorodnicyn computing center FRC SCS RAS, expert in the field of system programming and databases.

email: oli@ultimeta.ru

СЕРЕБРЯКОВ Владимир Алексеевич – специалист в области теории формальных языков и её приложений, доктор физ.-мат. наук, профессор, зав. отделом Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН. Руководитель и участник разработки ряда известных программных проектов, в частности, ИСИР РАН, «Научный портал РАН».

Vladimir Alekseevich SEREBRIAKOV – expert in the field of theory of formal languages and its applications, doctor of sciences, professor, head of Dorodnicyn computing center FRC SCS RAS department. Head and participant in the development of a number of well-known program projects, in particular, ISIR RAS, Scientific portal RAS.

email: serebr@ultimeta.ru

ТУЧКОВА Наталия Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук, окончила ВМиК МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU. The expert in the field of algorithmic languages and information technologies.

email: natalia_tuchkova@mail.ru

Материал поступил в редакцию 15 ноября 2019 года

УДК 004.55 + 004.21 + 004.4

СОЗДАНИЕ ИНСТРУМЕНТАЛЬНОЙ ПЛАТФОРМЫ МУЛЬТИМЕДИЙНОГО НАУЧНОГО ЖУРНАЛА

Н. В. Борисов[1], В. В. Захаркина[2], И. А. Мбого[3], Д. Е. Прокудин[4], П. П. Щербаков[5]

Санкт-Петербургский государственный университет, г. Санкт-Петербург

[1]nikborisov@gmail.com, [2]zakharkina@gmail.com, [3]irina.mbogo@gmail.com, [4]hogben.young@gmail.com, [5]paul.tscherbakov@gmail.com

Аннотация

Обсуждены подходы к созданию инструментальной платформы электронного научного журнала, обеспечивающей публикацию мультимедийных материалов через веб-интерфейс. Описаны проблемы, связанные с необходимостью включения мультимедиа данных различных типов, и представлен рабочий прототип мультимедиа научного журнала.

Ключевые слова: *научная публикация, электронный научный журнал, мультимедиа контент, электронная коллекция*

ВВЕДЕНИЕ

В настоящее время результаты научных исследований, в наиболее адекватном виде, часто представляются в мультимедийной форме. В качестве примера можно упомянуть электронные коллекции мультимедийных материалов, которые формируются в процессе обработки результатов гуманитарных исследований археологов, фольклористов, этнографов и др. Существенными для публикации результатов подобных исследований в интернете могут быть возможности отображения материалов различных медийных форматов: фото, видео, звук и другие. На сегодняшний день практически все научные журналы имеют свое электронное представление. Часть журналов является полностью электронной (сетевой). Это ставит проблему обеспечения технологических возможностей создания научных публикаций с мультимедийным контентом.

С точки зрения технологической реализации онлайн-представления статей к мультимедийным могут быть отнесены и сущности, характерные для публикаций в целом ряде предметных областей. Это, например, формулы, диаграммы,

графики и т. д. Адекватное отображение подобных сущностей в онлайн-статье требует разработки соответствующих программных модулей онлайн-журнала. При этом необходимо обеспечить либо импорт сущностей, созданных при помощи сторонних приложений (при этом возникает проблема совместимости форматов), либо создание соответствующего инструментария на базе инструментальной платформы онлайн-журнала.

На сегодняшний день практически все научные журналы имеют свое электронное представление. Часть журналов является полностью электронными (сетевыми). Это ставит проблему обеспечения технологических возможностей создания научных публикаций с мультимедийным контентом.

Современная научная статья помимо традиционных иллюстраций может быть обогащена динамическими объектами – 3D-моделями, видео, видео 360, галереями изображений, интерактивными элементами виртуальной реальности, диаграммами и графиками с элементами интерактивности, а также другими.

1. ВЕБ-ПУБЛИКАЦИИ

Для обеспечения долгосрочного хранения информационных ресурсов, обмена метаинформацией и др. создан ряд информационных систем (ИС), используемых научным и образовательным сообществами. Приведем несколько примеров систем, используемых для электронных научных публикаций:

1. Система поддержки цифровых хранилищ (институциональных репозиториев) широко используется для построения архивов открытого доступа и электронных библиотек, позволяющих создавать, хранить и распространять цифровые материалы. К ним относятся такие открытые программные платформы, как DSpace, EPrints, GreenStone, Fedora и другие. Нам представляется, что DSpace – самое популярное в академической среде программное обеспечение для создания архива электронных ресурсов (цифрового репозитория). Платформа DSpace разрабатывалась совместно компанией Hewlett-Packard и библиотеками MIT (Massachusetts Institute of Technology) [1, 2];

2. Некоторые электронные научные журналы используют электронные издательские системы, такие, как Open Journal Systems (OJS). Основным достоинством такого подхода является реализация полного издательского цикла для подготовки электронных публикаций. При этом формат научной публикации

остается традиционным. Она может быть представлена в двух форматах – PDF и статический HTML. Следует отметить, что HTML-верстка проводится вне системы, а в OJS загружаются уже готовые файлы. В этой системе не предусмотрена реализация процессов вёрстки, публикации и отображения мультимедийного контента. Отсутствие соответствующих веб-интерфейсов переводит процессы вёрстки и публикации в ручной режим, который подразумевает наличие соответствующих знаний и умений у специалистов, занятых в редакционно-издательском процессе;

3. Вопросы интеграции решаются на базе технологии обмена метаданными, основанной на протоколе OAI-PMH;

4. Развивается концепция «живых публикаций». Живая публикация – размещенная в интернете в свободном доступе научная работа, которая постоянно поддерживается ее автором в актуальном состоянии [3];

5. Начиная с 2014, несколько журналов издательства Elsevier начали публиковать отзывы рецензентов вместе со статьей. Такой подход, называемый открытым рецензированием (Open review), приводит к улучшению качества статей, признается вклад рецензента в публикационный процесс [4].

2. ОБЗОР ВОЗМОЖНОСТЕЙ МУЛЬТИМЕДИЙНЫХ ЖУРНАЛОВ

С начала 2000-х годов активно начинают развиваться такие сетевые электронные ресурсы, как электронные журналы. Электронные журналы публикуют статьи как в формате PDF, так и в HTML, который предоставляет больше возможностей для публикации, в том числе, и мультимедийных материалов. Ограничиваясь анализом журналов научной направленности, можно представить следующие выводы:

— приём редакцией текстов статей и сопутствующих мультимедийных материалов осуществляется посредством электронной почты;

— редакции просят указывать место в тексте статьи, куда необходимо вставить соответствующий мультимедийный материал, т. е. авторы непосредственно не принимают участие в вёрстке статьи;

— ни один из проанализированных нами журналов не публикует полный спектр мультимедийных файлов; основными публикуемыми форматами являются: графические файлы различных форматов (например, JPG, GIF, PNG),

видео (например, AVI, MPG, MOV), анимация (SWF, GIF), аудио различных форматов, презентации; есть отдельные журналы, публикующие, кроме всего перечисленного спектра мультимедийных форматов, виртуальную реальность в формате VRML (журнал «Научная визуализация»[1]);

— для отображения видео в статьях используются различные видео проигрыватели или встраиваются видео, размещённые в соответствующих облачных видео сервисах;

— сайты журналов либо являются оригинальными разработками, либо представлены на платформах издательств.

Наибольший спектр отображаемых мультимедийных материалов публикует журнал «Научная визуализация». Однако даже на страницах этого журнала для просмотра некоторых мультимедийных объектов (например, 3D[2]) требуется использование дополнительного программного обеспечения, установленного на компьютер читателя.

Несомненный интерес представляют исследования, проводимые в Институте прикладной математики им. М.В. Келдыша Российской академии наук по представлению анимации и видео в научной публикации [5]. В «Препринтах ИПМ им. М.В. Келдыша» некоторые статьи сопровождаются видео аннотациями на страницах метаданных публикации (например, [3]). В рамках этих исследований решены задачи встраивания видео в статью в формате PDF.

3. ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ МУЛЬТИМЕДИЙНОЙ СТАТЬИ

Мультимедийная статья, опубликованная в глобальной сети и отображаемая в браузере, может дать читателю возможности, качественно отличающие её от статьи в традиционном «печатном» формате. Принципиально возможна публикация видео- и звуковых фрагментов, 3D-объектов и сцен с элементами интерактивности, интерактивных схем и диаграмм.

Даже, казалось бы, простейший медийный элемент – изображение – при публикации в формате веб-документа (HTML) может обрести новое качество. Типовые программные модули с открытым кодом позволяют реализовать уве-

[1] http://sv-journal.org/
[2] http://sv-journal.org./example/index.html
[3] http://library.keldysh.ru/preprint.asp?id=2013-51

личение изображения, первоначально представленного в виде миниатюры, что обеспечивает комфортное восприятие. В ряде случаев иллюстрации имеют такие пропорции, которые в печатном формате вообще не позволяют их адекватно представить. Например, высота японских свитков-какэмоно в 2–2,5 раза превышает их ширину, что заставляет в печатной статье ограничиться публикацией фрагментов. В браузере эта проблема может быть решена с помощью программных модулей, обеспечивающих интерактивный просмотр фрагментов изображения высокого разрешения (рис. 1).

Рис. 1. Интерактивный просмотр фрагментов свитка-какэмоно

Новое качество отображения могут получить и иные традиционные элементы печатной статьи. Объёмные таблицы и схемы, не ограниченные размерами печатного листа, могут быть показаны с использованием горизонтальной прокрутки. Текст сноски может отображаться во всплывающем блоке при наведении курсора мыши, при этом у читателя не смещается фокус внимания. Возможность воспроизведения видеофрагмента или звука, представление 3D-сцен и т. д. позволяет автору донести свою идею на качественно новом уровне. Впрочем, подготовка мультимедийных материалов, естественно, требует от него дополнительных усилий.

4. ПРЕДСТАВЛЕНИЕ 3D-СЦЕН И ВИДЕО 360 В НАУЧНОЙ СТАТЬЕ

Одной из важных задач поддержки мультимедийных коллекций является расширение перечня поддерживаемых типов мультимедийных ресурсов. Наиболее распространенные типы ресурсов, такие, как тексты, изображения, давно получили поддержку в языке разметки веб-страниц, позволяющей отображать тексты, их фрагменты, цитаты из них, выделяя соответствующие объекты как на уровне оформления и представления, так и на логическом уровне, обеспечивая идентификацию, мета-описание, поисковую поддержку и другие возможности. Поддержка отображения pdf-документов в современных браузерах стала настолько прозрачной, что удобнее открыть такой виртуальный документ в браузере, чем в специализированном приложении Adobe Acrobat. Благодаря распространению стриминговых сервисов и видеохостингов, представление аудио и видео ресурсов на веб-страницах стало сводиться к вставке короткого фрагмента html-кода [10], предоставляемого владельцем ресурса, а создание виртуальных галерей, подборок, подкастов (фактически, составление горизонтально интегрированных мультимедийных коллекций) стало для социальных сетей обычным явлением.

Появление нового типа мультимедийного ресурса – видео в формате 360 градусов – повлекло сначала создание специализированных плейеров, позволяющих просматривать соответствующий контент в специализированном приложении, браузере, с использованием специальных гарнитур виртуальной реальности, на мобильных устройствах, затем появление возможности просмотра в универсальных плейерах, указывая в качестве параметра тип ресурса, затем встраивание в файл ресурса мета-данных, позволяющее плейерам автоматически распознавать тип ресурса и формировать элементы управления просмотром, подходящие для соответствующей ситуации (рис. 2.) На представленных скриншотах демонстрации ресурса видео 360 на youtube.com, в левом верхнем углу экрана помещается элемент управления виртуальной камерой, позволяющей изменять направление виртуальной камеры.

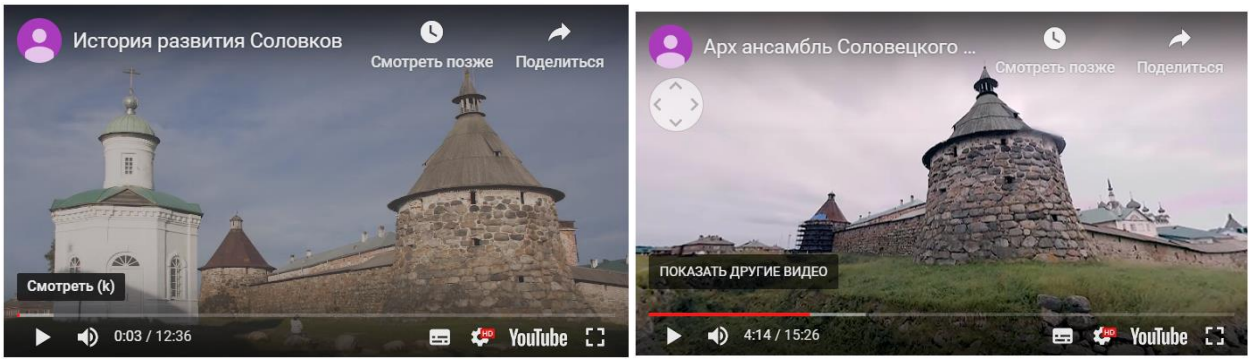


Рис. 2. Представление на YouTube видео различных форматов с изменением элементов управления просмотром

Использование 3D-моделей и интерактивных 3D-сцен для содержательной иллюстрации в документах привело к появлению автономных средств отображения таких объектов, в том числе и через интернет. Однако сложность и разнообразие инструментов создания таких ресурсов затрудняют их широкое использование в качестве элементов мультимедийных коллекций. Необходимость установки специализированного программного обеспечения на компьютер пользователя в общем случае не решает проблему отображения ресурса. Так, созданная в 3DS-max сцена может быть экспортирована в различные форматы (в том числе, и предназначенные для обмена 3D-моделями), однако при экспорте многие особенности модели будут утеряны, так как специфические модификаторы среды разработки не поддерживаются такими форматами. Более того, при использовании специальных плагинов, особенно коммерческих, воспроизведение сцены на компьютере пользователя может оказаться невозможным. Кроме того, для правильного воспроизведения сцены необходимо сохранить структуру проекта, содержащего для сложных сцен десятки каталогов и сотни, может быть, и тысячи отдельных файлов, причем в проекте могут сохраняться абсолютные пути к используемым ресурсам, так что даже при копировании всего дерева проекта в другое место правильное воспроизведение сцены может нарушиться.

Авторам представляется, что решение задачи полноценного использования мультимедийных ресурсов, типа 3D-моделей и интерактивных 3D-сцен, в качестве элементов мультимедийной коллекции может быть найдено, например, одним из следующих способов.

1. Использование специализированных веб-сервисов, позволяющих автору загрузить свой проект на сервер, при необходимости отредактировать и проверить адекватность его отображения и представить код для его внедрения в веб-страницу (как это сделано, например, на youtube.com). В качестве такого сервиса может, например, выступить сервис SketchFab[4]. Недостатком такого решения можно считать то, что ресурс хранится на внешнем сервисе, и возникновение проблем с аккаунтом владельца может сделать ресурс недоступным.

2. Экспорт проекта в формат WebGL. Последнее время поддержка этого формата браузерами постоянно улучшается, кроме того, имеются среды разработки, позволяющие реализовать или отредактировать 3D-проект непосредственно в формате WebGL. 3D-проект, подготовленный в такой технологии, может быть представлен непосредственно на веб-странице средствами исключительно веб-браузера. В качестве среды разработки, например, можно использовать среду, поддерживаемую сообществом ThreeJS[5].

5. ПРЕДСТАВЛЕНИЕ ФОРМУЛ В НАУЧНОЙ СТАТЬЕ

Немаловажным и востребованным компонентом научной статьи является наличие математического контента. При включении математической нотации в текст необходимо учитывать два аспекта – это написание формулы и их дальнейшее отображение.

Стандартом де-факто при вводе математических уравнений является TeX, который чаще всего используется в виде своего расширения LaTeX. Многие ведущие издательства, такие, как Nature, Elsevier, SAGE, принимают статьи именно в этом формате. В связи со сложностью написания команд TeX в текстовом редакторе и популярностью формата существуют специализированные системы подготовки таких документов. Рынок LaTeX-редакторов довольно большой, и все лучшие инструменты предлагают схожие функции с небольшими различиями. Можно отметить несколько лучших онлайн-платформ:

— Overleaf[6] – инструмент для написания статьи и совместной работы, который работает на LaTeX. Ключевой функцией системы является совместная

[4] https://sketchfab.com/
[5] http://threejs.org
[6] https://www.overleaf.com/

работа в режиме реального времени. Существуют варианты использования как LaTeX, так и режима расширенного текста, отображающего форматирование текста и графическое представление уравнений непосредственно в редакторе. Ценовая политика имеет бесплатный тарифный план.

— Authorea[7] – этот редактор известен как «Документы Google для ученых». Редактор позволяет писать рукопись на языках LaTeX и Markdown. Этот инструмент также помогает опубликовать статью в нескольких журналах с открытым доступом. Бесплатная версия сервиса обеспечивает неограниченный доступ к публичным документам. Authorea имеет встроенный поиск цитат, который позволяет искать и добавлять любую цитату из PubMed, CrossRef или путем вставки DOI. Бесплатный тарифный план предполагает ограниченное количество функций.

— Papeeria[8] – это простая в использовании онлайн-платформа для написания и редактирования LaTeX, имеющая очень мощный редактор. Он широко известен в сообществе LaTeX благодаря своей простоте и исключительно чистому пользовательскому интерфейсу. Это также один из немногих инструментов, которые поддерживаются на Android и iOS. Существует бесплатная версия, предоставляющая большинство функций.

— Lyx[9] – это процессор документов, который в основном работает над концепцией WYSIWYM (то, что вы видите, это то, что вы имеете в виду). Основное внимание уделяется структуре документа, а не только тому, как он будет выглядеть. Lyx имеет один из лучших редакторов математических формул. Уравнения можно вводить через графический интерфейс или введя команды LaTeX. Lyx – бесплатное программное обеспечение с лицензией Open Source.

Второй стороной использования формул в статье, опубликованной в научном on-line журнале, является их корректное отображение в браузере. Существует несколько подходов к решению этой проблемы:

— Отображение формул в виде растрового изображения (GIF или PNG). Такой подход хоть и является наиболее простым, но полностью стирает смысл формулы, исключает возможность ее дальнейшего анализа и считается малоэф-

[7] https://www.authorea.com/
[8] https://papeeria.com/
[9] https://www.lyx.org/

фективным и устаревшим. Небольшим улучшением ситуации может служить публикация LaTeX-формулы в качестве альтернативного текста к изображению.

— Конвертация LaTeX-формулы в векторный формат SVG.

— Использование специализированного языка математической разметки MathML. MathML предназначен для облегчения использования и обмена данными математической нотации в интернете (подробнее о MathML см., например в Wikipedia[10]). MathML является приложением XML и может обрабатываться и отображаться браузером. Однако не все браузеры имеют поддержку MathML, что влечет за собой необходимость использования дополнительных библиотек. Особого внимания заслуживает кросс-браузерная JavaScript библиотека MathJAX[11].

— Актуальным является использование автоматической конвертации из LaTeX в MathML. Реализовать такое преобразование можно как на стороне клиента, так и на стороне сервера. Существует довольно много математических инструментов, решающих эту задачу [12], из которых, с точки зрения онлайн журнала, интересны клиентские JavaScript библиотеки. Основными считаются fmath, jqMath и MathJax. MathJax может отображать математические обозначения, написанные в разметке LaTeX или MathML, а, начиная с версии 2.0, представлен рендеринг в SVG.

Реализация конвертации на стороне сервера представляется более затруднительной в связи с необходимостью установки дополнительных библиотек на уровне операционной системы, что зачастую невозможно осуществить на виртуальном хостинге, таким образом, теряется универсальность подхода.

Одним из вариантов решения задачи полноценного использования формул в журнале с мультимедиа контентом, на взгляд авторов, может являться следующий ряд подходов:

— использовать формат TeX/LaTeX для формул;

— рекомендовать авторам статей проводить предварительную подготовку формул к публикации с помощью мощных сторонних инструментов, таких, как Overleaf, Authorea, Papeeria или Lyx;

[10] https://ru.wikipedia.org/wiki/MathML
[11] https://www.mathjax.org/
[12] https://www.w3.org/wiki/Math_Tools

— отображать формулы в браузере с использованием MathML и в качестве инструмента кросс-браузерного отображения и инструмента автоматической конвертации из LaTeX в MathML можно предложить использовать MathJAX.

6. СТРУКТУРА МУЛЬТИМЕДИА-СТАТЬИ

Традиционная научная статья имеет сложившийся стиль оформления – текст и различные статические объекты (изображения, формулы и другие статические объекты), которые отображаются браузером без дополнительной обработки, например, изображения, таблицы, формулы и графики также в виде изображений. Визуальные элементы вставляются по ходу письма и снабжаются подписями. При проектировании мультимедийного журнала коллектив не стал отходить от традиций оформления, и мультимедиа объекты также должны встраиваться по ходу текста.

Следует особо подчеркнуть, что для различных типов мультимедиа недостаточно только средств браузера. Видео показывается с помощью плеера, для отображения видео 360 необходим соответствующий плагин к видеоплееру, для показа изображений в виде слайдера необходимо привязать слайдер к конкретному набору изображений, демонстрация изображений в виде галереи требует привязки других изображений к галереям и т. д. Таким образом, при формировании онлайн-статьи требуются не только различные обработчики: видео- и аудиоплееры, скрипты для слайдеров и галерей, но также необходимы инструменты привязки обработчиков к мультимедиа объектам. Дополнительной сложностью является то, что никогда заранее неизвестно, какие, в каком количестве и в каком порядке будут использоваться мультимедиа материалы.

Заранее неизвестная структура статьи не позволяет использовать в интерфейсе редактора просто набор фиксированных полей, так как в этом случае каждая статья будет формироваться по единому шаблону, с одинаковым порядком следования элементов. Для реализации более гибкого подхода к формированию структуры статьи были разработаны механизмы, позволяющие многократно использовать элементы статьи в любых порядке и количестве. Кроме этого, к каждому полю, описывающему свой объект, привязывается свой обработчик.

Важным является акцент на том, что медийные элементы должны быть представлены в виде отдельных сущностей на уровне базы данных. Это даёт

гибкие возможности их представления на веб-странице и обеспечивает эффективную работу технического редактора. Более того, при таком подходе не только статья в целом, но и её отдельные медийные элементы могут быть снабжены метаинформацией для потенциальных агрегаторов.

Совершенно очевидно, что «расширенная» публикация может быть просмотрена только в браузере, поэтому в этой ситуации первичной следует считать именно версию статьи в формате веб-документа. Однако электронная публикация может также иметь и свое статическое (бумажное) воплощение, где все динамические объекты конвертированы в изображения.

Работающий прототип электронного научного журнала с мультимедийным научным контентом, основанный на описанных подходах, реализован авторами для журнала «Культура и технологии»[13] на базе свободно распространяемой CMF Drupal [6]. В ходе реализации использованы административные интерфейсы Drupal, готовые модули, а также дополнительные модули на основе API.

7. ИНСТРУМЕНТЫ ОНЛАЙН-РЕДАКТИРОВАНИЯ МУЛЬТИМЕДИА-СТАТЬИ

Наиболее существенная проблема состоит в разработке инструментария технического редактора. Не обладая специфической квалификацией и навыками, редактор должен иметь эффективную возможность вёрстки статьи через простой и интуитивно понятный веб-интерфейс путём вставки в соответствующие поля текстовых фрагментов и загрузки файлов изображений, видео и т. д.

В варианте традиционной статьи может быть достаточно JavaScript-редактора, такого, как TinyMCE или CKEditor. Эти редакторы позволяют в стиле MSWord форматировать текст и вставлять изображения непосредственно в браузере. С помощью этих редакторов можно вставить и некоторые мультимедиа элементы, например, видео, элементы внешних мультимедиа-хранилищ, таких, как YouTube, и некоторые другие элементы. В предлагаемом подходе функциональности таких редакторов недостаточно, так как:

— в концепции журнала каждый мультимедиа-объект должен представлять собой отдельную сущность на уровне полей; агрегаторы могут рассматривать не только статью в целом, но и её отдельные медийные элементы, которые (опционно) имеют свою дополнительную метаинформацию; эти элементы

[13] http://cat.ifmo.ru

могут быть, например, отобраны по запросу для представления во внешних коллекциях с сохранением привязки к статье [6]; при использовании же встроенного редактора статья остается монолитной;

— возможности по загрузке файлов через веб-интерфейс сильно ограничены;

— в процессе создания статьи контент-менеджером необходимо подключать плагины видео плеера, JS-скрипты слайдеров, галерей и другие обработчики, что может потребовать дополнительных компетенций в веб-разработке.

Исходя из указанных предпосылок, были сформированы требования к инструментам редактирования статьи с мультимедиа-материалами:

1. Все элементы статьи разбиваются на поля: фрагмент текста, галерея изображений, видео, видео 360;
2. Должна существовать возможность вставить в статью любое количество мультимедиа-объектов, поэтому все поля должны быть множественными;
3. Необходим инструмент, позволяющий вставлять мультимедиа объекты в текст статьи и привязывать соответствующий обработчик.
4. Реализовать несколько вариантов демонстрации мультимедийных объектов, в частности, для изображений, кроме вставки одиночного объекта, нужно реализовать, например, слайдер и галерея;
5. Предусмотреть возможность загрузки больших файлов через веб-интерфейс, учитывая ограничения, традиционно связанные с настройками веб-сервера.

В предыдущей версии платформы полный текст статьи публикуется в общем текстовом поле. Для мультимедиа-элементов предусмотрены соответствующие поля со смыслом галереи изображений, слайдер, видео и др., находящиеся вне текста. Визуализация этих элементов в браузере осуществляется, благодаря использованию соответствующих JavaScript-библиотек. Таким образом, формируется статическая структура статьи.

Перемещение мультимедиа-элементов внутрь текста реализуется следующим образом – во время верстки статьи технический редактор должен дать имена всем мультимедиа-объектам, а в необходимых местах текста вставить соответствующие якоря на элементы. Во время отображения страницы статьи в

браузере выполняется JavaScript, который перемещает мультимедиа-объект по DOM-дереву на место соответствующего якоря (рис. 3).

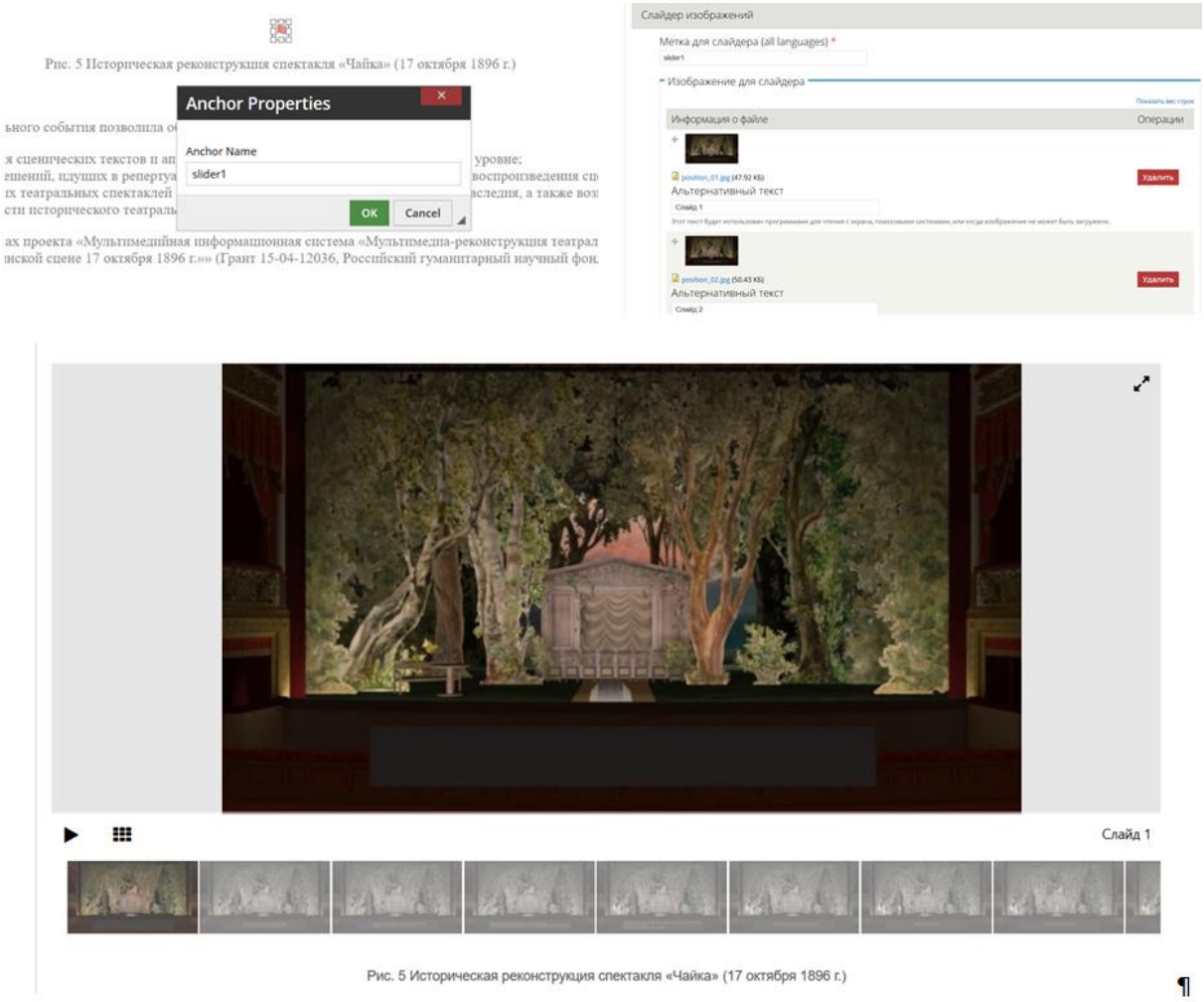


Рис. 3. Пример реализации административного и пользовательских интерфейсов динамического элемента «Слайдер»

На основании такого подхода реализована платформа журнала «Культура и технологии»[14]. Работа была выполнена на базе CMS Drupal 7. Основные инструменты редактирования статей были реализованы с использованием инструментов Drupal, позволяющих работать с полями.

С момента начала своего функционирования журнал «Культура и технологии» размещает статьи со следующими вариантами мультимедиа контента – галерея изображений, слайдер изображений, видео, видео360. В качестве примера можно привести несколько статей:

[14] http://cat.ifmo.ru/

— Статья «Application of Video 360° Technology for the Presentation of the Solovetsky Monastery Cultural Heritage», включающая 7 фрагментов водео360[15] [8];

— Статья «Мультимедиа-реконструкция театрального события. Премьера спектакля «Чайка» на Александринской сцене 17 октября 1896 г.», включающая слайдеры из изображений реконструированных сцен и моделей декораций. В статье представлено более 20 сцен из спектакля, демонстрирующих детальность проработки проекта. Использование такого количества изображений в традиционной статье представлялось бы затруднительным[16] [9].

Дальнейшее развитие инструментальной платформы идет по направлению большей универсальности конструктора статьи. В связи с тем, что структура и содержание статьи заранее неизвестны, необходим удобный инструмент работы с фрагментами, позволяющий компоновать их в произвольном порядке и в неограниченном количестве (рис. 4).

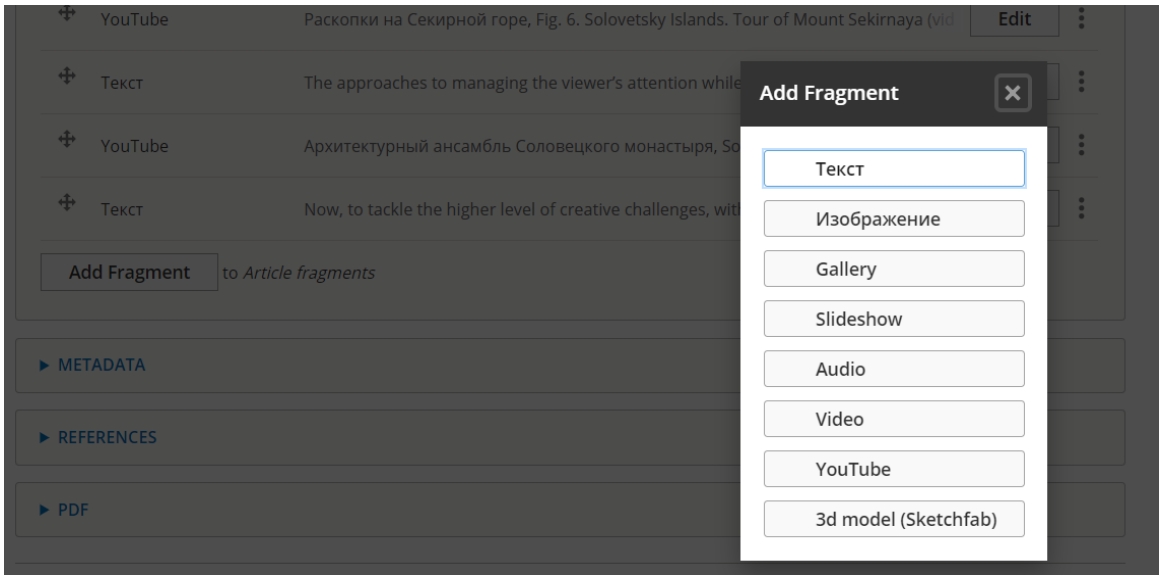


Рис. 4. Инструмент добавления произвольных фрагментов статьи

Основная идея заключается в использовании текста в качестве фрагмента. Таким образом, статья формируется по следующей схеме: добавление теста до мультимедиа-элемента, добавление мультимедиа-элемента, добавление следующего фрагмента текста и так далее (рис. 5).

[15] http://cat.ifmo.ru/ru/2016/v1-i1/88
[16] http://cat.ifmo.ru/ru/2016/v1-i1/65

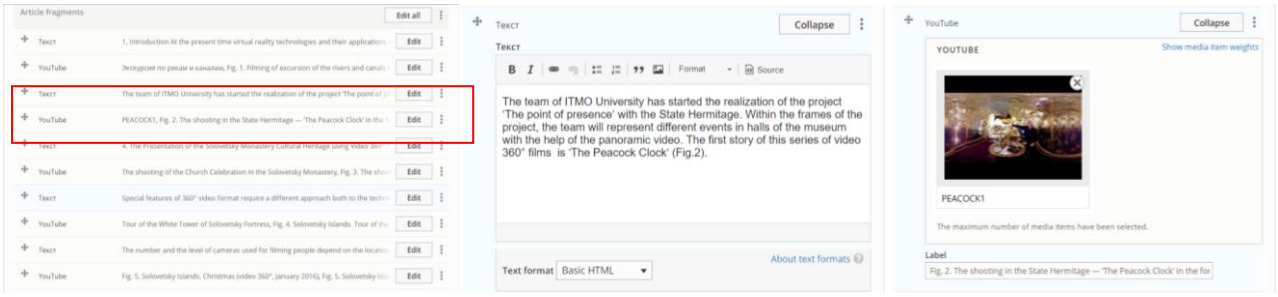


Рис. 5. Режим редактирования статьи при использовании фрагментов

В рамках развития инструментальной платформы мультимедиа электронных коллекций и, в частности, электронного журнала с мультимедиа-контентом развитие системы осуществляется на базе CMF Drupal 8. В основу конструктора положено использование модуля Paragraphs, который позволяет на лету выбирать заранее предопределенные фрагменты.

На основе конфигурации инструментальной платформы создан профиль для установки CMS Drupal. Разработан дополнительный установочный модуль CMS Drupal вместе со всеми необходимыми модулями и настроенной конфигурацией, позволяющий развернуть полнофункциональную систему на любом стандартном виртуальном хостинге, поддерживающим PHP 7 и MySQL.

Дальнейшее развитие функциональности платформы предполагается вести в нескольких направлениях:

1. Добавление мультимедиа элементов новых типов, которые можно включить в статью, например, 3D-объекты, формулы, диаграммы;
2. Реализация экспорта данных в наукометрические базы данных и научные репозитории: eLibrary, DOAJ;
3. Реализация взаимодействия журнала с научными агрегаторами по протоколу OAI-PMH и мета-тэги Google Scholar;
4. Создание on-line инструмента, позволяющего непосредственно использовать в статье элементы других мультимедийных ресурсов, реализуя, таким образом, «горизонтальную» интеграцию между электронными коллекциями и обеспечивая доступ к информации, хранящейся во внешних базах данных [10].

8. ИНСТРУМЕНТЫ РЕДАКТИРОВАНИЯ МЕТАДАННЫХ

В основном научные журналы в качестве метаданных используют простой набор Дублинского ядра, в том числе, и OJS. Такого набора метаданных недостаточно для формирования сведений о публикации в Российском индексе научного цитирования (РИНЦ). «Блок метаданных любой научной публикации обязательно включает ее библиографическое описание (авторы, название, источник (например, журнал), год издания, том, номер, начальная и конечная страницы), авторское резюме (аннотация, реферат) и ключевые слова, а также различную дополнительную информацию» [7]. При разработке ИС журнала «Культура и технологии» были реализованы необходимые поля метаописаний, а также учтены дополнительные метаданные, характерные для России, – УДК, ББК, ГРНТИ и др. (рис. 6).

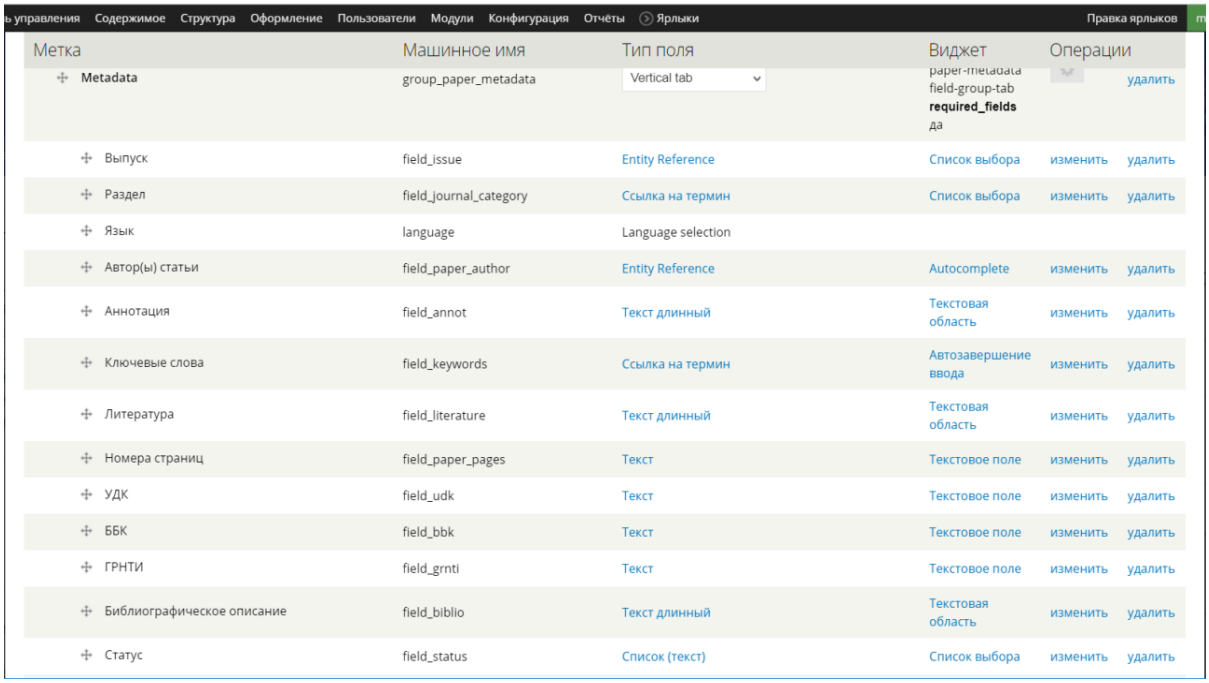


Рис. 6. Блок полей метаописаний ИС электронного журнала

ЗАКЛЮЧЕНИЕ

Целью создания инструментальной платформы поддержки веб-публикации является реализация онлайн-инструментария, позволяющего редакционной коллегии издавать научные журналы, включающие мультимедиа-контент, формировать расширенное метаописание, форматы обмена и изда-

тельский цикл. Основной отличительной особенностью разрабатываемой ИС является реализация онлайн рабочего места, позволяющего верстать статью из широкого набора полей-«кирпичиков» в произвольном порядке.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 19-07-01012.

СПИСОК ЛИТЕРАТУРЫ

1. DSpace: an open source solution for accessing, managing and preserving scholarly works. [Электронный ресурс] // MIT Libraries; HP Labs. 2007. URL: http://www.dspace.org/ (дата обращения: 29.10.2019)

2. *Федотов А.М., Байдавлетов А.Т., Жижимов О.Л., Самбетбаева М.А., Федотова О.А.* Цифровой репозиторий в научно-образовательной информационной системе // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2015. Т. 13. Вып. 3. С. 68–86.

3. *Горбунов-Посадов М.М.* Живая публикация [Электронный ресурс] // ИПМ им. М.В. Келдыша РАН. 2011, редакция от 02.10.2018. URL: http://www.keldysh.ru/gorbunov/live.htm (дата обращения: 29.10.2019)

4. Что такое научное рецензирование? [Электронный ресурс] // Elsevier. URL: https://www.elsevier.com/reviewers/what-is-peer-review (дата обращения: 29.10.2019)

5. *Горбунов-Посадов М.М., Ролдугин Д.С., Слепенков М.И., Тузов И.В.* Анимация и видео в научной публикации // Препринты ИПМ им. М.В. Келдыша. 2014. No 104. 32 с. URL: http://library.keldysh.ru/ preprint.asp?id =2014-104 (дата обращения: 29.10.2019)

6. *Борисов Н.В., Захаркина В.В., Мбого И.А., Прокудин Д.Е.* Проектирование программной платформы полного издательского цикла для издания сетевого мультимедийного журнала [электронный текст] // Культура и технологии. 2017. Т. 2. С. 21–28. URL: http://cat.ifmo.ru/ru/2017/v2-i1/100 (дата обращения: 29.10.2019)

7. *Герасимов А.Н., Елизаров А.М., Липачёв Е.К.* Формирование метаданных для международных баз цитирования с системе управления электронными научными журналами // Электронные библиотеки. 2015. Т. 18. №1-2. С. 6–31.

8. *Борисов Н.В., Волков О.Г., Никитина Л.Л., Николаев А.О., Смолин А.А., Столяров Д.А.* Application of Video 360° Technology for the Presentation of the Solovetsky Monastery Cultural Heritage [электронный текст] // Культура и технологии. 2016. Т. 1. С. 24–31. URL: http://cat.ifmo.ru/en/2016/v1-i1/88 (дата обращения: 29.10.2019)

9. *Борисов Н.В., Никитин А.В., Смолин А.А., Трушин В.А., Чепуров А.А., Чепурова О.А.* Мультимедиа-реконструкция театрального события. Премьера спектакля «Чайка» на Александринской сцене 17 октября 1896 г. [электронный текст] // Культура и технологии. 2016. Т. 1. Вып. 1. С. 15–23. URL: http://cat.ifmo.ru/ru/2016/v1-i1/65 (дата обращения: 29.10.2019)

10. *Борисов Н.В., Захаркина В.В., Мбого И.А., Щербаков П.П.* Проблемы интеграции сетевых электронных коллекций // Труды международной объединенной конференции «Интернет и современное общество», СПб, 2019 (в печати) URL: http://ims.ifmo.ru/file/pages/2/IMS-2019_program_final.pdf (дата обращения: 29.10.2019)

BUILDING A PUBLISHING TOOLKIT FOR MULTIMEDIA SCIECE JOURNALS

N.V. Borisov[1], V.V. Zakharkina[2], I.A. Mbogo[3], D.E. Prokudin[4], P.P. Shcherbakov[5]

St. Petersburg University, Saint-Petersburg

[1]nikborisov@gmail.com, [2]zakharkina@gmail.com, [3]irina.mbogo@gmail.com, [4]hogben.young@gmail.com, [5]paul.tscherbakov@gmail.com

Abstract

This article discusses approaches to the creation of an electronic scientific journal tool platform that provides the publication of multimedia materials through a web interface. The problems associated with the need to include multimedia data of different types are described and a working prototype of the multimedia of the scientific journal is presented.

Keywords: *scientific publication, digital scientific journal, multimedia content, digital collection*

REFERENCES

1. DSpace: an open source solution for accessing, managing and preserving scholarly works. [online] // MIT Libraries; HP Labs. 2007. URL: http://www.dspace.org/ (in English)

2. *Fedotov A.M., Bajdavletov A.T., ZHizhimov O.L., Sambetbaeva M.A., Fedotova O.A.* Digital Repository of Scientific and Educational Information Sysnem // Vestnik. Novosibirsk State Univercity. Series: Information Technologies. 2015. V. 13. No 3. P. 68–86 (in Russian).

3. *Gorbunov-Posadov M.M.* Zhivaya publikaciya [online] // IGM im. M.V. Keldysha RAN. 2011, redakciya ot 02.10.2018. URL: http://www.keldysh.ru/gorbunov/live.htm (in Russian)

4. What is peer review? [online] // Elsevier. URL: https://www.elsevier.com/reviewers/what-is-peer-review (in English)

5. *Gorbunov-Posadov M.M., Roldugin D.S., Slepenkov M.I., Tuzov I.V.* Animation and video in scientific publication // KIAM Preprint № 104, Moscow, 2014. No 104. 32 p. URL: http://library.keldysh.ru/preprint.asp?id=2014-104 (in Russian)

6. *Borisov N.V., Zaharkina V.V., Mbogo I.A., Prokudin D.E.* Proektirovanie programmnoj platformy polnogo izdatel'skogo cikla dlya izdaniya setevogo mul'timedijnogo zhurnala [online] // Kul'tura i tekhnologii. 2017. V. 2. P. 21–28. URL: http://cat.ifmo.ru/ru/2017/v2-i1/100 (data obrashcheniya: 29.3.2019) (in Russian)

7. *Gerasimov A.N., Elizarov A.M., Lipachyov E.K.* Subsystem of Formation Metadata for Science Index Databases on Management Platform Electronic Scientific Journals // Russian Digital Libraries J. 2015. V. 18. No 1-2. P. 6–31. URL: https://elbib.ru/en/article/366 (in Russian)

8. *Borisov N. V., Volkov O. G., Nikitina L. L., Nikolaev A.O., Smolin A.A., Stolyarov D.A.* Application of Video 360° Technology for the Presentation of the Solovetsky Monastery Cultural Heritage [electronic text] // International Culture & Technology Studies. 2016. V. 1. P. 24–31. URL: http://cat.ifmo.ru/en/2016/v1-i1/88 (in Russian)

9. *Borisov N.V., Nikitin A.V., Smolin A.A., Trushin V.A., Chepurov A.A., Chepurov O.A.* Multimedia-reconstruction of the theatrical event. The premiere of

performance "the Seagull" at Aleksandrinsky scene 17 Oct 1896 [online] // International Culture & Technology Studies. 2016. V. 1. Issue 1. P. 15–23. URL: http://cat.ifmo.ru/ru/2016/v1-i1/65 (in Russian)

10. *Borisov N.V., Zaharkina V.V., Mbogo I.A., Shcherbakov P.P.* Problemy integracii setevyh elektronnyh kollekcij // Trudy mezhdunarodnoj ob"edinennoj konferencii " Internet and Modern Society IMS-2019", SPb, 2019 (v pechati) URL: http://ims.ifmo.ru/file/pages/2/IMS-2019_program_final.pdf

СВЕДЕНИЯ ОБ АВТОРАХ

БОРИСОВ Николай Валентинович – профессор, заведующий кафедрой информационных систем в искусстве и гуманитарных науках Санкт-Петербургского государственного университета, специалист в области мультимедийных информационных систем, цифрового культурного наследия и телекоммуникационных систем для науки, образования и культуры.

Nikolay Valentinovich BORISOV – Professor, Head of the Department of information systems in the arts and humanities of St. Petersburg University, specialist in the field of multimedia information systems, digital cultural heritage and telecommunication systems for science, education and culture.

email: nikborisov@gmail.com, n.borisov@spbu.ru

ЗАХАРКИНА Валентина Валентиновна – доцент факультета искусств Санкт-Петербургского государственного университета, специалист в области мультимедийных информационных систем, веб-дизайна.

Valentina Valentinovna ZAKHARKINA – Associate professor of faculty of Arts *of Saint-Petersburg State University,* specialist in the field of multimedia information systems, web-design.

email: zakharkina@gmail.com

МБОГО Ирина Анатольевна – старший преподаватель факультета искусств Санкт-Петербургского государственного университета, специалист в области мультимедийных информационных систем, веб-дизайна.

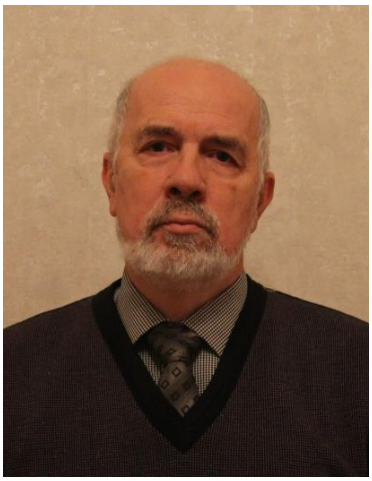
Irina Anatiljevna MBOGO -– Senior lecture of faculty of Arts of Saint-Petersburg State University, specialist in the field of multimedia information systems, web – design

email: irina.mbogo@gmail.com

ПРОКУДИН Дмитрий Евгеньевич – доктор философских наук, доцент кафедры культурологии, философии культуры и эстетики Санкт-Петербургского государственного университета, главный редактор электронного мультимедийного журнала «Культура и технологии», специалист в области информатизации образования научной деятельности, исследователь культуры информационного общества.

Dmitry E. PROKUDIN - Associate professor, Dr. Science (Philosophy), Saint-Petersburg State University, editor-in-chief of the online multimedia journal "International Culture and Technology Studies", a specialist in Informatization of education and scientific activity, researcher of information society culture.

email: hogben.young@gmail.com



ЩЕРБАКОВ Павел Петрович – доцент факультета искусств Санкт-Петербургского государственного университета**,** специалист в области мультимедийных информационных систем, баз данных, систем виртуальной реальности.

Pavel Shcherbakov SHCHERBAKOV - Associate Professor of faculty of Arts *of Saint-Petersburg State University,* specialist in the field of multimedia information systems, databases, virtual reality.

email: *paul.tscherbakov@gmail.com*

Материал поступил в редакцию 20 ноября 2019 года

УДК 004.021 + 004.42

МОДЕЛЬ САМОТРАНСФОРМАЦИИ ГРАФОВ, ОСНОВАННАЯ НА ОПЕРАЦИИ ИЗМЕНЕНИЯ КОНЦА РЕБРА

И. Б. Бурдонов

Институт системного программирования им. В.П. Иванникова Российской академии наук, г. Москва

igorburdonov@yandex.ru, igor@ispras.ru

Аннотация

Рассмотрена распределенная сеть, топология которой описана неориентированным графом. Сеть может сама изменять свою топологию, используя специальные «команды», подаваемые ее узлами. В работе предложена предельно локальная атомарная трансформация $a \to c \to b$ изменения конца c ребра ac, «движущегося» вдоль ребра cb от вершины c к вершине b. В результате этой операции ребро ac удаляется, а ребро ab добавляется. Такая трансформация выполняется по «команде» от общей вершины c двух смежных ребер ac и cb. Показано, что из любого дерева можно получить любое другое дерево с тем же множеством вершин, использовав только атомарные трансформации. Если степени вершин дерева ограничены числом d ($d \geq 3$), то трансформация не нарушает этого ограничения. В качестве примера цели такой трансформации рассмотрены задачи максимизации и минимизации индекса Винера дерева с ограниченной степенью вершин без изменения множества его вершин. Индекс Винера – это сумма попарных расстояний между вершинами графа. Максимальный индекс Винера имеет линейное дерево (дерево с двумя листовыми вершинами). Для корневого дерева с минимальным индексом Винера определены его вид и способ вычисления числа вершин в ветвях соседей корня. Предложены два распределенных алгоритма: трансформации дерева в линейное дерево и трансформации линейного дерева в дерево с минимальным индексом Винера. Доказано, что оба алгоритма имеют сложность не выше $2n–2$, где n – число вершин дерева. Также рассмотрена трансформация произвольных неориентированных графов, в которых могут быть циклы, кратные ребра и петли, без ограничения на степени вершин. Показано, что любой связный граф с n вершинами может быть преобра-

зован в любой другой связный граф с k вершинами и тем же числом ребер за время не более $2(n+k)–2$.

***Ключевые слова**: распределенная сеть, самотрансформация графов, индекс Винера*

ВВЕДЕНИЕ

Индекс Винера [1] – топологический индекс молекулярных графов, используемый во многих приложениях, особенно в математической и компьютерной химии и хемоинформатике.

Рассмотрим распределенную сеть, топология которой – динамически изменяющееся дерево. Динамические графы [2] моделируют самоорганизующиеся сети [3–5], в том числе, социальные сети, нейронные сети [6] и роевой интеллект [7]. Особенность этих сетей – их однородность, без разделения узлов на коммутаторы, хосты и контроллеры. Основное внимание уделяется вопросам маршрутизации, пропускной способности, помехоустойчивости, безопасности, распределения нагрузки и сетевых ресурсов и т. п. Изменение топологии сети понимается как внешний фактор, который надо учитывать, но которым сама сеть не управляет или управляет лишь частично [8, 9].

С другой стороны, в литературе много работ, посвященных как раз целенаправленной трансформации графа, в частности, деревьев, с целью оптимизации по тем или иным критериям. В настоящей статье предложена атомарная трансформация, которая предельно локальна, затрагивая минимум вершин и ребер, максимально близких друг другу. Ранее [10–12] рассматривались другие трансформации деревьев, но они недостаточно локальны и сводятся к цепочкам атомарных трансформаций.

Предлагаемые в статье алгоритмы распределенные и параллельные. Дерево трансформирует само себя по «командам» от вычислительных единиц, соотносимых с вершинами. Для этого и нужна локальность трансформации.

Структура статьи следующая. В разделе 1 определены модель распределенной сети и атомарная трансформация. Раздел 2 содержит основные понятия и утверждения, связанные с индексом Винера. В разделе 3 предложен алгоритм трансформации дерева в линейное дерево, а в разделе 4 – из линейного дерева в дерево с минимальным индексом Винера и заданным ограничением на степе-

ни вершин. Даны оценки сложности. Доказательства утверждений можно найти в [15].

В разделе 5 рассмотрен общий случай трансформации неориентированных связных графов, в которых могут быть циклы, кратные ребра и петли, без ограничения на степени вершин. С помощью атомарной трансформации изменения конца ребра смоделированы удаление и добавление вершин. Показано, что любой связный граф с n вершинами может быть преобразован в любой другой связный граф с k вершинами и тем же числом ребер за время не более $2(n+k)-2$.

1. МОДЕЛЬ

Пусть G – неориентированное дерево без кратных ребер и петель с ограничением d ($d \geq 3$) на степени вершин, лежащее в основе распределенной сети. Дерево *упорядоченное*: ребрам, инцидентным вершине, присвоены различные ненулевые номера. Ребро ab имеет два номера: e_{ab} в вершине a и e_{ba} в вершине b.

Вершины отождествляются с вычислительными единицами, которые посылают друг другу сообщения по ребрам графа. Память вершины – набор переменных. Вначале в каждой вершине a переменная $E(a)$ инициализирована множеством номеров ребер, инцидентных a. Далее при трансформации графа вершина a сама корректирует переменную $E(a)$.

Сообщение задается типом и параметрами: $Тип(p_1, \ldots, p_k)$. Вершина a, посылая сообщение по ребру ab, указывает его номер e_{ab}. Вершина b получает сообщение вместе с номером ребра e_{ba}.

Дерево трансформируется по «командам» от своих вершин. Атомарная трансформация $a \rightarrow c \rightarrow b$ – это замена ребра ac на ребро ab при наличии ребра cb (рис. 1).

Выполняется по команде ***Изменить***(e_{ca}, e_{cb}, P), которую подает вершина c, где P – дополнительные параметры. Ребро ab получает в вершине a тот же номер, который имело удаляемое ребро ac, т. е. e_{ac}, а в вершине b – любой «свободный» номер e_{bc}. Для того чтобы вершина b «узнала» этот номер, по ребру из a в b автоматически посылается сообщение *Изменение*(P). Другие сообщения, передаваемые по изменяемому ребру, не теряются, но сообщение, направлявшееся в вершину c, получит вершина b. Вершина c сама удаляет номер e_{ca} из $E(c)$,

а вершина *b* сама добавляет номер e_{ba} в $E(b)$. Атомарная трансформация не изменяет множество вершин дерева и оставляет его деревом.

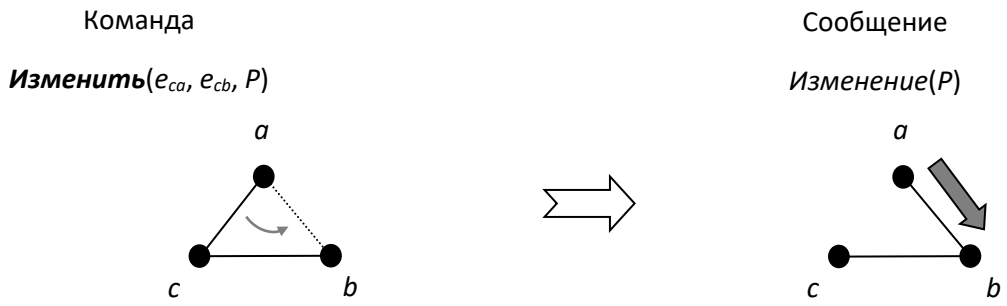


Рис. 1. Трансформация $a \to c \to b$: замена ребра *ac* на ребро *ab*

Для оценки времени работы алгоритмов будем пренебрегать временем вычислений в вершине и предполагать, что время пересылки сообщения по ребру и время атомарной трансформации, включая пересылку сообщения *Изменение*, не превышает 1 такта.

Утверждение 1: Для любых двух деревьев с ограничением d ($d \geq 3$) на степени вершин одно можно получить из другого с помощью цепочки атомарных трансформаций, причем в процессе трансформации ограничение d на степени вершин не будет нарушено.

2. ИНДЕКС ВИНЕРА

Индекс Винера – это сумма всех попарных расстояний между вершинами. Для данного числа вершин максимальный индекс Винера имеет линейное дерево (дерево с двумя листьями) (A000292 в [13]). Вид дерева с ограничением на степени вершин и минимальным индексом Винера определен в [14]. Это разновидность сбалансированного дерева (высоты листьев отличаются не более чем на 1) с жестким требованием к степени вершины, что отличает его от B-деревьев (в которых все листья находятся на одной высоте, а степени вершин могут быть разными) и от АВЛ-деревьев (которые двоичны).

Дерево *G* с выделенной вершиной – *корнем* – называется *корневым*. *Высота вершины* – расстояние от нее до корня. *Высота дерева* – максимальная высота вершины. *Ветвь вершины v* – подграф $G(v)$, порожденный множеством вершин, связанных с корнем путем, проходящим через *v*. Для ребра *ab* вершина

a – *отец* вершины *b*, а вершина *b* – *сын* вершины *a*, если путь из корня в *b* проходит через *a*. У каждой вершины, кроме корня, ровно один отец.

В упорядоченном корневом дереве вершины одной высоты линейно упорядочены: вершина *v левее* вершины *w* (*w правее v*), если после максимального общего префикса путей, ведущих из корня в *v* и *w*, номер следующего ребра на пути в *v* меньше номера следующего ребра на пути в *w*.

Корневое дерево высотой *h* с *n* вершинами *почти хорошее*, если степень корня равна min{*d*–1, *n*–1}, для *h*≥3 все вершины на высоте 1 .. *h*–2 имеют степень *d*, и дерево можно так упорядочить, что для *h*≥2 на высоте *h*–1 самая правая внутренняя вершина *u* имеет степень не больше *d*, вершины левее *u* имеют степень *d*, а вершины правее *u* – листья. *Хорошее* дерево отличается только степенью корня, она равна min{*d*, *n*–1}. Примеры даны на рис. 2.

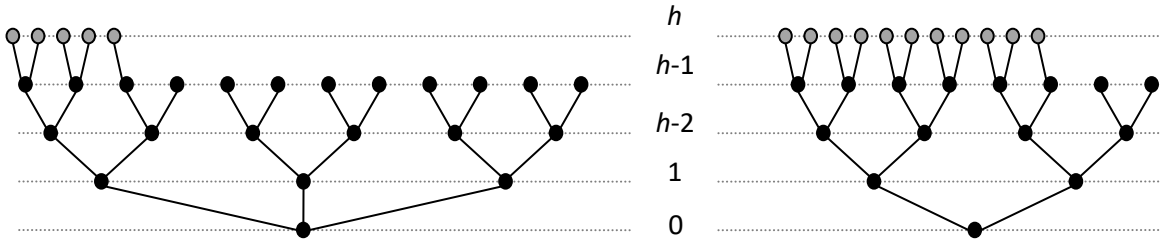


Рис. 2. Хорошее дерево (слева) и почти хорошее дерево (справа)

Утверждение 2 (Теорема 2.2 в [14]). Дерево со степенью вершин не более *d* (*d*≥3) имеет минимальный индекс Винера тогда и только тогда, когда это хорошее дерево.

Пусть в почти хорошем дереве высотой *h* степень корня равна 0 или *d*–1, а степени всех вершин на высоте *h*–1 равны *d*. Число вершин этого дерева обозначим $N(d, h)=1+(d–1)+(d–1)^2+…+(d–1)^h = ((d–1)^{+1} –1)/(d–2)$. Пусть в хорошем дереве высотой *h* степень корня равна 0 или *d*, а степени всех вершин на высоте *h*–1 равны *d*. Число вершин этого дерева обозначим *M*(*d*, *h*): *M*(*d*, 0)=1 и $M(d, h)=1+d+d(d–1)+…+d(d–1)^{h-1}=1+dN(d, h–1)$ для *h*≥1. Примеры – на рис. 2 при удалении «серых» вершин на высоте *h*.

Пусть задано (почти) хорошее дерево с *n* вершинами. Ветвь соседа корня – почти хорошее дерево. Упорядочим соседей корня по невозрастанию числа вершин в их ветвях и обозначим эти числа:

для почти хорошего дерева: $N(d, n, 1)≥…≥N(d, n, \min\{d–1, n–1\})$,

для хорошего дерева: $M(d, n, 1) \geq \ldots \geq M(d, n, \min\{d, n-1\})$.

Утверждение 3. Пусть $L(d, i)=N(d, i)$, если G – почти хорошее дерево, и $L(d, i)=M(d, i)$, если G – хорошее дерево. Пусть число вершин $n=L(d, h-1)+m(d-1)+s<L(d, h)$, где $0 \leq s<d-1$ и $m=p(d-1)^{h-2}+q$, где $0 \leq q<(d-1)^{h-2}$. Тогда для p самых левых соседей корня их ветви имеют по $N(d, h-1)$ вершин, для следующего справа соседа корня его ветвь имеет $N(d, h-2)+q(d-1)+s$ вершин, а для остальных соседей корня их ветви имеют по $N(d, h-2)$ вершин.

3. АЛГОРИТМ $\mathcal{A}$ ТРАНСФОРМАЦИИ В ЛИНЕЙНОЕ ДЕРЕВО

Если для вершины v ветвь ее сына w – линейное дерево, то путь от v через w до листа назовем *линейкой из v*. *Звездообразное* дерево – это корневое дерево, состоящее из линеек, ведущих из корня.

Пусть G – дерево с n вершинами с корнем r. Для удобства будем считать, что у корня есть фиктивное ребро с номером 0, ведущее к фиктивному отцу. Алгоритм стартует при получении корнем от его (фиктивного) отца сообщения *Старт*() и завершается посылкой из корня по этому ребру сообщения *Линия*(n). Алгоритм рекурсивный, на уровне рекурсии h алгоритм работает на каждой ветви $G(v)$, где v имеет высоту h, и состоит из трех этапов.

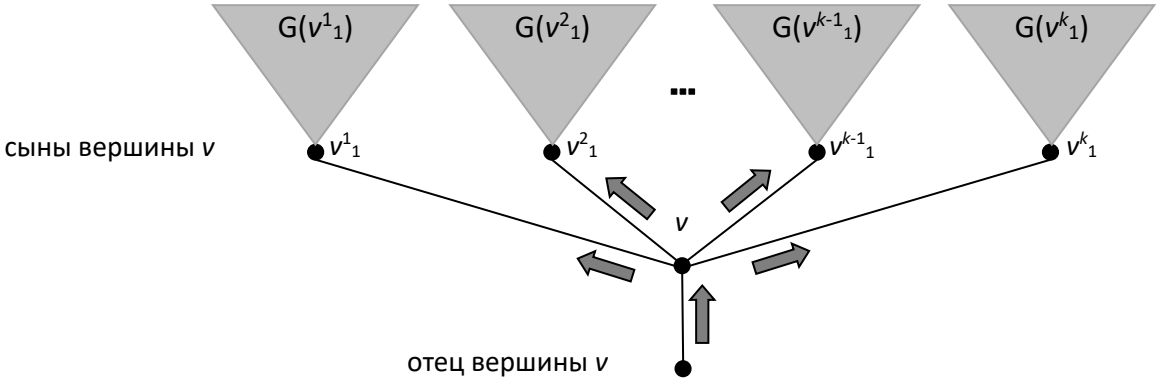


Рис. 3. Этап 1: Сообщения *Старт* и ветви дерева

Этап 1 (рис. 3). Вершина v получает от своего отца сообщение *Старт*, запоминает номер ребра, ведущего к отцу, и рассылает *Старт* всем своим сынам.

Этап 2 (рис. 4). Вершина v ожидает от своих сынов получения сообщений *Линия*, подсчитывая число вершин ветви $G(v)$ как 1 + сумма параметров получаемых сообщений *Линия*.

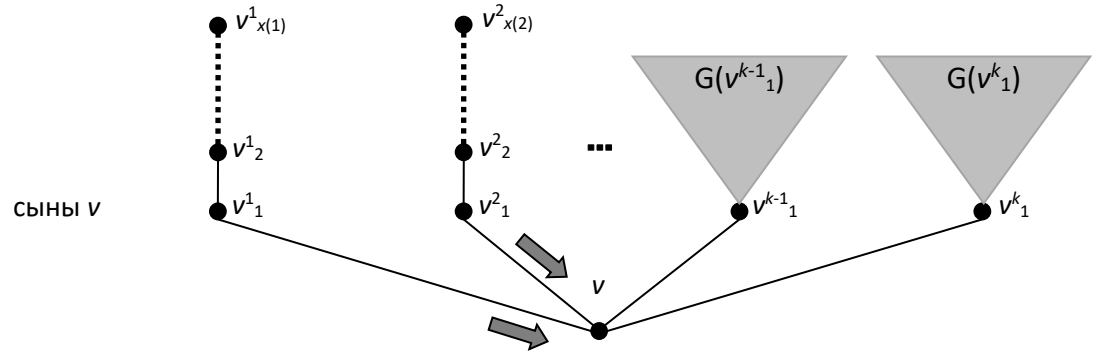


Рис. 4. Этап 2: Сообщения *Линия* и линейки

В начале этапа 3 (рис. 5) ветвь $G(v)$ – звездообразное дерево с k линейками. Длина i-й линейки равна $x(i)$. Вершина v запускает k–1 параллельных цепочек атомарных трансформаций так, что для i=1 .. k–1 у первого ребра i+1-ой линейки его конец $v^{i+1}{}_1$ фиксируется, а другой конец двигается по i-ой линейке от v до листа $v^i_{x(i)}$. Для этого вершина v выполняет сразу k–1 трансформаций $v^{i+1}{}_1 \to v \to v^i{}_1$ в последовательности $i=k$–1, ..., 1. Каждая из этих трансформаций – первая в цепочке трансформаций вдоль i-й линейки: $v^{i+1}{}_1 \to v \to v^i{}_1$, $v^{i+1}{}_1 \to v^i{}_1 \to v^i{}_2$, …, $v^{i+1}{}_1 \to v^i{}_{x(i)-1} \to v^i{}_{x(i)}$. Эти цепочки трансформаций выполняются параллельно вдоль k–1 линеек. Когда цепочка трансформаций вдоль i-й линейки заканчивается в ее листе, происходит конкатенация i-й и i+1-й линеек. Когда это случится для всех i=1..k–1, ветвь $G(v)$ станет линейным деревом. Об этом вершину v извещает сообщение *Финиш*(). Оно посылается после конкатенации k-й и k–1-й линеек, и далее проходит по линейкам k–1,…,1, причем i-я линейка проходится от конца $v^i_{x(i)}$ к началу v^i_1. Если *Финиш* переходит на i–1-ю линейку до конкатенации ее с i-й линейкой, пересылка приостанавливается до завершения конкатенации. В конце *Финиш* посылается по ребру (v^1_1, v). Вершина v посылает своему отцу сообщение *Линия*, завершая работу на ветви $G(v)$.

Утверждение 4. Алгоритм $\mathcal{A}$ трансформирует дерево с n вершинами в линейное дерево, не нарушая ограничения d на степени вершин, за время $t(n) \le 2n-2$. Из корня по его фиктивному отцу послано сообщение *Линия*(n).

Оценка $2n$–2 достижима на линейном дереве, когда корень – один из листов. Оценка не улучшаема: чтобы выяснить, что дерево линейное, сообщение

должно дойти от корня до другого листа и обратно, т. е. пройти путь длиной 2*n*–2.

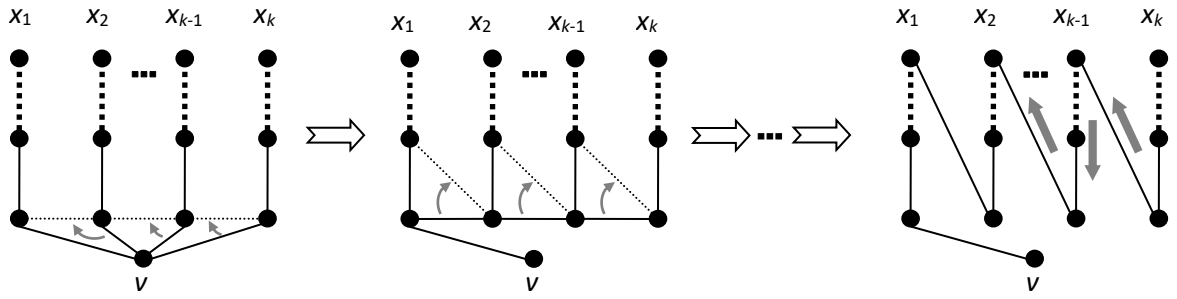


Рис. 5. Этап 3: Сообщения *Финиш* и трансформации

4. АЛГОРИТМ $\mathcal{B}$ – ТРАНСФОРМАЦИЯ ЛИНЕЙНОГО ДЕРЕВА В ХОРОШЕЕ ДЕРЕВО

Пусть *G* – линейное дерево с *n* вершинами и корнем *r* в листе. Алгоритм стартует при получении корнем по фиктивному ребру с номером 0 сообщения *Начало*(*d*, *n*), а завершается посылкой из корня по этому ребру сообщения *Конец*().

Алгоритм выполняется рекурсивно, уровень рекурсии равен высоте вершины в целевом хорошем дереве. На уровне рекурсии *h* построена часть хорошего дерева на высотах от 0 до *h*. Вначале *h*=0, и построенная часть состоит из одного корня. На уровне *h* алгоритм выполняется на каждой ветви *G*(*v*), где вершина *v* имеет в *G* высоту *h*, и состоит из двух этапов.

(*Почти*) *хорошим звездообразным* деревом назовем звездообразное дерево, в котором степень корня и числа вершин ветвей соседей корня такие же, как у (почти) хорошего дерева с тем же числом вершин.

Этап 1. Ветвь *G*(*v*) – линейка из *v*. Строим звездообразное дерево с корнем в *v*: хорошее, если *v*=*r*, и почти хорошее, если *v*≠*r*. Число вершин ветви *G*(*v*) – параметр сообщения *Начало*, с получения которого стартует этап 1 на ветви *G*(*v*).

Этап 2. Ветвь *G*(*v*) – хорошее (*v*=*r*) или почти хорошее (*v*≠*r*) звездообразное дерево. Вершина *v* посылает каждому сыну *w* сообщение *Начало*(*d*, *l*), где *l* – число вершин на ветви *G*(*w*), инициируя работу алгоритма на следующем уровне рекурсии. Это можно делать, как только построена линейка нужной длины из *w*. Вершина *v* ожидает от всех своих сынов сообщений *Конец*(), а затем посылает своему отцу сообщение *Конец*(). Если *v*=*r*, алгоритм заканчивается.

Как строить звездообразное дерево на этапе 1? Используем понятие *текущей вершины* (вначале вершина v) и две операции: *перемещение* и *трансформация*. Параметр t – число трансформаций, которые осталось сделать для построения линейки звездообразного дерева.

Перемещение $c \dashrightarrow b$: c – текущая вершина, есть ребро $\{c, b\}$. Вершина c посылает в вершину b сообщение *Перемещение*(t). Получив сообщение, вершина b становится текущей.

Трансформация $a \to c \to b$: c – текущая вершина, есть ребра $\{a, c\}$ и $\{c, b\}$. Величина t уменьшается на 1: $t:=t-1$. Вершина c подает команду ***Изменить***(e_{ca}, e_{cb}, t). Получив сообщение *Изменение*(t), вершина b становится текущей.

Построение показано на рис. 6: серая стрелка указывает текущую вершину, белый кружок – вершину v. Пусть $l>2$ – число вершин ветви $G(v)$. Вначале есть линейка $v=v_1, v_2, \ldots, v_l$, текущая вершина v. Обозначим:

$x=\min\{d, l-1\}$ – степень вершины v в хорошем звездообразном дереве;

$SM(d, l, 0)=1$;

$SM(d, l, j) = 1+M(d, l, 1)+M(d, l, 2)+\ldots+M(d, l, j)$, для $j=1 .. x$, – число вершин в хорошем звездообразном дереве на первых j линейках плюс единица (корень);

$v^j_i=v_{SM(d, l, j-1)+i}$, для $j=1 .. x-1$ и $i=1 .. M(d, l, j)$ – i-я вершина j-й линейки;

$v^x_i=v_{SM(d, l, x)-i+1}$ для $i=1 .. M(d, l, x)$ – i-я вершина x-й линейки.

Строим линейки «справа налево», начиная от x-й и заканчивая 2-й.

Построение j-й линейки хорошего звездообразного дерева для $j=x .. 2$:

1. $t=M(d, l, j)$.
2. Перемещение $v \dashrightarrow v^j_1$.
3. Цепочка $M(d, l, j)-1$ трансформаций:

$$v \to v^j_1 \to v^j_2, \quad v \to v^j_2 \to v^j_3, \quad \ldots, v \to v^j_{M(d, l, j)-1} \to v^j_{M(d, l, j)};$$

$$t=M(d, l, j)-1, \quad t=M(d, l, j)-2, \quad \ldots, \quad t=1.$$

4. $M(d, l, j)$-я трансформация: $v^{j+1}_1 \to v^j_{M(d, l, j)} \to v$.
5. Пуск алгоритма построения почти хорошего звездообразного дерева на j-й ветви: вершина v посылает вершине $v^j_{M(d, l, j)}$ сообщение *Начало*(d, $M(d, l, j)$).

Когда построена 2-я линейка, построена и 1-я линейка, поэтому одновременно запускается алгоритм на 1-й линейке: вершина v посылает вершине $v^1_{M(d, l, 1)}$ сообщение *Начало*(d, $M(d, l, 1)$).

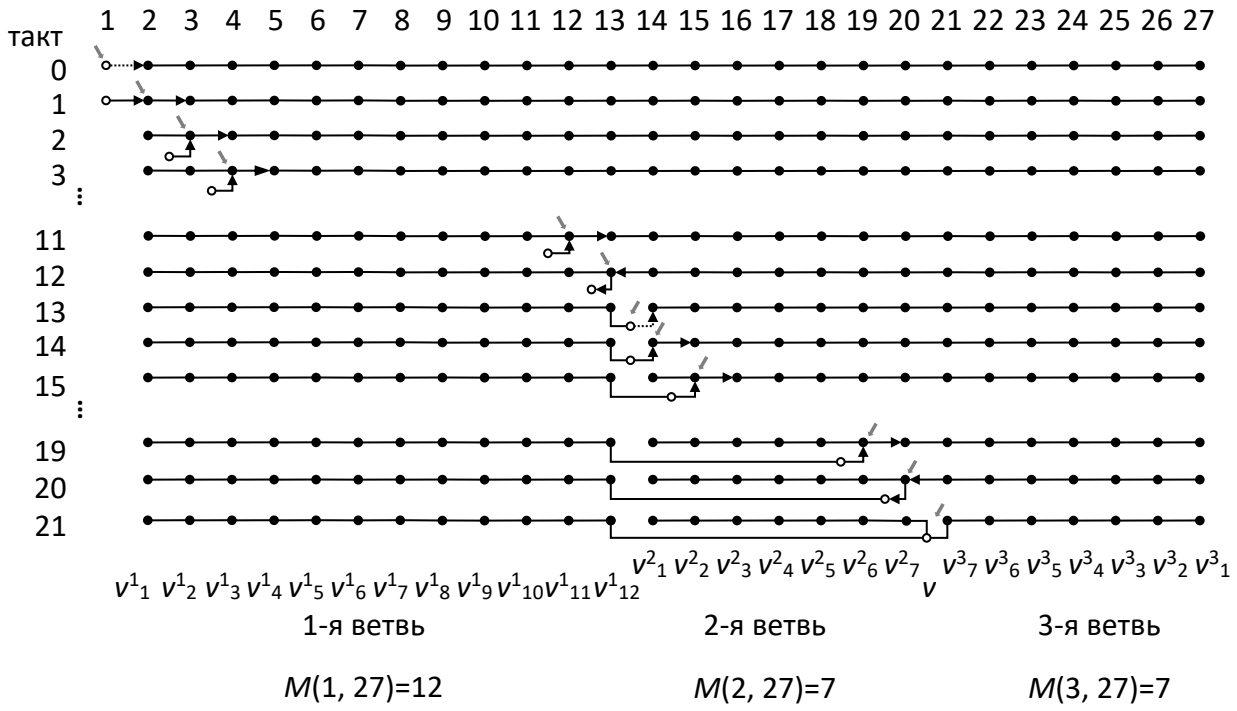


Рис. 6. Построение звездообразного хорошего дерева для n=27 и d=3

Почти хорошее звездообразное дерево строится так же, но число линеек x равно не $\min\{d, l–1\}$, а $\min\{d–1, l–1\}$, и число вершин в j-й линейке равно не $M(d, l, j)$, а $N(d, l, j)$.

Утверждение 5. Алгоритм $\mathcal{B}$ трансформирует линейное дерево с n вершинами в хорошее дерево без нарушения ограничения d на степени вершин за время $t(n)\leq 2n–2$.

5. ТРАНСФОРМАЦИЯ ПРОИЗВОЛЬНЫХ НЕОРИЕНТИРОВАННЫХ ГРАФОВ

В этом разделе рассмотрена трансформация произвольных неориентированных графов, в которых могут быть циклы, кратные ребра и петли, без ограничения на степени вершин.

В предыдущих разделах неявно предполагалось, что атомарная трансформация $a\rightarrow c\rightarrow b$ выполняется тогда, когда ребро cb не изменяется в результате одновременного выполнения других атомарных трансформаций. Здесь мы это ограничение снимаем.

Кроме асинхронной сети, в которой время выполнения атомарных трансформаций и пересылки сообщений по ребрам недетерминировано (но ограни-

чено сверху), мы также будем рассматривать синхронную сеть, в которой это время фиксировано.

В синхронной сети две трансформации $a \to c \to b$ и $c \to b \to d$ могут выполняться одновременно: конец c ребра ac перемещается по ребру cb от вершины c к вершине b, и одновременно конец b ребра cb перемещается по ребру bd от вершины b к вершине d. В результате этих двух трансформаций ребро bd не изменяется, а рёбра ac и cb заменяются рёбрами ad и cd. (рис. 7 справа). Тот же результат получается в асинхронной сети, если трансформации выполняются последовательно в порядке $c \to b \to d$, $a \to c \to b$. Однако при другом порядке последовательного выполнения трансформаций результат получается другой (рис. 7 слева).

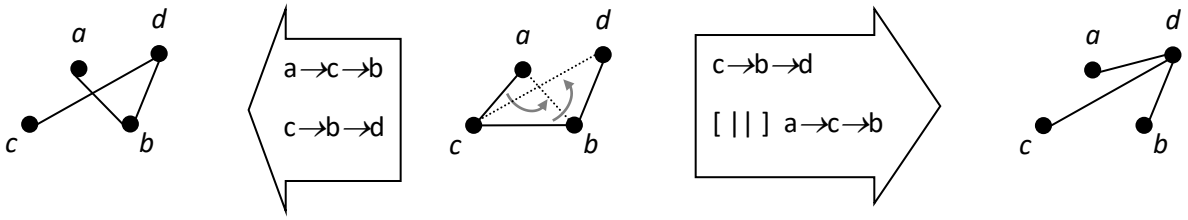


Рис. 7. Порядок трансформаций влияет на результат

При трансформациях могут появляться кратные рёбра: либо в результате одной трансформации при наличии циклов (рис. 8 слева), либо в результате одновременного выполнения одной и той же вершиной нескольких трансформаций (рис. 8 справа).

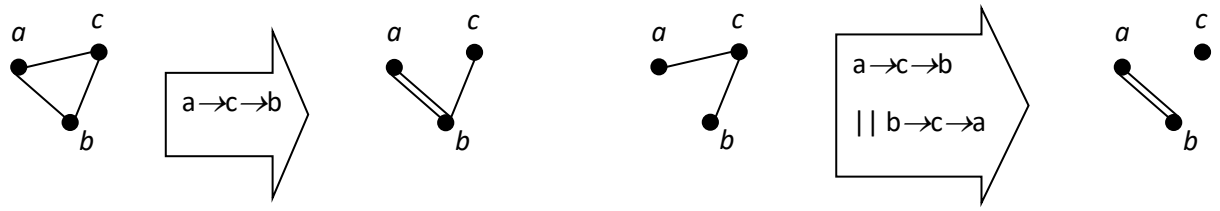


Рис. 8. Появление кратных рёбер

Два кратных ребра образуют цикл длины 2, но могут возникать и циклы большей длины (рис. 9).

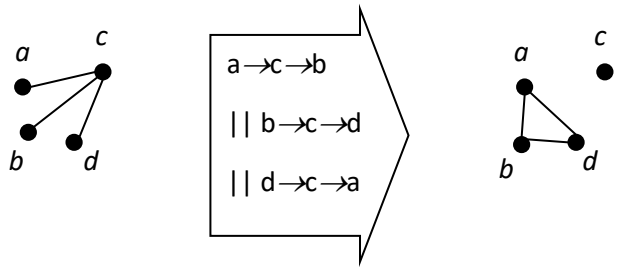


Рис. 9. Появление циклов

Рис. 8 (справа) и рис. 9 демонстрируют также, что в результате трансформаций граф может стать несвязным, в частности, могут появляться изолированные вершины без петель. В таких случаях естественно считать, что результатом трансформации связного корневого графа является компонента связности результирующего графа, которой принадлежит корень.

Особая ситуация возникает при одновременном выполнении трансформаций по циклу (рис. 10). Это можно интерпретировать как рождение новой вершины, соответствующей циклу; в результате трансформаций цикл исчезает, а его рёбра будут вести в эту новую вершину.

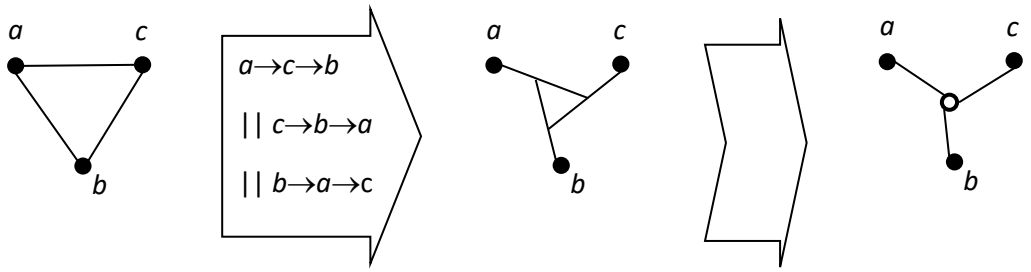


Рис. 10. Рождение вершины из цикла

Конечно, одновременной выполнение трансформаций по циклу требует синхронного выполнения, поскольку команды на эти трансформации подают разные вершины, лежащие на цикле. Однако есть одно исключение, когда цикл имеет длину 1, т. е. состоит из одной петли в одной вершине (рис. 11 справа). Петля, если её не было с самого начала, порождается трансформацией вида $a \to b \to a$ (рис. 11 слева), когда конец b ребра ab двигается вдоль самого этого ребра к вершине a, в которой и образуется петля. Обратная трансформация формально должна была бы записываться как $a \to a \to b$, однако, во-первых, вер-

шины *a* и *b* могут быть не соединены ребром, а, во-вторых, речь идёт о проведении ребра не в старую вершину *b*, а в новую вершину. Поэтому эта трансформация имеет вид $a \to a \to a$, она порождает новую вершину, например, *c,* и превращает петлю *aa* в ребро *ac* (рис. 11 справа). Мы будем записывать эту трансформацию как $a \to a \to a[c]$.

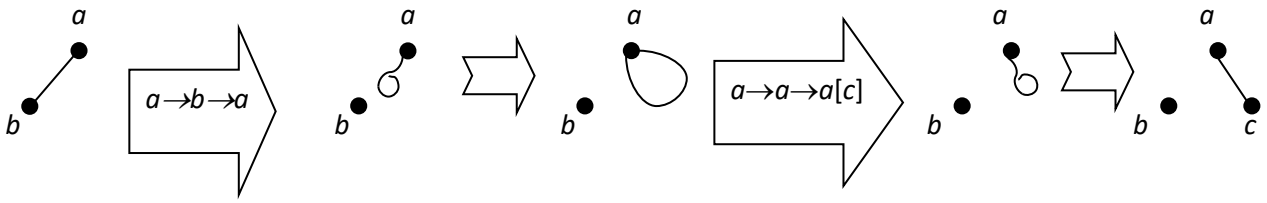


Рис. 11. Превращение ребра в петлю и петли в ребро, ведущее в новую вершину

В качестве примера предложим алгоритм «схлопывания» графа в изолированную вершину с петлями. Рассмотрим произвольный граф *G* с *m* рёбрами и корнем *r*, в котором могут быть циклы, петли и кратные рёбра. При этом петля имеет два номера в вершине. При «схлопывании» граф *G* с *m* рёбрами превращается в граф с *m* петлями в одной вершине — корне *r*.

Идея алгоритма «схлопывания» заключается в следующем: строится остов графа, после чего концы всех рёбер двигаются к корню по этому остову, пока не превратятся в петли в корне.

В вершинах имеются следующие переменные:

d(*v*)=степень вершины *v* в начале (до трансформаций);

s(*v*) – число сообщений *Старт*, полученных вершиной *v*, вначале *s*(*v*)=0;

e(*v*) – число рёбер, инцидентных вершинам ветви *G*(*v*), причем ребра остова считаются дважды, вначале *e*(*v*)=0.

Кроме того, в корне *r* имеется переменная *M* – число полученных корнем *r* сообщений *Изменение* плюс степень корня до трансформаций, вначале *M*=*d*(*r*).

Используются сообщения двух типов: 1) *Старт*(*p*), где *p* – целочисленный параметр в диапазоне [0..*m*], 2) *Изменение* – сообщение, автоматически посылаемое в вершину *b* при трансформации $a \to c \to b$ по ребру *ab*, в которое превращается ребро *ac*.

Алгоритм «схлопывания»:

1. Вначале из корня *r* посылается сообщение *Старт*(1) по всем инцидентным корню рёбрам.

2. Если $c \neq r$, $s(c)=0$ и вершина c получает (первый раз) сообщение *Старт* по ребру cb, вершина c запоминает ребро cb как *обратное* и посылает сообщение *Старт*(1) по всем остальным инцидентным ей рёбрам: $s(c):=1$, $e(c):=2$.
3. Если $c \neq r$, $s(c)>0$ и вершина c получает (повторно) сообщение *Старт*(p) по ребру ca, вершина c делает трансформацию $a \to c \to b$, где ребро cb — обратное. Посылается сообщение *Изменение* по ребру ab, в которое превращается ребро ac, $s(c)=(c)+1$, $e(c):=e(c)+p$.
4. Если $c \neq r$ и вершина c получает сообщение *Изменение* по ребру ac, вершина c делает трансформацию $a \to c \to b$, где ребро cb — обратное. Посылается сообщение *Изменение* по ребру ab, в которое превращается ребро ac.
5. Если $c \neq r$ и $s(c)=d(c)$ (вершина c получила сообщение *Старт* по всем инцидентным ей рёбрам), вершина c посылает сообщение *Старт*($e(c)$) по обратному ребру cb и делает трансформацию $b \to c \to b$. Посылается сообщение *Изменение* по ребру bb, в которое превращается ребро bc.
6. Если корень r получает сообщение *Изменение*, $M:=M+1$.
7. Если корень r получает сообщение *Старт*(p), $e(r):=e(r)+p$.
8. Если $s(r)=d(r)$ (корень r получил сообщение *Старт* по всем инцидентным ему рёбрам) и $e(r)=M$, конец алгоритма.

Время работы алгоритма не превосходит $2n-1$, где n – число вершин. Действительно, время достижения сообщением *Старт* любой вершины не превосходит $n-1$. Тогда сообщение *Старт* пройдёт по каждому ребру ac из вершины a в вершину c через время не более n от начала работы алгоритма. После этого начинается цепочка трансформаций, перемещающая конец c этого ребра до корня. Число трансформаций в этой цепочке равно расстоянию по остову от вершины c до корня, которое не превосходит $n-1$. Тем самым, через время не более $2n-1$, каждый конец каждого ребра перемещается в корень. Следовательно, каждое ребро становится петлёй в корне.

Каждая вершина c, получая сообщения *Старт*, подсчитывает число $e(c)$ ребер, инцидентных вершинам поддерева $G(c)$, причём рёбра остова считаются дважды. После этого вершина c посылает по обратному ребру сообщение *Старт*($e(c)$). Когда корень r получит сообщение *Старт* по каждому инцидентному ему ребру, переменная $e(r)$ будет содержать удвоенное число рёбер графа $2m$. С другой стороны, от каждого из двух концов каждого ребра, кроме конца r

ребра, инцидентного корню до трансформаций, в корень придёт одно сообщение *Изменение*. Тем самым корень получит 2*m*–*d*(*r*) сообщений *Изменение*, и переменная *M* станет равной 2*m*. Поэтому конец алгоритма правильно определяется по равенству *e*(*r*)=*M*.

Обратная последовательность обратных трансформаций алгоритма «схлопывания» графа *G* с *m* ребрами выполняет «расхлопывание» графа, т. е. превращение изолированного корня с *m* петлями в граф *G*. На рис. 12 показан пример «схлопывания» и «расхлопывания» графа.

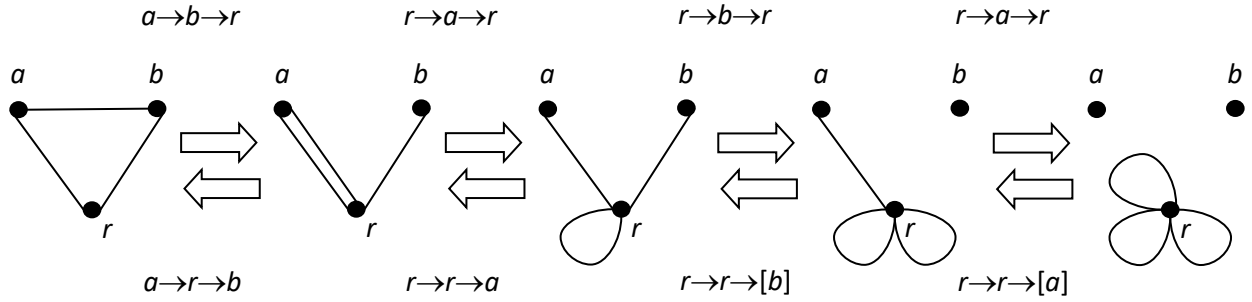


Рис. 12. «Схлопывание» и «расхлопывание» графа

Таким образом, любой связный граф с *n* вершинами может быть преобразован в любой другой связный граф с *k* вершинами и тем же числом ребер за время не более (2*n*–1)+(2*k*–1)=2(*n*+*k*)–2.

ЗАКЛЮЧЕНИЕ

В статье предложены два алгоритма самотрансформации дерева, лежащего в основе распределенной сети, с помощью локальных атомарных трансформаций, выполняемых по «командам» от узлов дерева. Это алгоритм трансформации любого дерева в линейное дерево и алгоритм трансформации линейного дерева в хорошее дерево, имеющее минимальный индекс Винера при том же числе вершин. Трансформации не изменяют множество вершин дерева и не нарушают ограничение *d* (*d*≥3) на степени вершин.

Оба алгоритма имеют верхнюю оценку времени работы 2*n*–2, где *n* – число вершин дерева. Однако, если для алгоритма $\mathcal{A}$ трансформации в линейное дерево эта оценка достижима, то для алгоритма $\mathcal{B}$ трансформации в хорошее дерево это не так. Это объясняется тем, что для алгоритма $\mathcal{A}$ оценка достигается на линейном дереве, а хорошее дерево алгоритм $\mathcal{A}$ трансформирует в линейное за меньшее время. Соответственно, и «обратный» алгоритм $\mathcal{B}$ трансформирует ли-

нейное дерево в хорошее меньше чем за 2*n*–2 тактов. Поэтому верхняя оценка для алгоритма $\mathcal{B}$ нуждается в уточнении (в сторону уменьшения).

Другое отличие алгоритмов $\mathcal{A}$ и $\mathcal{B}$ состоит в следующем. Алгоритм $\mathcal{A}$ (если из него удалить функцию вычисления числа вершин дерева) не зависит от *n* – числа вершин дерева. Поэтому алгоритм $\mathcal{A}$ может выполняться конечными автоматами в вершинах дерева. Однако алгоритм $\mathcal{B}$ существенно зависит от *n*. Остаются вопросы:

1) Можно ли сделать алгоритм трансформации линейного дерева в хорошее дерево не зависящим от *n*?

2) Можно ли сделать алгоритм трансформации дерева в хорошее дерево «напрямую», т.е. не через линейное дерево?

В статье также рассмотрена трансформация произвольных неориентированных графов, в которых могут быть циклы, кратные ребра и петли. Такая трансформация сохраняет число ребер графа. В качестве примера предложен алгоритм преобразования графа в изолированную вершину с петлями. Ограничение на степени вершин при этом, очевидно, не сохраняется. Интерес представляет исследование таких трансформаций, которые сохраняют ограничение на степени вершин. В частности, трансформации, имеющие целью минимизацию индекса Винера для графов с циклами.

СПИСОК ЛИТЕРАТУРЫ

1. *Wiener H.* Structural determination of paraffin boiling points // J. Am. Chem. Soc. 1947. No 69 (1). P. 17–20.

2. *Кочкаров А.А., Сенникова Л.И., Кочкаров Р.А.* Некоторые особенности применения динамических графов для конструирования алгоритмов взаимодействия подвижных абонентов // Известия ЮФУ. Технические науки, раздел V, системы и пункты управления. 2015. №1. С. 207–214.

3. *А.В. Проскочило, А.В. Воробьев, М.С. Зряхов, А.С. Кравчук*. Анализ состояния и перспективы развития самоорганизующихся сетей // Научные ведомости, серия экономика, информатика, выпуск 36/1. 2015. № 19 (216). С. 177–186.

4. *Pathan A.S.K.* (ed.). Security of self-organizing networks: MANET, WSN, WMN, VANET. CRC press, 2010. 638 p.

5. *Boukerche A.* (ed.). Algorithms and protocols for wireless, mobile Ad Hoc networks // John Wiley & Sons, 2008. 496 p.

6. *Chen Z., Li S., Yue W*. SOFM Neural Network Based Hierarchical Topology Control for Wireless Sensor Networks // Hindawi Publishing Corporation. J. of Sensors. 2014. article ID 121278. 6 p. http://dx.doi.org/10.1155/2014/121278

7. *Mo S., Zeng J.-C., Tan Y.* Particle Swarm Optimization Based on Self-organizing Topology Driven by Fitness // International Conference on Computational Aspects of Social Networks, CASoN 2010, Taiyuan, China, 10.1109/CASoN. 2010. No 13. P. 23–26.

8. *Wen C.-Y., Tang H.-K.* Autonomous distributed self-organization for mobile wireless sensor networks // Sensors (Basel, Switzerland). 2009. V. 9, 11. P. 8961–8995.

9. *Llorca J., Milner S.D., Davis C.* Molecular System Dynamics for Self-Organization in Heterogeneous Wireless Networks // EURASIP J. on Wireless Communications and Networking. 2010. 10.1155/2010/548016. 13 p.

10. *Wai-kai C*. Net Theory And Its Applications: Flows In Networks. Imperial College Press (26 May 2003). 672 p.

11. *Wang H.* On the Extremal Wiener Polarity Index of Hückel Graphs // Computational and Mathematical Methods in Medicine. 2016. article ID 3873597. 6 p. http://dx.doi.org/10.1155/2016/3873597

12. *Xu X., Gao Y., Sang Y., Liang Y.* On the Wiener Indices of Trees Ordering by Diameter-Growing Transformation Relative to the Pendent Edges // Mathematical Problems in Engineering. 2019. article ID 8769428. 11 p. https://doi.org/10.1155/2019/8769428

13. The On-Line Encyclopedia of Integer Sequences (OEIS). http://oeis.org/

14. *Fischerman M., Hoffmann A., Rautenbach D., Székely L., Volkmann L.* Wiener index versus maximum degree in trees // Discrete Applied Mathematics. 2002. V. 122. Is. 1–3. P. 127–137.

15. *Бурдонов И.* Самотрансформация деревьев с ограниченной степенью вершин с целью минимизации или максимизации индекса Винера// Труды ИСП РАН. 2019. Т. 31. Вып. 4. С. 189–210.

GRAPH SELF-TRANSFORMATION MODEL BASED ON THE OPERATION OF CHANGE THE END OF THE EDGE

I. B. Burdonov

Ivannikov Institute for System Programming of the RAS, Moscow

igorburdonov@yandex.ru,igor@ispras.ru

Abstract

We consider a distributed network whose topology is described by an undirected graph. The network itself can change its topology, using special “commands” provided by its nodes. The work proposes an extremely local atomic transformation $a \rightarrow c \rightarrow b$ of a change the end c of the edge ac, “moving” along the edge cb from vertex c to vertex b. As a result of this operation, the edge ac is removed, and the edge ab is added. Such a transformation is performed by a “command” from a common vertex c of two adjacent edges ac and cb. It is shown that from any tree you can get any other tree with the same set of vertices using only atomic transformations. If the degrees of the tree vertices are bounded by the number d ($d \geq 3$), then the transformation does not violate this restriction. As an example of the purpose of such a transformation, the problems of maximizing and minimizing the Wiener index of a tree with a limited degree of vertices without changing the set of its vertices are considered. The Wiener index is the sum of pairwise distances between the vertices of a

graph. The maximum Wiener index has a linear tree (a tree with two leaf vertices). For a root tree with a minimum Wiener index, its type and method for calculating the number of vertices in the branches of the neighbors of the root are determined. Two distributed algorithms are proposed: transforming a tree into a linear tree and transforming a linear tree into a tree with a minimum Wiener index. It is proved that both algorithms have complexity no higher than 2*n*–2, where *n* is the number of tree vertices. We also consider the transformation of arbitrary undirected graphs, in which there can be cycles, multiple edges and loops, without restricting the degree of the vertices. It is shown that any connected graph with *n* vertices can be transformed into any other connected graph with *k* vertices and the same number of edges in no more than 2(*n*+*k*)–2.

Keywords*: distributed network, self-transformation of graphs, Wiener index*

REFERENCES

1. *Wiener H.* Structural determination of paraffin boiling points // J. Am. Chem. Soc. 1947. No 69 (1). P. 17–20.

2. *Kochkarov A.A., Sennikova L.I., Kochkarov R.A.* Nekotorye osobennosti primeneniia dinamicheskikh grafov dlia konstruirovaniia algoritmov vzaimodeistviia podvizhnykh abonentov // Izvestiia IuFU. Tekhnicheskie nauki, razdel V, cistemy i punkty upravleniia. 2015. No 1, S. 207–214 (in Russian).

3. *Proskochilo A.V., Vorobev A.V., Zriakhov M.S., Kravchuk A.S.* Analiz sostoianiia i perspektivy razvitiia samoorganizuiushchikhsia setei // Nauchnye vedomosti, seriia ekonomika, informatika. 2015. Vypusk 36/1. No 19 (216). S. 177–186 (in Russian).

4. *Pathan A.S.K.* (ed.). Security of self-organizing networks: MANET, WSN, WMN, VANET. CRC press, 2010. 638 p.

5. *Boukerche A.* (ed.). Algorithms and protocols for wireless, mobile Ad Hoc networks // John Wiley & Sons, 2008. 496 p.

6. *Chen Z., Li S., Yue W*. SOFM Neural Network Based Hierarchical Topology Control for Wireless Sensor Networks // Hindawi Publishing Corporation. J. of Sensors. 2014. article ID 121278. 6 p. http://dx.doi.org/10.1155/2014/121278

7. *Mo S., Zeng J.-C., Tan Y.* Particle Swarm Optimization Based on Self-organizing Topology Driven by Fitness // International Conference on Computational

Aspects of Social Networks, CASoN 2010, Taiyuan, China, 10.1109/CASoN. 2010. No 13. P. 23–26.

8. *Wen C.-Y., Tang H.-K.* Autonomous distributed self-organization for mobile wireless sensor networks // Sensors (Basel, Switzerland). 2009. V. 9, 11. P. 8961–8995.

9. *Llorca J., Milner S.D., Davis C.* Molecular System Dynamics for Self-Organization in Heterogeneous Wireless Networks // EURASIP J. on Wireless Communications and Networking. 2010. 10.1155/2010/548016. 13 p.

10. *Wai-kai C*. Net Theory And Its Applications: Flows In Networks. Imperial College Press (26 May 2003). 672 p.

11. *Wang H.* On the Extremal Wiener Polarity Index of Hückel Graphs // Computational and Mathematical Methods in Medicine. 2016. article ID 3873597. 6 p. http://dx.doi.org/10.1155/2016/3873597

12. *Xu X., Gao Y., Sang Y., Liang Y.* On the Wiener Indices of Trees Ordering by Diameter-Growing Transformation Relative to the Pendent Edges // Mathematical Problems in Engineering. 2019. article ID 8769428. 11 p. https://doi.org/10.1155/2019/8769428

13. The On-Line Encyclopedia of Integer Sequences (OEIS). http://oeis.org/

14. *Fischerman M., Hoffmann A., Rautenbach D., Székely L., Volkmann L.* Wiener index versus maximum degree in trees // Discrete Applied Mathematics. 2002. V. 122. Is. 1–3. P. 127–137.

15. *Bourdonov I.* Self-transformation of trees with a limited degree of vertices in order to minimize or maximize the Wiener index// Proceedings of ISP RAS. 2019. V. 31. No 4. P. 189–210 (in Russian).

СВЕДЕНИЯ ОБ АВТОРЕ

БУРДОНОВ Игорь Борисович – ведущий научный сотрудник Института системного программирования им. В.П. Иванникова РАН. Сфера научных интересов – моделирование и верификация программных систем, теория графов, теория автоматов

Igor Borisovich BURDONOV – *Leading Researcher, Ivannikov Institute for System Programming of the RAS. Research interests - modeling and verification of software systems, graph theory, automata*

email: igorburdonov@yandex.ru, igor@ispras.ru

Материал поступил в редакцию 30 октября 2019 года

УДК 004.021 + 004.42

БАЗОВЫЕ СЕРВИСЫ ФАБРИКИ МЕТАДАННЫХ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ LOBACHEVSKII-DML

П. О. Гафурова[1][0000-0002-1544-155X], А. М. Елизаров[2][0000-0003-2546-6897], Е. К. Липачёв[3][0000-0001-7789-2332]

[1-3]*Высшая школа информационных технологий и интеллектуальных систем Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

[2, 3]*Институт математики и механики им. Н.И. Лобачевского Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

[1]pogafurova@gmail.com, [2]amelizarov@gmail.com, [3]elipachev@gmail.com

Аннотация

Решен ряд задач, связанных с построением фабрики метаданных цифровой математической библиотеки Lobachevskii-DML. Под фабрикой метаданных понимается система взаимосвязанных программных инструментов, направленных на создание, обработку, хранение и управление метаданными объектов цифровых библиотек и позволяющих интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки. С целью выбора оптимальных таких программных инструментов из существующих и их модернизации:

- обсуждены особенности представления метаданных документов различных электронных коллекций, связанные как с применяемыми форматами, так и с изменениями состава и полноты набора метаданных в течение всего времени издания соответствующего научного журнала;
- представлены и охарактеризованы программные инструменты управления научным контентом и методы организации автоматизированной интеграции репозиториев математических документов с другими информационными системами;
- обсуждена такая важная функция фабрики метаданных цифровой библиотеки, как нормализация метаданных в соответствии с форматами других агрегирующих библиотек.

В результате разработки фабрики метаданных цифровой математической библиотеки Lobachevskii-DML предложена система сервисов автоматизированного формирования метаданных электронных математических коллекций; разработан xml-язык представления метаданных, основанный на Journal Archiving and Interchange Tag Suite (NISO JATS); созданы программные инструменты нормализации метаданных электронных коллекций научных документов в форматах, разработанных международными организациями – агрегаторами ресурсов по математике и Computer Science; разработан алгоритм приведения метаданных к формату oai_dc и генерации структуры архивов для импорта в цифровое хранилище DSpace; предложены и реализованы методы интеграции электронных математических коллекций Казанского университета в отечественные и зарубежные цифровые математические библиотеки.

***Ключевые слова**: цифровые библиотеки, цифровая математическая библиотека, формирование метаданных, извлечение метаданных, нормализация метаданных, фабрика метаданных, NISO JATS, семантические связи, Lobachevskii-DML.*

ВВЕДЕНИЕ

Как известно, навигация в научном информационном пространстве в значительной степени обеспечивается сегодня наличием и полнотой набора метаданных документов, представленных в сети [1–5]. Важной составляющей информационного научного пространства являются цифровые библиотеки (см., например, [6]). В области математических наук создано значительное число цифровых библиотек, выполняющих разнообразные функции интеграции математических знаний [7–12]. Обзор специфики и функциональных возможностей ряда существующих цифровых математических библиотек содержится в [13].

Комитетом по электронной информационной коммуникации (Committee on Electronic Information Communication, CEIC) Международного математического союза (International Mathematical Union, IMU) в 2002 году был подготовлен документ [14], определяющий лучшие практики распространения результатов математических исследований, которые учитывают новые достижения в области цифровых коммуникаций. Российских математиков в CEIC представлял академик А.Б. Жижченко. Среди первых цифровых математических библиотек, созданных в

соответствии с практиками, обозначенными в [14], наиболее крупными являются MathNet.Ru (http://www.mathnet.ru/), Numdam (http://www.numdam.org/) и DML-CZ (Czech Digital Mathematics Library, https://dml.cz/). Результаты их формирования и развития к 2013 году представлены в работах [15–19]. Одновременно для информационной поддержки цифровых библиотек, в том числе математических, стали разрабатываться методы обработки документов, основанные на семантических связях объектов, выделенных из их контента [20–27].

Заметным шагом в цифровизации математических знаний стали инициативы “World Digital Mathematics Library” (WDML, «Всемирная цифровая математическая библиотека», https://www.mathunion.org/ceic/library/world-digital-mathematics-library-wdml, 2012 год; предысторию этой инициативы подробно описал Peter J. Olver, см. http://www-users.math.umn.edu/~olver/t_/wdmlb.pdf) и “The Global Digital Mathematics Library” (GDML, «Глобальная цифровая математическая библиотека», 2014 год). Эти инициативы направлены на создание методологии интеграции математических знаний [28–31]. Основная их цель состоит в разработке принципов построения информационной среды, предоставляющей доступ ко всем когда-либо опубликованным работам по математике с помощью системы интеллектуальных агентов, обеспечивающих навигацию в информационном пространстве.

В направлении, обозначенном названными выше инициативами, инициирован и реализован ряд значимых международных проектов. Например, проект “The European Digital Mathematics Library” – «Европейская Цифровая Математическая Библиотека» (EuDML, https://initiative.eudml.org/ – направлен на интеграцию математических ресурсов европейских цифровых библиотек [32–34]. Далее, благодаря реализации проекта MathNet.Ru (http://www.math-net.ru/) оцифрованы, снабжены метаданными и представлены в открытый доступ архивы многих российских математических научных журналов и других изданий. На портале этого проекта предложены к использованию разработанные в рамках проведенных исследований методы навигации и расширенного поиска по математическому контенту, а также указаны возможности организации системы связей с международными библиографическими базами данных [15–17]. В Казанском университете,

начиная с 2017 года, в соответствии с базовыми принципами WDML и GDML создается цифровая математическая библиотека Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/) [35–37].

В настоящее время в связи со значительным ростом объемов научных публикаций появилась необходимость разработки специализированных методов автоматизированной обработки больших массивов документов [20, 25, 38–41]. Одним из результатов, полученных в этом направлении, является ряд разработанных семантических методов обработки математического контента [42–48]. Семантическая составляющая большинства используемых здесь методов основана на применении специализированных онтологий. Отметим, в частности, спроектированную и развиваемую в настоящее время цифровую экосистему OntoMath [45, 46]: она включает семантические поисковые инструменты [47, 48], онтологии профессиональной [42] и образовательной математики [49–51], рекомендательные системы, ориентированные на специфику математического контента [52–57], и набор инструментов для работы с математическими документами. Все перечисленное составляет основу фабрики метаданных цифровой математической библиотеки, разрабатываемой нами в настоящее время.

Мы используем термин *фабрика метаданных цифровой библиотеки* (*metadata factory of digital library*) в том же смысле, в каком он использован в [25], а именно: фабрика метаданных – это система взаимосвязанных программных инструментов, направленных на создание, обработку, хранение и управление метаданными объектов цифровых библиотек и позволяющих интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки. С помощью этих инструментов преимущественно в автоматизированном режиме выполняются такие операции, как выделение объектов и связей между ними, экстракция метаданных из различных источников и конкретных документов, верификация, уточнение, улучшение, нормализация в различных форматах и гармонизация метаданных с помощью ручного редактирования или автоматизированных агентов, хранение и связывание метаданных с внешними базами данных. В случае цифровой математической библиотеки к перечисленным инструментам добавляется ряд специализированных, таких, например, как преобразование в формат MathML, разметка математических формул и организация поиска по ним.

Для обозначения методов формирования и преобразования метаданных документов в соответствии с правилами и XML-схемами цифровых библиотек и наукометрических баз данных мы используем термин *нормализация* (см. [25, 58–60]).

В настоящей статье представлены базовые сервисы фабрики метаданных цифровой математической библиотеки Lobachevskii-DML. Для решения задачи интеграции информационных ресурсов предложены методы преобразования метаданных электронных коллекций и содержащихся в них документов по DTD-правилам и XML-схемам Journal Archiving and Interchange Tag Suite (NISO JATS) различных версий [61–66]. Так, например, для интеграции с ресурсами, сформированными в рамках проекта EuDML, необходимо, чтобы метаданные интегрируемых электронных коллекций были представлены в соответствии с EuDML-схемой на основе NLM JATS версии 1.0. Одновременно должен быть создан отдельный OAIPMH-сервер для реализации EuDML-харвестинга. Отметим также, что разработанный метод нормализации метаданных электронных коллекций цифровой библиотеки Lobachevskii-DML по правилам NISO JATS послужил основой для формирования обязательного и фундаментального наборов метаданных по схемам EuDML. Кроме того, в статье приведен алгоритм автоматизированной подготовки метаданных электронной коллекции статьей журнала «Электронные библиотеки» (“Russian Digital Libraries Journal”, https://elbib.ru/) по правилам библиографической базы по компьютерным наукам “dblp computer science bibliography” (DBLP, https://dblp.uni-trier.de/).

Статья организована следующим образом.

В Разделе 1 обсуждены особенности представления метаданных документов различных электронных коллекций, связанные не только с применяемыми форматами, но и с изменениями состава и полноты набора метаданных в течение всего времени издания соответствующего научного журнала.

В Разделе 2 представлены программные инструменты управления научным контентом. Эти инструменты используются и фабрикой метаданных для создания, обработки, хранения и управления метаданными электронных документов и позволяют интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки.

В разделе 3 затронуты вопросы организации автоматизированной интеграции репозиториев математических документов с другими информационными системами.

В разделе 4 обсуждена такая важная функция фабрики метаданных цифровой библиотеки, как нормализация метаданных в соответствии с форматами других агрегирующих библиотек. Реализация этой функции позволяет организовать взаимодействие сервисов как в рамках самой цифровой библиотеки, так и с внешними библиотеками и базами данных, для чего необходимо учитывать используемые в них форматы метаданных.

В разделе 5 решен ряд задач, связанных с построением фабрики метаданных цифровой математической библиотеки Lobachevskii-DML.

В Заключении сформулированы выводы из проведенных исследований и намечены планы их развития.

1. ОСОБЕННОСТИ ПРЕДСТАВЛЕНИЯ МЕТАДАННЫХ ДОКУМЕНТОВ В РАЗЛИЧНЫХ ЦИФРОВЫХ МАТЕМАТИЧЕСКИХ БИБЛИОТЕКАХ

В настоящее время многие наукометрические базы данных осуществляют индексацию статей, опубликованных в ведущих математических журналах за достаточно большие временные периоды их издания. Эти базы данных предъявляют различные требования к набору метаданных включаемых документов, а также схемам их представления (см., например, [67, 68]). Вместе с тем, необходимо отметить, что, как правило, пока еще не индексируются такие новые формы публикаций, как презентации, научные блоги и видеолекции – важные компоненты современных цифровых библиотек [69–71]. В последнее время электронные научные журналы стали поддерживать мультимедийный контент – многие их статьи содержат, помимо текста, формул и графики, также анимационные вставки [72, 73].

Цифровые математические библиотеки при формировании электронных коллекций, входящих в них, используют различные форматы метаданных. Это объясняется тем, что многие такие коллекции образуются из статей, опубликованных в академических журналах и оформленных в соответствии с правилами, уста-

новленными в них и отличающимися требованиями к используемым метаданным. Эти отличия касаются, прежде всего, состава метаданных и их формата и более всего заметны в архивных коллекциях научных журналов.

Наборы метаданных статей даже одного научного журнала в зависимости от года их публикации существенно отличаются. Примером служит архив статей журнала «Известия высших учебных заведений. Математика» (“Russian Mathematics (Izvestiya VUZ. Matematika)”, https://kpfu.ru/science/nauchnye-izdaniya/ ivrm), статьи которого составляют одну из электронных коллекций цифровой библиотеки Lobachevskii-DML. Этот журнал издается с 1957 года, а относительно полным набором метаданных сопровождены только статьи, опубликованные в нем, начиная с 2010 года. В статьях, опубликованных до 2008 года, отсутствуют ключевые слова и аннотации. История расширения набора метаданных указанного журнала приведена в Таблице 1.

Таблица 1. Состав метаданных статей журнала «Известия вузов. Математика» в период 1957–2020 годов

Год	Состав метаданных
1957–1965	Только основные метаданные: ФИО авторов, название статьи, организация (университет, институт) или город (например, Москва), дата поступления статьи
1965–1969	Начиная со второго номера 1965 года, статьи дополнительно снабжены классификаторами УДК
1970–2007	Из многих статей можно выделить блок с кратким описанием (аннотацию)
2008–2020	Начиная с третьего номера 2008 года, появились блоки «Аннотация» и «Ключевые слова», указаны e-mail адреса авторов и адреса организаций (не всегда)

Отметим, что архивы журнала «Известия вузов. Математика», как и ряда других отечественных научных журналов, оцифрованы и размещены на портале

Mathnet.Ru, причем набор метаданных включенных статей расширен по сравнению с тем, который был изначальным: в дополнение к метаданным, приведенным в Таблице 1, указаны ссылки на английские переводы статей (с 1975 года), выделены блоки литературы, списки статей, ссылающихся на данную статью, а также ссылки на описания статьи в реферативных и наукометрических базах данных MathSciNet и Scopus.

Расширение набора используемых метаданных, описанное выше, является типичным практически для всех научных журналов. В последнее время в ряде научных журналов обязательным является указание ORCID (Open Researcher and Contributor ID) каждого автора [74]. Этот идентификатор предназначен для решения проблемы возможной неоднозначности представления имен и фамилий авторов публикаций в сетях научной коммуникации.

Для пополнения набора метаданных разрабатываются методы извлечения метаданных из документов [20, 25, 39, 68]. Также возникла необходимость в методах нормализации метаданных, позволяющих преобразовать уже созданные метаданные в форматы наукометрических баз данных (методы нормализации метаданных подробно описаны ниже). Отметим также, что обязательным условием участия в интеграционных проектах, таких, например, как EuDML, является предоставление наборов метаданных, сформированных по схемам, установленным в этих проектах.

Статьи из ведущих российских математических журналов переводятся на английский язык и, как правило, имеют библиографическое описание, не совпадающее с имеющимся в русскоязычном издании. В Таблице 2 приведен пример, характеризующий возникающие при этом проблемы. Русскоязычный источник в списке литературы англоязычной статьи либо транслитерируется, либо указывается ссылка на переводную версию документа. Однако, как представлено в Таблице 2, библиографическое описание переведенной статьи может отличаться от транслитерированного описания не только в названии статьи – могут быть совсем иными название журнала, а также номер выпуска и диапазоны страниц статьи. Таким образом, транслитерированная статьи и ее переводная версия могут быть восприняты как разные документы. В настоящее время схемы метаданных, основанные на NISO JATS и используемые в цифровой библиотеке EuDML, не позво-

ляют соединить в рамках одного метаописания статью, опубликованную на русском языке, и ее переводную версию на английском языке [75, 76]. В коллекциях Lobachevskii-DML, а также цифровых библиотеках eLibrary.ru и MathNet.Ru такие статьи представлены как дубликаты одного документа.

При формировании электронной коллекции статей научного журнала приходится также учитывать, что в определенные периоды времени этот журнал мог выпускаться под другим названием. Так, например, переводная версия журнала «Известия высших учебных заведений. Математика» до 1991 года имела название “Soviet Mathematics (Izvestiya VUZ. Matematika)“, а, начиная с 1992 года, – “Russian Mathematics (Izvestiya VUZ. Matematika)“.

Таблица 2. Как отличаются метаданные в статье и её переводных версиях

Цитирование на русском языке	А.М. Елизаров, А.Б. Жижченко, Н.Г. Жильцов, А.В. Кириллович, Е.К. Липачёв, «Онтологии математического знания и рекомендательная система для коллекций физико-математических документов», Докл. РАН, 467:4 (2016), 392–395
Транслитерированное цитирование (русскоязычный источник в англоязычной статье)	A.M. Elizarov, A.B. Zhizhchenko, N.G. Zhiltsov, A.V. Kirillovich, E.K. Lipachev, “Ontologii matematicheskogo znaniya i rekomendatelnaya sistema dlya kollektsiy fiziko-matematicheskikh dokumentov”, Dokl. RAN, 467:4 (2016), 392–395
Цитирование переведенной статьи	A.M. Elizarov, A.B. Zhizhchenko, N.G. Zhiltsov, A.V. Kirillovich, E.K. Lipachev, “Mathematical knowledge ontologies and recommender systems for collections of documents in physics and mathematics”, Dokl. Math., 93:2 (2016), 231–233

Представление правильного варианта цитирования статьи важно при включении таких документов в международные агрегирующие базы данных (например, Scopus), так как между транслитерированным и переведенным названиями

практически нет связи. Отметим, что сегодня во многих журналах для каждой публикуемой статьи приводится готовый вариант ее библиографического описания для последующего цитирования другими авторами. Это описание также является частью метаданных этого документа.

Одним из недостатков формата представления метаданных, предложенного EuDML, является отсутствие средств поддержки нескольких версий одной и той же публикации. Поэтому с использованием этого формата нельзя описать не только новые формы публикации, например, динамические и живые публикации, но и приходится описывать переведенные статьи как различные статьи в разных журналах. В Таблице 2, приведенной выше, показано, как статья и ее перевод на английский язык могут быть представлены как совершенно разные документы. В этом контексте также встает вопрос о препринтах: считать ли ссылку на препринт статьи, размещенный в свободном доступе, ссылкой на статью в цифровой библиотеке; возможно ли ассоциированное размещение в цифровой библиотеке ссылок на оригинальный текст статьи и ее препринт.

Таким образом, при формировании инструментов фабрики метаданных, отвечающих за выбор формата метаданных для описания электронной коллекции, включаемой в цифровую библиотеку, обязательно нужно учитывать историю развития индексируемого научного издания и связи исходных и переводных версий соответствующих документов.

Другой важной составляющей любой цифровой библиотеки являются программные инструменты управления научным контентом. Эти инструменты используются и фабрикой метаданных для создания, обработки, хранения и управления метаданными электронных документов и позволяют интегрировать создаваемые электронные коллекции в агрегирующие цифровые научные библиотеки. Опишем подробнее существующие решения.

2. ПРОГРАММНЫЕ ИНСТРУМЕНТЫ УПРАВЛЕНИЯ НАУЧНЫМ КОНТЕНТОМ

Создание цифровой библиотеки и последующее расширение её функциональных возможностей предполагают решение целого ряда трудоемких задач, связанных, в том числе, с управлением научным контентом (см., например, [6]).

Структура наиболее известных цифровых математических библиотек и разработанные в них сервисы управления математическим контентом обсуждены в [12, 14].

Существующие цифровые библиотеки, а также агрегаторы научных знаний предлагают ряд программных инструментов работы с контентом, прежде всего, сервисы поиска в электронных коллекциях. Например, средства семантического поиска документов представлены на сайте проекта EuDML (https://initiative.eudml.org/). Здесь же размещены демонстрационные версии инструментов, разработанных для обслуживания EuDML (см. Таблицу 3).

Таблица 3: Демонстрационные версии инструментов EuDML

Название инструмента	Цель создания	Режим доступа и возможности демо-версии
PDF Text Extractor	Экстракция и распознавание текста PDF-документа	Демоверсия (https://initiative.eudml.org/EuDmlAnalysisDemo/) извлекает текст только из распознанных документов
Maxtract	Создан для анализа PDF-документа и выделения в нем математических конструкций в виде списка MathML	В демоверсию (https://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/eudml-demo.php) включены 10 страниц (с. 11–20) книги “Semi-Riemann Geometry and General Relativity – Schlomo Sternberg, 2003”. Есть возможность получить в MathML-формате всю страницу или только определенные строки
TeX2NLM	Принимает на входе T_EX-строку в UTF-8-кодировке и возвращает то же содержи-	Доступ не предоставлен

	мое в TEX- и MathML-представлениях в соответствии со структурой EuDML NLM	
Enhance NLMTeXw-MathML	Пакетный инструмент, работающий с действительными XML-документами и предназначенный для обновления метаданных с помощью (представления) MathML для любой формулы, написанной на TEX, при условии, что функции TEX известны компилятору	Доступ не предоставлен
Plain Text Reference Segmenter	Выделяет библиографические ссылки из текста	Совмещен с сервисом PDF Text Extractor. Выделение не всегда корректно (https://initiative.eudml.org/EuDmlAnalysisDemo/)
Bibliographic Reference Parser	Разделяет библиографические ссылки, выделяет такие метаданные, как имена и фамилии авторов, названия публикации, год публикации и т. д.	Совмещен с сервисом PDF Text Extractor. Работает не всегда корректно (https://initiative.eudml.org/EuDmlAnalysisDemo/)
Find similar articles via Gensim	Поиск сходства между статьями в коллекциях arXiv.org с использованием библиотеки Gensim	https://mir.fi.muni.cz/eudmldemo/gensim-arxiv/)
MIaS4gensim demo	Формирует списки терминов и математических формул по их встречаемости в тек-	Словари не полностью очищены от стоп-слов. Пример из полного списка терминов данной библиотеки.

	сте, анализируя статьи физико-математической направленности в arxiv.org	23659 anybody 35 49385 anyhow 91 67484 anymore 518 82158 anyone 157 88230 anything 999 4840 anytime 18 18032 anyway 716 86702 anywhere 408 https://mir.fi.muni.cz/eudmldemo/mias4gensim/

Более подробно назначение и функциональные возможности приведенных в таблице программных инструментов описаны в [77–79]. Как видно из Таблицы 3, часть перечисленных инструментов имеет непосредственное отношение к соответствующей фабрике метаданных.

Рассмотрим теперь программные инструменты, представленные в национальных цифровых математических библиотеках.

В рамках проекта «Общероссийский математический портал MathNet.Ru» разработаны сервисы поиска статей, персоналий и организаций, обеспечен полный доступ к статьям. Статьи связаны ссылками с их англоязычными версиями и дополнены соответствующими метаданными. Разработан пакет amsbib, предназначенный для оформления библиографии в TEX-нотации, а также сервис MiRef для работы со списками литературы. Выделены связи со статьями, цитирующими данную статью. Построена гибкая система учета дублирований при цитировании. Разработана и поддерживается система видеопубликаций. Предложена система электронного документооборота редакций журналов [15–17].

Новая платформа французской цифровой математической библиотеки Numdam (http://www.numdam.org/) содержит фабрику метаданных, включающую сервисы нормализации метаданных, их корректировки и улучшения, а также инструменты работы с формулами в TEX- и MathML-форматах [25].

С 2005 года развивается проект чешской цифровой математической библиотеки DML-CZ (https://dml.cz/). Как отмечено в [19], при создании этой библиотеки были учтены успешные решения, реализованные ранее в проекте Numdam. Первоочередная задача проекта DML-CZ состояла в создании цифрового архива

математической литературы, изданной в Чехии. Эта цифровая библиотека содержит практически все математические журналы, изданные чешскими издателями с девятнадцатого века. С самого начала проекта разработчики этой цифровой библиотеки рассматривали её как один из строительных блоков для предполагаемой глобальной DML. Программные инструменты были созданы с учётом этого обстоятельства, и это можно заметить по списку разработчиков сервисов EuDML (см., например, [80]). Названная цифровая библиотека интегрирована в EuDML.

Одной из важнейших задач проекта DML-CZ было также создание комплексной программной системы управления контентом цифровой библиотеки. Эта система содержит программные инструменты, адаптированные к потребностям DML-CZ. Основным инструментом является редактор метаданных [18], объединяющий все действия, связанные с обработкой цифрового контента, созданием метаданных и взаимосвязанными источниками информации.

Отметим, что в DML-CZ реализованы сервисы, учитывающие математическую специфику контента, в частности, имеется возможность классификации документов по кодам AMS Math Classification [38, 80].

3. ОРГАНИЗАЦИЯ ХРАНИЛИЩА ЦИФРОВЫХ КОЛЛЕКЦИЙ

Одной из важных задач при работе с цифровыми математическими библиотеками является автоматизированная интеграция репозиториев математических документов с другими информационными системами. Такой процесс основан на модели агрегирования и распространений метаданных. Модель OAI Protocol for Metadata Harvesting (далее OAI-PMH) [81] поддерживается большинством систем, предназначенных для хранения информационных ресурсов.

Для организации работы с OAI-PMH необходимо использовать систему поддержки цифрового хранилища. Обзор различных цифровых хранилищ приведен в [82]. Наиболее известными из них являются DSpace, Eprints, Fedora и Greenstone. Некоторые библиотеки имеют специализированные методы харвестинга метаданных из других хранилищ, в этом случае необходимо, чтобы у поставщиков данных были инструменты и сервисы, которые позволяют распространять метаданные.

Цифровые математические библиотеки DML-CZ, EuDML и Numdam базируются на информационной платформе DSpace. Это инструмент с открытым исходным кодом, предназначенный для реализации цифровых библиотек и репозиториев [83]. DSpace предлагает большинство базовых функций цифровых библиотек и сервисов, включая пользовательский интерфейс, индексирование, поиск документов, просмотр источников информации, постоянная идентификация документов, предоставление метаданных для сбора данных по протоколу OAI-PMH, поддержка долгосрочного сохранения цифровых данных.

Для решения наших задач важным обстоятельством является совместимость DSpace с издательской платформой Open Journal System (OJS) [84], что позволяет реализовать модель «хранилище + система» работы научного журнала, включая операции с архивом журнала. Обмен данными происходит через OAI-PHM сервер, что позволяет автоматически производить харвестинг метаданных.

4. НОРМАЛИЗАЦИЯ МЕТАДАННЫХ

Для организации взаимодействия сервисов как в рамках цифровой библиотеки, так и с внешними библиотеками и базами данных, прежде всего, требуется учитывать используемые в них форматы метаданных. Даже в одной цифровой библиотеке программные инструменты работают с несколькими форматами метаданных, что связано с особенностями формирования цифрового контента (например, методов оцифровки), а также требованиями агрегирующих цифровых библиотек и наукометрических баз данных. Отметим лишь наиболее распространенные форматы метаданных, с которыми приходится иметь дело при организации взаимодействия сервисов в цифровых математических библиотеках.

Прежде всего, это формат Dublin Core и его расширения [85], формат каталогизации MARC [86, 87], формат библиографических ссылок RIS (Research Information Systems), AMSBib [15], XML-форматы РИНЦ [88–90], NISO JATS [61–66], форматы на основе схем DBLP [91].

Одной из функций фабрики метаданных является нормализация метаданных в соответствии с форматами других агрегирующих библиотек. Как было отмечено во Введении, термином «нормализация метаданных» обозначают систему сервисов преобразования или формирования метаданных в соответствии с регла-

ментирующими XML-схемами или DTD-правилами целевой цифровой библиотеки [25, 58–60, 68, 92–95]. Так, например, протокол OAI-PMH требует обязательного включения в описание ресурса набора метаданных в нотации oai_dc, которая основана на Dublin Core. Данная нотация использует только названия ограниченного количества тегов Dublin Core [85].

Для описания статей из математических журналов в цифровой математической библиотеке EuDML применены XML-схемы NISO JATS V1.0 [62], а общая схема метаданных этой цифровой библиотеки описана в [75, 76]. Выделены три набора метаданных: обязательные (obligatory metadata), фундаментальные (fundamental metadata) и дополнительные (supplemental metadata). Из них минимальным по составу является обязательный набор метаданных, в который включены: название статьи на языке оригинала, список авторов, библиография, уникальный идентификатор статьи, например, doi и URL полного текста статьи. Фундаментальный набор метаданных дополнительно к обязательным метаданным включает аннотацию статьи и ключевые слова.

Специализированной базой данных по компьютерным наукам является DBLP (https://dblp.uni-trier.de/) [91]. Необходимым условием включения электронных коллекций в эту базу данных являются реорганизация и нормализация метаданных документов включаемой цифровой библиотеки. Передача метаданных электронных коллекций (отдельные статьи не индексируются) в базу DBLP осуществляется автоматическими средствами либо вручную. Схема описания статьи из электронной коллекции включает в себя такие метаданные, как название, ФИО авторов, номера страниц, год выпуска и название конференции/журнала, в котором она была размещена. Индексируемый сборник конференции имеет такие метаданные: ФИО редакторов, название конференции (полное и сокращенное), место приведения конференции.

Определенные затруднения при подготовке метаданных в DBLP возникают с коллекциями, содержащими статьи, не имеющие переводов или, хотя бы, транслитерацию на английский язык таких важных метаданных, как название статьи, список авторов, набор ключевых слов, аффилиация и аннотация. Соответствующие сервисы, автоматизирующие эти процессы, либо являются частью фабрики метаданных цифровой библиотеки, либо онлайн-сервисами других цифровых библиотек.

Отметим, что процесс подготовки метаданных в формате eLibrary.ru автоматизирован (см. [87–90, 96]), а соответствующие сервисы включены в фабрику метаданных цифровой математической библиотеки Lobachevskii-DML, в частности, с их помощью формируются метаданные журнала «Электронные библиотеки».

5. ФАБРИКА МЕТАДАННЫХ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ

В настоящем разделе мы предлагаем решения ряда задач, связанных с построением фабрики метаданных в рамках проекта создания цифровой математической библиотеки Lobachevskii-DML [35–37]. Как и в случае любой цифровой научной библиотеки, формирование библиотеки Lobachevskii-DML и соответствующей фабрики метаданных потребовало привлечения ранее созданных, а также разработки новых технологических решений управления научным контентом.

На Рис. 1 представлена IDEF0-схема фабрики метаданных цифровой библиотеки Lobachevskii-DML.

На этапе препроцессорной обработки выполняется отсев тех документов, которые не получилось обработать в автоматическом режиме, с указанием возникших проблем в файле отчета, сформированном автоматически. Также на этом этапе производятся исправление некоторых ошибок орфографии, неправильного выбора регистра, а также удаление лишних пробелов и знаков.

На этапе экстракции метаданных обрабатываются полные тексты документов, используются шаблоны поиска обязательных метаданных.

На этапе верификации метаданных выполняется проверка полноты и соответствия состава выделенных метаданных установленным правилам, записанным в виде DTD-файлов или XML-схем. После прохождения этапа верификации возможны три варианта дальнейших действий: дополнительная экстракция необходимых и дополнительных метаданных; повторная верификация метаданных и выдача отчета о том, что документ недостаточен для получения требуемых метаданных; переход к финальному действию – нормализации метаданных.

Экстракция дополнительных метаданных направлена на извлечение метаданных не только из самого документа, но и с помощью внешних ресурсов (например, персональных страниц авторов).

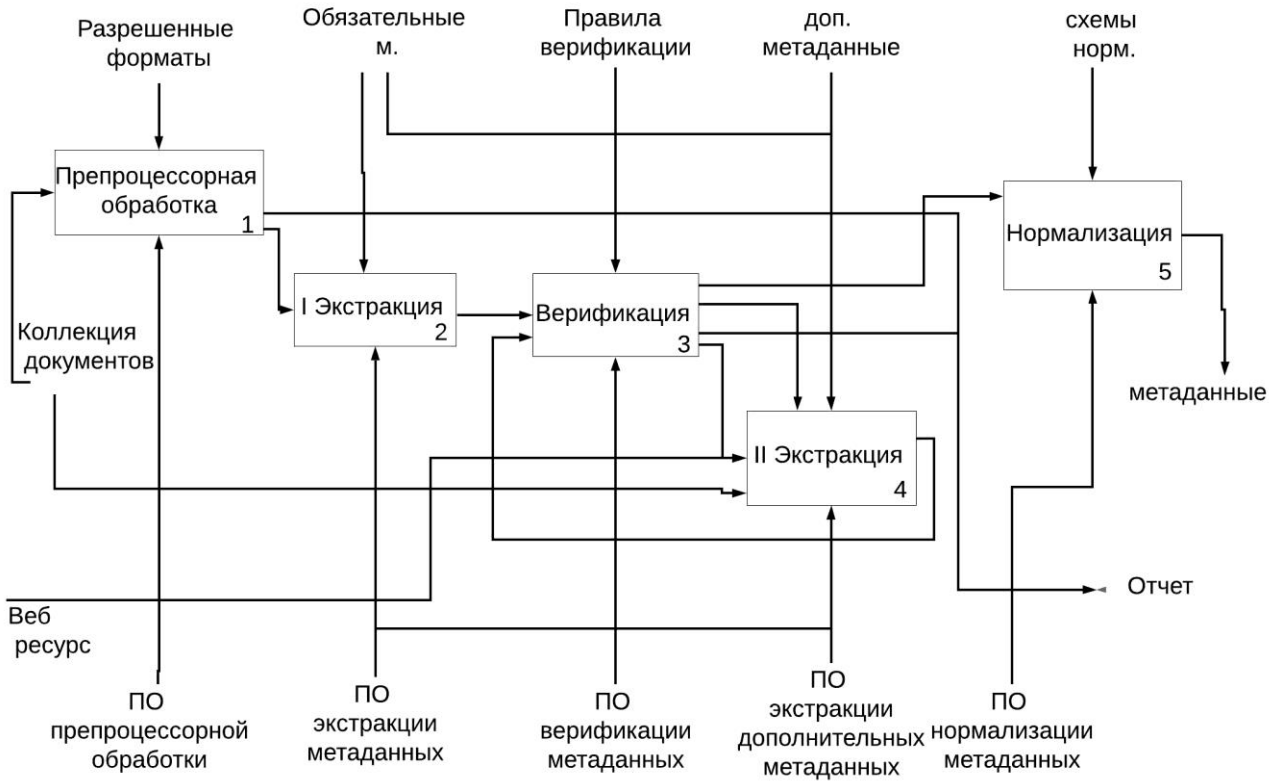


Рис. 1. Этапы формирования метаданных в фабрике метаданных цифровой математической библиотеки

Фабрика метаданных подразумевает, что последние собираются для конкретной цифровой библиотеки. Однако модель такой фабрики может быть распространена на процесс подготовки метаданных и для сайтов-агрегаторов математических знаний. В этом случае изменяются правила верификации метаданных, а также форматы, в которых будет происходить нормализация (некоторые результаты в этом направлении представлены в работах [92, 93]).

Ниже представлены некоторые из уже реализованных нами инструментов фабрики метаданных.

5.1. ЭКСТРАКЦИЯ И НОРМАЛИЗАЦИЯ МЕТАДАННЫХ

Размещение метаданных в интернете привело к тому, что одним из их источников могут стать веб-страницы сайта-агрегатора метаданных или самой цифровой библиотеки. Таким образом, при формировании фундаментального набора метаданных электронных коллекций, а также при получении дополнительных метаданных необходимо использовать метаданные, хранящиеся на

внешних ресурсах. Эта задача сопряжена с задачами поиска информации в агрегирующих базах данных и цифровых библиотеках, некоторые из которых частично закрыты для доступа или прерывают соединение, позволяя скачивать только ограниченное количество метаданных. При поиске метаданных на страницах сайтов-агрегаторов нужно также понимать и учитывать, что выбор и порядок поиска в таких источниках должны быть определены заранее, так как некоторые источники хранят информацию только по конкретной тематике (например, библиографическая база данных DBLP) или же неполный список метаданных.

Один из частных случаев экстракции метаданных с сайтов-агрегаторов разработан нами на основе сайта проекта MathNet. Цель созданного программного приложения – выделение и запись метаданных статьи на русском и английском языках с дальнейшей нормализацией по формату, принятому в EuDML. Основные шаги алгоритма экстракции и нормализации метаданных на примере коллекции журнала «Известия вузов. Математика» (“Russian Mathematics”) приведены в [94].

Задача перевода метаданных из одного формата в другой часто связана с задачами дополнения или улучшения метаданных. Так, например, в цифровой библиотеке Lobachevskii DML возникла необходимость перевода метаданных электронной коллекции статей журнала «Электронные библиотеки» (“Russian Digital Libraries Journal”, https://elbib.ru/) в формат базы данных DBLP. Процесс перевода включал семантическую транслитерацию имен и фамилий авторов статей. Исходные наборы метаданных, использованных при переводе в названной формат, были сформированы автоматически с помощью программных инструментов, разработанных в редакции журнала «Электронные библиотеки» (http://ojs.kpfu.ru/index.php/elbib), и средств программной платформы OJS [84], на которой функционирует данный журнал. Алгоритм перевода этих метаданных в формат DBLP был успешно реализован, подробно он подробно в [94, 95].

Одним из самых распространенных форматов метаданных, принятых в цифровых библиотеках, является формат Dublin Core. В соответствии с концепцией гармонизации метаданных, цифровые библиотеки должны иметь возможность создавать для своих электронных коллекций метаданные в различных форматах. Поэтому была поставлена задача создания инструмента для автоматизированного перевода имеющихся метаданных в формат Dublin Core. Данная задача была

решена на примере коллекции «Трудов Математического центра им. Н.И. Лобачевского» (далее – «Труды»). Эта коллекция насчитывает более 60 томов на русском и английском языках, имеющиеся метаданные статей разнородны.

Прежде всего, была произведена кластеризация, в результате которой соответствующие тома «Трудов» были разделены на классы по сходству их структуры и оформления. Для каждого класса был разработан набор паттернов регулярных выражений, задающих правила поиска информационных блоков. Далее производилась обработка массива файлов «Трудов» с целью выделения метаданных, описывающих как том в целом, так и статьи, входящие в него. В частности, определялись номера страниц всех статей каждого тома. С помощью методов текстового анализа из документов электронной коллекции были выделены термины, из которых были образованы наборы ключевых слов для включения в состав метаданных [97, 98]. Следующий этап включал процедуры разделения каждого тома «Трудов» на отдельные статьи. Далее следовал этап нормализации метаданных. Алгоритмы 1 и 2 нормализации метаданных в различные форматы представлены ниже.

Алгоритм 1: Нормализация метаданных по схемам DBLP

```
1: Загрузить VolCollection     \\ коллекция xml файлов метаописаний
2: for each volume in VolCollection do
3     for each paper  in volume do
4.      Считать из paper  значения тегов: author's_names,  title,
page_numbers, year_of_issue, url, volume
5.      Считать cite_page из https://elbib.ru/en/year/+year
6.      Split cite_page
7.      Выделить metadata: author's_names in English, url in
elbib.kpfu.ru.
8.     Split author's_names
9.   Answer:=Form(author's_name, Transliteration(name), ti-
tle, page_numbers, url, volume);
10.     Записать Answer in dblp.xml
11.    end for
12. end for
```

Алгоритм 2: Нормализация метаданных в формат Dublin Core

```
1: Загрузить VolCollection                    \\ коллекция xml файлов
2: for each volume from VolCollection do
3:     Прочитать numbervolume из названия файла volume
4:     Считать из файла info.csv переменные publisher, issue, year
5:     Papers:=new string List
6.     for each paper from volume do
7.         Считать из paper переменные: author's names, title,
page numbers
8.         Paper:=Formoai_dc(authors's names, title, issue
year, page numbers, publisher)
\\ функция формирует строковое описание статьи в формате Dublin Core
9.         Papers.Add(Paper)
10.    end for
11.    Создать файл с названием numbervolume
12.    for each paper from Papers do
13.        Записать paper в файл numbervolume.txt
14.    end for
15. end for
```

5.2. ХРАНЕНИЕ ЭЛЕКТРОННЫХ КОЛЛЕКЦИЙ

Одной из основных функций фабрики метаданных является подготовка электронных коллекций для загрузки в цифровую библиотеку. При использовании системы DSpace нужно было решить задачу автоматического формирования файлов и их загрузки в DSpace. Загрузка метаданных в DSpace производилась следующим образом.

Формируется файл-таблица в формате csv (Comma-Separated Values), в который записываются метаданные, подготовленные по схеме Dublin Core. Используется также способ передачи архива в формате Simple Archive Format. Кроме того, используются возможности загрузки метаданные через консоль и ручного ввода метаданных на сайте цифровой библиотеки.

Как показал наш опыт, наиболее рационально использовать загрузку архивом: основными достоинствами являются простота загрузки всех файлов одним архивом, а также возможность загрузить не только метаданные, но и сами файлы.

Далее встает задача нормализации метаданных в тот формат, который принят в DSpace. Подробный обзор методов подготовки метаданных приведен [94, 99].

ЗАКЛЮЧЕНИЕ

С целью интеграции электронных математических коллекций Казанского университета в международное научное информационное пространство разработаны алгоритмы и инструменты создания, обработки, хранения и управления метаданными объектов этих электронных коллекций, позволяющие включать создаваемые электронные коллекции в цифровую математическую библиотеку Lobachevskii-DML. Названные алгоритмы и инструменты составляют фабрику метаданных библиотеки Lobachevskii-DML и обеспечивают формирование метаданных этих коллекций и документов, входящих в них, в соответствии с форматами международных цифровых математических библиотек и наукометрических баз данных, а также дают возможность организовать взаимодействие названных сервисов как в рамках самой цифровой библиотеки, так и с внешними библиотеками и базами данных. Дальнейшее направление развития заключается в совершенствовании созданной фабрики метаданных и разработке возможности ее использования в любых научных цифровых библиотеках.

Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований и Правительства Республики Татарстан в рамках проекта № 18-47-160012 и в рамках программы развития Регионального научно-образовательного математического центра Приволжского федерального округа, номер соглашения № 075-02-2020-1478/1. Настоящая статья содержит также результаты, полученные в рамках проекта «Мониторинг и стандартизация развития и использования технологий хранения и анализа больших данных в цифровой экономике Российской Федерации», выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору Московского государственного университета имени М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 15.08.2019 № 7/1251/2019.

СПИСОК ЛИТЕРАТУРЫ

1. *Sicilia M.-A. (Ed)* Handbook of Metadata, Semantics and Ontologies. World Scientific Publishing Co. Pte. Ltd., 2014. 579 p.

2. *Alemu G., Stevens B.* An Emergent Theory of Digital Library Metadata. Enrich then Filter. Chandos Publishing is an imprint of Elsevier. 2015, 121 p. URL: http://store.elsevier.com/An-Emergent-Theory-of-Digital-Library-Metadata/ Getaneh-Alemu/isbn-9780081003855/.

3. *Gartner R.* Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer International Publishing Switzerland, 2016. 118 p. URL: https://doi.org/ 10.1007/978-3-319-40893-4.

4. *Когаловский М.Р.* Метаданные, их свойства, функции, классификация и средства представления // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15–18 октября 2012 г. Переславль-Залесский, 2012. С. 3–14. URL: http://rcdl.ru/doc/ 2012/paper3.pdf.

5. *Когаловский М.Р.* Метаданные в компьютерных системах // Программирование. 2013. Т. 39 (4). С. 28–46. URL: http://www.ipr-ras.ru/articles/ kogalov13-03.pdf.

6. *Xie I., Matusiak K.K.* Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016. 388 p.

7. *Jackson A.* The Digital Mathematics Library // Notices Amer. Math. Soc. 2003. V. 50. P. 918–923.

8. *Bouche T.* Introducing the mini-DML project // ECM4 Satellite Conference EMANI/DML. 2004. P. 19–29.

9. *Borwein J.M., Rocha E.M., Rodrigues J.F.* (eds.) Communicating Mathematics in the Digital Era. Taylor & Francis, 2008. 325 p.

10. *Bouche T*. Some Thoughts on the Near-Future Digital Mathematics Library // Towards a Digital Mathematics Library. Masaryk University, 2008. P. 3–15. URL: https://eudml.org/doc/221606.

11. *Bouche T.* Digital Mathematics Libraries: The Good, the Bad, the Ugly // Math. Comput. Sci. 2010. V. 3. P. 227–241. URL: https://doi.org/10.1007/s11786-010-0029-2.

12. *Bouche T.* The Digital Mathematics Library as of 2014 // Notices Amer. Math. Soc. 2014. V. 61. No 9. P. 1085–1088.

13. *Elizarov A.M., Lipachev E.K., Zuev D.S.* Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. V. 2022. P. 317–325.

14. Committee on Electronic Information Communication of the International Mathematical Union, Best current practices: Recommendations on electronic information communication // Notices of the AMS. 2002. 49(8), pp. 922–925. URL: http://ams.org/notices/200208/commpractices.pdf.

15. *Жижченко А.Б., Изаак А.Д.* Информационная система Math-Net.Ru. Применение современных технологий в научной работе математика // Успехи математических наук. 2007. Т. 62, №5 (377). С. 107–132. URL: https://doi.org/10.4213/rm8147. URL: http://www.mathnet.ru/links/c59aff2f134382372f88aa415a76755f/rm8147.pdf.

16. *Жижченко А.Б., Изаак А.Д.* Информационная система Math-Net.Ru. Современное состояние и перспективы развития. Импакт-факторы российских математических журналов // Успехи математических наук. 2009. Т. 64, №4 (388). С. 195–204. URL: https://doi.org/10.4213/rm9312; http://www.mathnet.ru/links/e27ab619eaefe03fe79d663468ddd3a0/rm9312.pdf.

17. *Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A., Zhizhchenko A.B.* Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics // Lecture Notes in Computer Science. 2013. V. 7961. P. 344–348. URL: https://doi.org/10.1007/978-3-642-39320-4_26.

18. *Bartošek M., Kovář P., Šárfy M.* DML-CZ Metadata Editor // In: Sojka P. (ed.) Towards Digital Mathematics Library. Masaryk University, 2010. P. 139–151. URL: https://dml.cz/bitstream/handle/10338.dmlcz/702537/DML_001-2008-1_17.pdf.

19. *Bartošek M., Rákosník J.* DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library // Notices of the AMS. 2013. V. 60. No 8. P. 1028–1033. URL: http://dx.doi.org/10.1090/noti1031.

20. *Bouche T.* Toward a digital mathematics library? // In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds) Communicating Mathematics in the Digital Era. Taylor & Francis, 2008. P. 47–73. URL: https://hal.archives-ouvertes.fr/hal-00347682.

21. *Lange C.* Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web // Semantic Web. 2013. V. 4. No. 2. P. 119–158. URL: https://content.iospress.com/articles/semantic-web/sw059.

22. *Серебряков В.А.* Что такое семантическая цифровая библиотека // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г. Дубна: ОИЯИ, 2014. С. 1–5. URL: http://rcdl.ru/doc/2014/paper/RCDL2014_021-25.pdf.

23. *Елизаров А.М., Кириллович А.В., Липачёв Е.К., Невзорова О.А.* Управление математическими знаниями: онтологические модели и цифровые технологии // Аналитика и управление данными в областях с интенсивным использованием данных: сборник статей XVIII Междун. конф. DAMDID/RCDL'2016. М.: ФИЦ ИУ РАН, 2016. С. 95–101.

24. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Mathematical Knowledge Management: Ontological Models and Digital Technology // CEUR Workshop Proceedings. 2016. V. 1752. P. 44–50. URL: http://ceur-ws.org/Vol-1752/paper08.pdf

25. *Bouche T., Labbe O.* The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. Vol. 10383. Springer, Cham, 2017. P. 70–82. URL: https://doi.org/10.1007/978-3-319-62075-6_6. https://zenodo.org/record/581405/.

26. *Sadegh A., Lange C., Vidal M.E., Auer S.* Integration of Scholarly Communication Metadata using Knowledge Graphs // International Conference on Theory and Practice of Digital Libraries. 2017. P. 328–341.

27. *Ataeva O.M., Serebryakov V.A.* Information Model of LibMeta Digital Library // Lobachevskii J. of Mathematics. 2019. V. 40. No 7. P. 861–875. URL: https://doi.org/10.1134/S1995080219070035.

28. Developing a 21st Century Global Library for Mathematics Research // Washington: The National Academies Press, 2014. 142 p. doi:10.17226/18619.

29. *Olver P.J.* The World Digital Mathematics Library: Report of a Panel Discussion // Proceedings of the International Congress of Mathematicians, August 13–

21, 2014, Seoul, Korea. Kyung Moon SA, 1. 2014. P. 773–785.

30. *Watt S.* How to Build a Global Digital Mathematics Library // 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). 2016. P. 37–40.

31. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. 2017. V. 10383. P. 56–69. URL: https://doi.org/10.1007/978-3-319-62075-6_5.

32. *Bouche T.* Reviving the free public scientific library in the digital age? The EuDML project // In: Kaiser K., Krantz S.G., Wegner B. (Eds) Topics and Issues in Electronic Publishing JMM/AMS Special Session, FIZ Karlsruhe. 2013. P. 57–80. URL: https://www.emis.de/proceedings/TIEP2013/05bouche.pdf.

33. *Bouche T., Rákosník J.* Report on the EuDML External Cooperation Model // In: Kaiser K., Krantz S.G., Wegner B. (Eds) Topics and Issues in Electronic Publishing, JMM, Special Session. San Diego. 2013. P. 99–108. URL: https://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf.

34. *Sylwestrzak W., Borbinha J., Bouche T., Nowiński A., Sojka P.* EuDML – Towards the European Digital Mathematics Library // In: Sojka P.(ed.) Towards a Digital Mathematics Library. Masaryk University, 2010. P. 11–26. URL: https://eudml.org/doc/220786.

35. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. V. 2022. P. 326–333.

36. *Елизаров А.М., Липачёв Е.К.* Семантические методы и инструменты электронной математической библиотеки Lobachevskii-DML // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2017. С. 130–136. URL: https://doi.org/10.20948/abrau-2017-73. http://keldysh.ru/ abrau/ 2017/73.pdf.

37. *Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Структура и сервисы цифровой математической библиотеки Lobachevskii-DML // Ученые записки Института социально-гуманитарных знаний. 2017. № 1 (15). С. 215–220.

38. *Růžička M., Sojka P., Krejčíř V.* Towards Machine-Actionable Modules of a Digital Mathematics Library // In: Carette J., Aspinall D., Lange C., Sojka P., Windsteiger

W. (eds) Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science. 2013. V. 7961. P. 263–277. URL: https://doi.org/10.1007/ 978-3-642-39320-4_17.

39. *Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Семантический анализ больших коллекций научных документов // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2016. Казань: Изд-во Казан. ун-та, 2016. С. 21–25.

40. *Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Автоматизированная система сервисов обработки больших коллекций научных документов // Аналитика и управление данными в областях с интенсивным использованием данных: сборник статей XVIII Междун. конф. DAMDID/RCDL'2016. М.: ФИЦ ИУ РАН, 2016. С. 109–115.

41. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // Proc. of the 2nd Russia and Pacific Conf. on Computer Technology and Applications. 2017. P. 1–5. URL: https://doi.org/10.1109/RPC.2017.8168064.

42. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Klinov P., Mouromstev D. (eds.) Proceedings of the 5th International Conference on Knowledge Engineering and Semantic Web (KESW 2014). Communications in Computer and Information Science, Springer, Cham, 2014. V. 468. P. 105–119. URL: https://doi.org/10.1007/ 978-3-319-11716-4_9.

43. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O., Solovyev V., Zhiltsov N.* Mathematical Knowledge Representation: Semantic Models and Formalisms // Lobachevskii J. of Mathematics. 2014. V. 35. No 4. P. 347–353. URL: https://doi.org/10.1134/S1995080214040143.

44. *Елизаров А.М., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы и средства семантического структурирования электронных математических документов // Доклады РАН. 2014. Т. 457, № 6. С. 642–645. URL: https://doi.org/10.7868/ S0869565214240049.

45. *Елизаров А.М., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К., Невзорова О.А.* Экосистема ONTOMATH и проект Всемирной цифровой математической

библиотеки // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2016. Казань: Изд-во Казан. ун-та, 2016. С. 25–28.

46. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management // Communications in Computer and Information Science. Springer. 2017. V. 70. P. 33–46. URL: https://doi.org/10.1007/978-3-319-57135-5_3.

47. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O.* Semantic Formula Search in Digital Mathematical Libraries // Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE, 2017. P. 39–43. URL: https://doi.org/10.1109/RPC.2017.8168063.

48. *Birialtsev E., Gusenkov A., Zhibrik O., Gusenkova P., Palacheva Y.* Search in Collections of Mathematical Articles // In: Rocha Á., Adeli H., Reis L.P., Costanzo S. (eds) Trends and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing, Springer, Cham, 2018. V. 745. P. 561–567. URL: https://doi.org/10.1007/978-3-319-77703-0_55.

49. *Шакирова Л.Р., Фалилеева М.В., Кириллович А.В., Липачев Е.К., Невзорова О.А., Невзоров В.Н.* Образовательная математическая онтология OntoMathEdu: структура и отношения // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23–28 сентября 2019 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2019. С. 653–661. URL: https://doi.org/10.20948/abrau-2019-84. http://keldysh.ru/abrau/2019/ theses/84.pdf.

50. *Shakirova L., Falileeva M., Kirillovich A., Lipachev E.* Modeling and evaluation of the mathematical educational ontology //CEUR Workshop Proceedings. 2020. V. 2543. P. 305–319.

51. *Kirillovich A., Shakirova L., Falileeva M., Lipachev E.* Towards an Educational Mathematical Ontology // L. Gómez Chova, et al. (eds). Proceedings of the 13th International Technology, Education and Development Conference (INTED2019), Valencia, Spain, March 11th–13th, 2019. IATED, 2019. P. 6823–6829.

52. *Елизаров А.М., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Семантическое аннотирование в системе управления физико-математическим контентом // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (21–26 сентября 2015 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2015. С. 98–103. URL: https://kpfu.ru//staff_files/ F1890276653/Elizarov_at_all_.pdf.

53. *Елизаров А.М., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Терминологическое аннотирование и рекомендательный сервис в системе управления физико-математическим контентом // Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных». Обнинск: ИАТЭ НИЯУ МИФИ, 2015. С. 357–350. URL: https://kpfu.ru//staff_files/F684099727/damdid2015_paper_ Elizarov_new.pdf.

54. *Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Доклады РАН. 2016. Т. 467, № 4. С. 392–395. URL: https://doi.org/10.7868/S0869565216100042.

55. *Хайдаров Ш.М., Ямалутдинова Г.Ш.* Алгоритм формирования словарей рекомендующей системы подбора классификаторов научной информации // Ученые записки Института социально-гуманитарных знаний. 2017. № 1 (15). С. 552–557.

56. *Хайдаров Ш.М., Ямалутдинова Г.Ш.* Рекомендательная система классификации физико-математических документов // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 480–486. URL: https://doi.org/ 10.20948/abrau-2018-57. http://keldysh.ru/abrau/2018/theses/57.pdf.

57. *Khaydarov S.M., Yamalutdinova G.S.* Recommender system of physical and mathematical documents classification // CEUR Workshop Proceedings. 2018. V. 2260. P. 480–486.

58. *Harper C.* Metadata normalization: a case study in Primo and linked open data in libraries // Metadata Working Group Forum, Cornell, 2008. URL: https://ecommons.cornell.edu/handle/1813/10920.

59. *Koh J., Hong D., Gupta R., Whitehouse K., Wang H., Agarwal Y.* Plaster: An Integration, Benchmark, and Development Framework for Metadata Normalization Methods // BuildSys'18: Proceedings of the 5th Conference on Systems for Built Environments. 2018. P. 1–10. URL: https://doi.org/10.1145/ 3276774.3276794.

60. *Koh J., Balaji B., Sengupta D., McAuley J., Gupta R., Agarwal Y.* Scrabble: Transferrable Semi-Automated Semantic Metadata Normalization using Intermediate

Representation // BuildSys'18: Proceedings of the 5th Conference on Systems for Built Environments. 2018. P. 11–20. URL: https://doi.org/10.1145/ 3276774.3276795.

61. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html.

62. ANSI/NISO Z39.96-2012, JATS: Journal Article Tag Suite (ver. 1.0). URL: https://groups.niso.org/apps/group_public/download.php/10904/z39.96-2012.pdf.

63. Journal Archiving and Interchange Tag Set, ver. 1.1. URL: https://jats. nlm. nih.gov/publishing/1.1/.

64. Journal Archiving and Interchange Tag Set, ver. 1.2. URL: https://jats. nlm.nih.gov/archiving/1.2/.

65. JATS: Journal Article Tag Suite. National Information Standards Organization, ver. 1.2 (ANSI/NISO Z39.96-2019). 8 Febr. 2019. 652 p. URL: https:// groups.niso.org/apps/group_public/download.php/21030/ANSI-NISO-Z39.96-2019.pdf.

66. Journal Publishing Tag Library NISO JATS, version 1.3d1 (ANSI/NISO Z39.96-2019). October 2019. URL: https://jats.nlm.nih.gov/archiving/1.3d1/.

67. *Heller L., The R., Bartling S.* Dynamic Publication Formats and Collaborative Authoring // In: Bartling S. and Friesike S. (eds.). Opening Science. Springer Cham, 2014. P. 191–211. URL: https://doi.org/10.1007/978-3-319-00026-8_13.

68. *Елизаров А.М., Зайцева Н.В., Зуев Д.С., Липачёв Е.К., Хайдаров Ш.М.* Сервисы формирования метаданных цифровых документов в форматах международных наукометрических баз данных // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. P. 75–185. URL: https://doi.org/ 10.20948/abrau-2018-53. http://keldysh.ru/abrau/2018/ theses/53.pdf.

69. *Кириллович А.В.* Информационная архитектура блогов // Электронные библиотеки. 2017. Т. 20. № 2. С. 147–162. URL: https://elbib.ru/article/view/ 418/504.

70. *Елизаров А.М., Кириллович А.В., Липачёв Е.К.* Блоги в системе научных коммуникаций // Ученые записки Института социально-гуманитарных знаний. 2017. №1 (15). С. 209–214.

71. *Гафурова П.О.* Форумы в системе научных коммуникаций // Сборник конференции: Информационные технологии в образовании и науке (ИТОН-2018). 2018. С. 102–103.

72. *Борисов Н.В., Захаркина В.В., Мбого И.А., Прокудин Д.Е., Щербаков П.П.* Проблемы создания онлайн научного журнала с мультимедиа контентом // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23–28 сентября 2019 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2019. С. 153–165. URL: https://doi.org/10.20948/abrau-2019-87. http://keldysh.ru/abrau/2019/theses/87.pdf.

73. *Borisov N.V., Zakharkina N.V., Mbogo I.A., Prokudin D.E., Scherbakov P.P.* Challenges of Publishing Online Scholarly Journals with Multimedia Content // CEUR Workshop Proceedings. 2020. V. 2543. P. 93–102. URL: http://ceur-ws.org/Vol-2543/rpaper09.pdf.

74. ORCID. Connecting Research and Researchers. URL: https://orcid.org/.

75. EuDML metadata schema specification (v2.0–final). URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final.

76. *Jost M., Bouche T., Goutorbe C., Jorda J.P.* D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010. URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf.

77. D7.2: Toolset for Image and Text Processing and Metadata Editing – Initial release. URL: http://www.mathdoc.fr/publis/d7.2-v1.0.pdf.

78. D7.3: Toolset for Image and Text Processing and Metadata Enhancements — Value release. URL: http://www.mathdoc.fr/publis/d7.3.pdf.

79. D7.4: Toolset for Image and Text Processing and Metadata Enhancements — Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/ images/D7.4-v1.0.pdf.

80. *Řehůřek R., Sojka P.* Automated Classification and Categorization of Mathematical Knowledge // In: Autexier S., Campbell J., Rubio J., Sorge V., Suzuki M., Wiedijk F. (Eds) Intelligent Computer Mathematics. CICM 2008. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2008. V. 5144. P. 543–557. URL: https://doi.org/10.1007/978-3-540-85110-3_44.

81. Open Archives Initiative Protocol for Metadata Harvesting. URL: http://www.openarchives.org/OAI/openarchivesprotocol.html.

82. *Федотов А.М., Байдавлетов А.Т., Жижимов О.Л., Самбетбаева М.А., Федотова О.А.* Цифровой репозиторий в научно-образовательной информационной системе // Вестник НГУ. Серия: Информационные технологии. 2015. Т. 13.

№ 3. С. 68–86. URL: https://cyberleninka.ru/article/n/tsifrovoy-repozitoriy-v-nauchno-obrazovatelnoy-informatsionnoy-sisteme.

83. *Krejčíř V.* Building the Czech Digital Mathematics Library upon DSpace System // In: Sojka P. (Ed) DML 2008. Towards Digital Mathematics Library. Brno: Masaryk University, 2008. P. 117–126. URL: https://dml.cz/bitstream/handle/10338.dmlcz/702539/DML_001-2008-1_14.pdf.

84. *MacGregor J., Stranack K., Willinsky J.* The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication // In: Bartling S., Friesike S. (Eds) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. P. 165–175. URL: https://doi.org/10.1007/978-3-319-00026-8_3.

85. Expressing Dublin Core metadata using XML. URL: https://www.dublincore.org/specifications/dublin-core/dc-xml/.

86. What is a MARC record and why is it important? URL: https://www.loc.gov/marc/umb/um01to06.html.

87. *MARC Standarts*. URL: http://www.loc.gov/marc/.

88. *Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Программа автоматизированного формирования метаданных в формате Российского индекса научного цитирования для статей журнала «Электронные библиотеки». Свидетельство о регистрации программы для ЭВМ RU 2018612458, 16.02.2018. Заявка № 2017663206 от 19.12.2017. URL: https://www.elibrary.ru/item.asp?id =39291526.

89. *Герасимов А.Н., Елизаров А.М., Липачёв Е.К.* Формирование метаданных для международных баз цитирования в системе управления электронными научными журналами // Электронные библиотеки. 2015. Т. 18. № 1–2. С. 6–31. URL: https://elbib.ru/article/view/356/447.

90. *Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К.* Сервис-ориентированная информационная система научного журнала «Электронные библиотеки» // Электронные библиотеки. 2016. Т. 19. № 1. С. 2–39. URL: https://elbib.ru/article/view/377/468.

91. *Ley M.* DBLP – Some Lessons Learned // Proceedings of the VLDB Endowment. 2009. V. 2 (2). P. 1493–1500.

92. *Гафурова П.О.* Методы нормализации метаданных цифровых матема-

тических библиотек // В книге: Ломоносов-2019. Сборник тезисов XXVI Международной научной конференции студентов, аспирантов и молодых ученых; секция «Вычислительная математика и кибернетика». 2019. С. 162–164.

93. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Методы нормализации метаданных электронных математических коллекций // Ученые записки Института социально-гуманитарных знаний. 2019. Т. 17. № 1. С. 141–148.

94. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 136–148.

95. *Гафурова П.О., Елизаров А.М., Липачёв Е.К*., Хамматова Д.М. Методы формирования и нормализации метаданных в цифровой математической библиотеке // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23–28 сентября 2019 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2019. С. 234–244. URL: https://doi.org/10.20948/abrau-2019-28. http://keldysh.ru/abrau/2019/theses/28.pdf.

96. *Елизаров А.М., Зайцева Н.В., Липачёв Е.К., Хайдаров Ш.М.* Программа автоматизированного формирования метаданных документов цифровой математической библиотеки Lobachevskii DML. Свидетельство о регистрации программы для ЭВМ RU 2019611328, 24.01.2019. Заявка № 2019610406 от 15.01.2019. URL: https://www.elibrary.ru/item.asp?id=39309871.

97. *Батыршина Р.Р.*, *Елизаров А.М., Липачёв Е.К.* Организация коллекций цифровой математической библиотеки методами семантического анализа // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23–28 сентября 2019 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2019. С. 85–90. URL: https://doi.org/10.20948/abrau-2019-97. http://keldysh.ru/abrau/2019/theses/ 97.pdf.

98. *Elizarov A.M., Lipachev E.K.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proseedings. 2020. V. 2543. P. 354–360.

99. *Гафурова П.О., Липачёв Е.К.* Методы семантического представления математических коллекций цифровой библиотеки Lobachevskii-DML // Труды Математического центра им. Н.И. Лобачевского. 2018. Т. 56. С. 90–93.

BASIC SERVICES OF FACTORY METADATA DIGITAL MATHEMATICAL LIBRARY LOBACHEVSKII-DML

P. O. Gafurova [1], A. M. Elizarov [2], E. K. Lipachev [3]

[1-3] *Higher School of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[2, 3] *N. I. Lobachevskii Institute of Mathematics and Mechanics, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[1]pogafurova@gmail.com, [2]amelizarov@gmail.com, [3]elipachev@gmail.com

Abstract

A number of problems related to the construction of the metadata factory of the digital mathematical library Lobachevskii-DML have been solved. By metadata factory we mean a system of interconnected software tools aimed at creating, processing, storing and managing metadata of digital library objects and allowing integrating created electronic collections into aggregating digital scientific libraries. In order to select the optimal such software tools from existing ones and their modernization:

- we discussed the features of the presentation of the metadata of documents of various electronic collections related both to the formats used and to changes in the composition and completeness of the set of metadata throughout the entire publication of the corresponding scientific journal;
- we presented and characterized software tools for managing scientific content and methods for organizing automated integration of repositories of mathematical documents with other information systems;
- we discussed such an important function of the digital library metadata factory as the normalization of metadata in accordance with the formats of other aggregating libraries.

As a result of the development of the metadata factory of the digital mathematical library Lobachevskii-DML, we proposed a system of services for the automated generation of metadata for electronic mathematical collections; we have developed an xml metadata presentation language based on the Journal Archiving and

Interchange Tag Suite (NISO JATS); we have created software tools for normalizing metadata of electronic collections of scientific documents in formats developed by international organizations – aggregators of resources in mathematics and Computer Science; we have developed an algorithm for converting metadata to oai_dc format and generating the archive structure for import into DSpace digital storage; we have proposed and implemented methods for integrating electronic mathematical collections of Kazan University into domestic and foreign digital mathematical libraries.

Keywords*: digital libraries, digital mathematical library, metadata generation, metadata extraction, metadata normalization, metadata factory, NISO JATS, semantic relationships, Lobachevskii-DML.*

REFERENCES

1. *Sicilia M.-A. (Ed)* Handbook of Metadata, Semantics and Ontologies. World Scientific Publishing Co. Pte. Ltd., 2014. 579 p.

2. *Alemu G., Stevens B.* An Emergent Theory of Digital Library Metadata. Enrich then Filter. Chandos Publishing is an imprint of Elsevier. 2015, 121 p. URL: http://store.elsevier.com/An-Emergent-Theory-of-Digital-Library-Metadata/Getaneh-Alemu/isbn-9780081003855/.

3. *Gartner R.* Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer International Publishing Switzerland, 2016. 118 p. URL: https://doi.org/10.1007/978-3-319-40893-4.

4. *Kogalovsky Mikhail.* Metadata, their Properties, Functions and Classifications // CEUR Workshop Proceedings. 2012. V. 934. P. 3–14. URL: http://ceur-ws.org/Vol-934/paper3.pdf.

5. *Kogalovsky M.R.* Metadata in Computer Systems // Programming and Computer Software. 2013. V. 39 (4). P. 182–193.

6. *Xie I., Matusiak K.K.* Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016. 388 p.

7. *Jackson A.* The Digital Mathematics Library // Notices Amer. Math. Soc. 2003. V. 50. P. 918–923.

8. *Bouche T.* Introducing the mini-DML project // ECM4 Satellite Conference EMANI/DML. 2004. P. 19–29.

9. *Borwein J.M., Rocha E.M., Rodrigues J.F.* (eds.) Communicating

Mathematics in the Digital Era. Taylor & Francis, 2008. 325 p.

10. *Bouche T*. Some Thoughts on the Near-Future Digital Mathematics Library // Towards a Digital Mathematics Library. Masaryk University, 2008. P. 3–15. URL: https://eudml.org/doc/221606.

11. *Bouche T.* Digital Mathematics Libraries: The Good, the Bad, the Ugly // Math. Comput. Sci. 2010. V. 3. P. 227–241. URL: https://doi.org/10.1007/s11786-010-0029-2.

12. *Bouche T.* The Digital Mathematics Library as of 2014 // Notices Amer. Math. Soc. 2014. V. 61. No 9. P. 1085–1088.

13. *Elizarov A.M., Lipachev E.K., Zuev D.S.* Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. V. 2022. P. 317–325.

14. Committee on Electronic Information Communication of the International Mathematical Union, Best current practices: Recommendations on electronic information communication // Notices of the AMS. 2002. 49(8), pp. 922–925. URL: http://ams.org/notices/200208/commpractices.pdf.

15. *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Application of contemporary technologies in the scientific work of mathematicians // Russian Math. Surveys. 2007. V. 62 (5). P. 943–966. http://dx.doi.org/10.1070/RM2007v062n05ABEH004455.

16. *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Current state and prospects. The impact factors of Russian mathematics journals // Russian Math. Surveys. 2009. V. 64 (4). P. 775–784. http://dx.doi.org/10.1070/RM2009v064n04ABEH004638.

17. *Chebukov D.E., Izaak A.D., Misyurina O.G., Pupyrev Yu.A., Zhizhchenko A.B.* Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics // Lecture Notes in Computer Science. 2013. V. 7961. P. 344–348. https://doi.org/ 10.1007/978-3-642-39320-4_26.

18. *Bartošek M., Kovář P., Šárfy M.* DML-CZ Metadata Editor // In: Sojka P. (ed.) Towards Digital Mathematics Library. Masaryk University, 2010. P. 139–151. URL: https://dml.cz/bitstream/handle/10338.dmlcz/702537/DML_001-2008-1_17.pdf.

19. *Bartošek M., Rákosník J.* DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library // Notices of the AMS. 2013. V. 60. No 8. P. 1028–1033.

URL: http://dx.doi.org/10.1090/noti1031.

20. *Bouche T.* Toward a digital mathematics library? // In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds) Communicating Mathematics in the Digital Era. – Taylor & Francis, 2008. P. 47–73. URL: https://hal.archives-ouvertes.fr/hal-00347682.

21. *Lange C.* Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web // Semantic Web. 2013. V. 4. No. 2. P. 119–158. URL: https://content.iospress.com/articles/semantic-web/sw059.

22. *Serebryakov Vladimir*. Semantic digital libraries. What is it? // CEUR Workshop Proceedings. 2014. V. 1297. P. 1–5. URL: http://ceur-ws.org/Vol-1297/1-5_paper-1.pdf.

23. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Mathematical knowledge management: ontological models and digital technology // Analitika i upravlenie danny`mi v oblastyax s intensivny`m ispol`zovaniem danny`x: sbornik statej XVIII Mezhdun. konf. DAMDID/RCDL'2016. M.: FICz IU RAN, 2016. S. 95–101..

24. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Mathematical Knowledge Management: Ontological Models and Digital Technology // CEUR Workshop Proceedings. 2016. V. 1752. P. 44–50. URL: http://ceur-ws.org/Vol-1752/paper08.pdf

25. *Bouche T., Labbe O.* The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. Vol. 10383. Springer, Cham, 2017. P. 70–82. URL: https://doi.org/10.1007/978-3-319-62075-6_6. https://zenodo.org/record/581405/.

26. *Sadegh A., Lange C., Vidal M.E., Auer S.* Integration of Scholarly Communication Metadata using Knowledge Graphs // International Conference on Theory and Practice of Digital Libraries. 2017. P. 328–341.

27. *Ataeva O.M., Serebryakov V.A.* Information Model of LibMeta Digital Library // Lobachevskii J. of Mathematics. 2019. V. 40. No 7. P. 861–875. URL: https://doi.org/10.1134/S1995080219070035.

28. Developing a 21st Century Global Library for Mathematics Research // Washington: The National Academies Press, 2014. 142 p. doi:10.17226/18619.

29. *Olver P.J.* The World Digital Mathematics Library: Report of a Panel

Discussion // Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, 1. 2014. P. 773–785.

30. *Watt S.* How to Build a Global Digital Mathematics Library // 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). 2016. P. 37–40.

31. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. 2017. V. 10383. P. 56–69. URL: https://doi.org/10.1007/978-3-319-62075-6_5.

32. *Bouche T.* Reviving the free public scientific library in the digital age? The EuDML project // In: Kaiser K., Krantz S.G., Wegner B. (Eds) Topics and Issues in Electronic Publishing JMM/AMS Special Session, FIZ Karlsruhe. 2013. P. 57–80. URL: https://www.emis.de/proceedings/TIEP2013/05bouche.pdf.

33. *Bouche T., Rákosník J.* Report on the EuDML External Cooperation Model // In: Kaiser K., Krantz S.G., Wegner B. (Eds) Topics and Issues in Electronic Publishing, JMM, Special Session. San Diego. 2013. P. 99–108. URL: https://www.emis.de/ proceedings/TIEP2013/07bouche_rakosnik.pdf.

34. *Sylwestrzak W., Borbinha J., Bouche T., Nowiński A., Sojka P.* EuDML – Towards the European Digital Mathematics Library // In: Sojka P.(ed.) Towards a Digital Mathematics Library. Masaryk University, 2010. P. 11–26. URL: https://eudml.org/ doc/220786.

35. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. V. 2022. P. 326–333.

36. *Elizarov A.M., Lipachev E.K.* Semanticheskie metody` i instrumenty` e`lektronnoj matematicheskoj biblioteki Lobachevskii-DML // Nauchny`j servis v seti Internet: trudy` XIX Vserossijskoj nauchnoj konferencii (18–23 sentyabrya 2017 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2017. S. 130–136. https://doi.org/ 10.20948/abrau-2017-73. URL: http://keldysh.ru/abrau/2017/ 73.pdf.

37. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Structure and Services Digital Mathematical Library Lobachevskii-DML // Ucheny`e zapiski ISGZ. 2017. № 1 (15). S. 215–220..

38. *Růžička M., Sojka P., Krejčíř V.* Towards Machine-Actionable Modules of a

Digital Mathematics Library // In: Carette J., Aspinall D., Lange C., Sojka P., Windsteiger W. (eds) Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science. 2013. V. 7961. P. 263–277. URL: https://doi.org/10.1007/ 978-3-642-39320-4_17.

39. *Elizarov A.M., Lipachev E.K., Khaidarov Sh.M.* Semanticheskij analiz bol`shix kollekcij nauchny`x dokumentov // Trudy` mezhdunarodnoj konferencii po komp`yuternoj i kognitivnoj lingvistike TEL-2016. Kazan`: Izd-vo Kazan. un-ta, 2016. S. 21–25.

40. *Elizarov A.M., Lipachev E.K., Khaidarov Sh.M.* Automated Processing Service System of Large Collections of Scientific Documents // CEUR Workshop Proceedings. 2016. V. 1752. P. 58–64.

41. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // Proc. of the 2nd Russia and Pacific Conf. on Computer Technology and Applications. 2017. P. 1–5. URL: https://doi.org/10.1109/RPC.2017.8168064.

42. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Klinov P., Mouromstev D. (eds.) Proceedings of the 5th International Conference on Knowledge Engineering and Semantic Web (KESW 2014). Communications in Computer and Information Science, Springer, Cham, 2014. V. 468. P. 105–119. URL: https://doi.org/10.1007/ 978-3-319-11716-4_9.

43. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O., Solovyev V., Zhiltsov N.* Mathematical Knowledge Representation: Semantic Models and Formalisms // Lobachevskii J. of Mathematics. 2014. V. 35. No 4. P. 347–353. URL: https://doi.org/ 10.1134/S1995080214040143.

44. *Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solov'ev V.D.* Methods and Means for Semantic Structuring of Electronic Mathematical Documents // Doklady Mathematics. 2014. 90 (1). P. 521–524. URL: https://doi.org/0.1134/S1064562414050275.

45. *Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* E`kosistema ONTOMATH i proekt Vsemirnoj cifrovoj matematicheskoj biblioteki // Trudy` mezhdunarodnoj konferencii po komp`yuternoj i kognitivnoj lingvistike TEL-

2016. Kazan`: Izd-vo Kazan. un-ta, 2016. S. 25–28.

46. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management // Communications in Computer and Information Science. Springer. 2017. V. 70. P. 33–46. URL: https://doi.org/10.1007/978-3-319-57135-5_3.

47. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O.* Semantic Formula Search in Digital Mathematical Libraries // Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE, 2017. P. 39–43. URL: https://doi.org/10.1109/RPC.2017.8168063.

48. *Birialtsev E., Gusenkov A., Zhibrik O., Gusenkova P., Palacheva Y.* Search in Collections of Mathematical Articles // In: Rocha Á., Adeli H., Reis L.P., Costanzo S. (eds) Trends and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing, Springer, Cham, 2018. V. 745. P. 561–567. URL: https://doi.org/10.1007/978-3-319-77703-0_55.

49. *Shakirova L.R., Falileeva M.V., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Nevzorov V.N.* OntoMathEdu – Educational Mathematical Ontology: Structure and Relationships // Nauchny`j servis v seti Internet: trudy` XXI Vserossijskoj nauchnoj konferencii (23–28 sentyabrya 2019 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2019. S. 653–661. https://doi.org/10.20948/abrau-2019-84.

50. *Shakirova L., Falileeva M., Kirillovich A., Lipachev E.* Modeling and evaluation of the mathematical educational ontology //CEUR Workshop Proceedings. 2020. V. 2543. P. 305–319.

51. *Kirillovich A., Shakirova L., Falileeva M., Lipachev E.* Towards an Educational Mathematical Ontology // L. Gómez Chova, et al. (eds). Proceedings of the 13th International Technology, Education and Development Conference (INTED2019), Valencia, Spain, March 11th–13th, 2019. IATED, 2019. P. 6823–6829.

52. *Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* Semanticheskoe annotirovanie v sisteme upravleniya fiziko-matematicheskim kontentom // Nauchny`j servis v seti Internet: trudy` XVII Vserossijskoj nauchnoj konferencii (21–26 sentyabrya 2015 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2015. S. 98–103. URL: https://kpfu.ru//staff_files/F1890276653/ Elizarov_at_all_.pdf.

53. *Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* Mathematical Scholarly Papers Recommendation Service // Trudy` XVII Mezhdunarodnoj konferencii

DAMDID/RCDL'2015 «Analitika i upravlenie danny`mi v oblastyax s intensivny`m ispol`zovaniem danny`x». Obninsk: IATE` NIYaU MIFI, 2015. S. 357–350. URL: https://kpfu.ru//staff_files/F684099727/damdid2015_ paper_Elizarov_new.pdf.

54. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Zhizhchenko A.B., Zhil'tsov N.G.* Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics // Doklady Mathematics. 2016. 93 (2). P. 1–3. https://doi.org/10.1134/S1064562416020174.

55. *Khaydarov S.M., Yamalytdinova G.S.* Algorithm for Forming the Dictionary of the Recommender System of Selecting Classifiers of Scientific Information // Ucheny`e zapiski ISGZ. 2017. № 1 (15). S. 552–557.

56. *Khaydarov Sh.M., Yamalutdinova G.Sh.* Rekomendatel`naya sistema klassifikacii fiziko-matematicheskix dokumentov // Nauchny`j servis v seti Internet: trudy` XX Vserossijskoj nauchnoj konferencii (17–22 sentyabrya 2018 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2018. S. 480–486 https://doi.org/ 10.20948/abrau-2018-57. URL: http://keldysh.ru/abrau/2018/ theses/57.pdf.

57. *Khaydarov S.M., Yamalutdinova G.S.* Recommender system of physical and mathematical documents classification // CEUR Workshop Proceedings. 2018. V. 2260. P. 480–486.

58. *Harper C.* Metadata normalization: a case study in Primo and linked open data in libraries // Metadata Working Group Forum, Cornell, 2008. URL: https://ecommons.cornell.edu/handle/1813/10920.

59. *Koh J., Hong D., Gupta R., Whitehouse K., Wang H., Agarwal Y.* Plaster: An Integration, Benchmark, and Development Framework for Metadata Normalization Methods // BuildSys'18: Proceedings of the 5th Conference on Systems for Built Environments. 2018. P. 1–10. URL: https://doi.org/10.1145/ 3276774.3276794.

60. *Koh J., Balaji B., Sengupta D., McAuley J., Gupta R., Agarwal Y.* Scrabble: Transferrable Semi-Automated Semantic Metadata Normalization using Intermediate Representation // BuildSys'18: Proceedings of the 5th Conference on Systems for Built Environments. 2018. P. 11–20. URL: https://doi.org/10.1145/ 3276774.3276795.

61. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html.

62. ANSI/NISO Z39.96-2012, JATS: Journal Article Tag Suite (ver. 1.0). URL: https://groups.niso.org/apps/group_public/download.php/10904/z39.96-2012.pdf.

63. Journal Archiving and Interchange Tag Set, ver. 1.1. URL: https://jats. nlm. nih.gov/publishing/1.1/.

64. Journal Archiving and Interchange Tag Set, ver. 1.2. URL: https://jats. nlm.nih.gov/archiving/1.2/.

65. JATS: Journal Article Tag Suite. National Information Standards Organization, ver. 1.2 (ANSI/NISO Z39.96-2019). 8 Febr. 2019. 652 p. URL: https:// groups.niso.org/apps/group_public/download.php/21030/ANSI-NISO-Z39.96-2019.pdf.

66. Journal Publishing Tag Library NISO JATS, version 1.3d1 (ANSI/NISO Z39.96-2019). October 2019. URL: https://jats.nlm.nih.gov/archiving/1.3d1/.

67. *Heller L., The R., Bartling S.* Dynamic Publication Formats and Collaborative Authoring // In: Bartling S. and Friesike S. (eds.). Opening Science. Springer Cham, 2014. P. 191–211. URL: https://doi.org/10.1007/978-3-319-00026-8_13.

68. *Elizarov A.M., Zaitseva N.V., Zuev D.S., Lipachev E.K., Khaidarov S.M*. Services for Formation of Digital Documents Metadata in the Formats of International Science-based Databases // CEUR Workshop Proceedings. 2018. V. 2260. P. 175–185.

69. *Kirillovich A.V.* Infromation Architecture of Blogs // Russian Digital Libraries Journal. 2017. V. 20. No 2. P. 147–162. URL: https://elbib.ru/article/view/418/504

70. *Elizarov A.M., Kirillovich A.V., Lipachev E.K.* Blogs in Scientific Communications Systems // Ucheny`e zapiski ISGZ. 2017. №1 (15). S. 209–214.

71. *Gafurova P.O.* Forumy` v sisteme nauchny`x kommunikacij // Sbornik konferencii: Informacionny`e texnologii v obrazovanii i nauke (ITON-2018). 2018. S. 102–103.

72. *Borisov N.V., Zakharkina N.V., Mbogo I.A., Prokudin D.E., Scherbakov P.P.* Problems of creating an online scientific journal with multimedia content // Nauchny`j servis v seti Internet: trudy` XXI Vserossijskoj nauchnoj konferencii (23–28 sentyabrya 2019 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2019. S. 153–165. https://doi.org/10.20948/abrau-2019-87. URL: http://keldysh.ru/abrau/ 2019/theses/87.pdf

73. *Borisov N.V., Zakharkina N.V., Mbogo I.A., Prokudin D.E., Scherbakov P.P.* Challenges of Publishing Online Scholarly Journals with Multimedia Content // CEUR Workshop Proceedings. 2020. V. 2543. P. 93–102. URL: http://ceur-ws.org/Vol-

2543/rpaper09.pdf.

74. ORCID. Connecting Research and Researchers. URL: https://orcid.org/.

75. EuDML metadata schema specification (v2.0–final). URL: https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final.

76. *Jost M., Bouche T., Goutorbe C., Jorda J.P.* D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010. URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf.

77. D7.2: Toolset for Image and Text Processing and Metadata Editing – Initial release. URL: http://www.mathdoc.fr/publis/d7.2-v1.0.pdf.

78. D7.3: Toolset for Image and Text Processing and Metadata Enhancements — Value release. URL: http://www.mathdoc.fr/publis/d7.3.pdf.

79. D7.4: Toolset for Image and Text Processing and Metadata Enhancements — Final Release. URL: https://wiki.eudml.eu/mediawiki/eudml/ images/D7.4-v1.0.pdf.

80. *Řehůřek R., Sojka P.* Automated Classification and Categorization of Mathematical Knowledge // In: Autexier S., Campbell J., Rubio J., Sorge V., Suzuki M., Wiedijk F. (Eds) Intelligent Computer Mathematics. CICM 2008. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2008. V. 5144. P. 543–557. URL: https://doi.org/10.1007/978-3-540-85110-3_44.

81. Open Archives Initiative Protocol for Metadata Harvesting. URL: http://www.openarchives.org/OAI/openarchivesprotocol.html.

82. *Fedotov A.M., Baidavletov A.T., Zhizhimov O.L., Sambetbayeva M.A., Fedotova O.A.* Digital repository of scientific and educational information system // Vestnik NSU. Series: Information Technologies. 2015. V. 13. No 3. P. 68–86. URL: https://cyberleninka.ru/article/n/tsifrovoy-repozitoriy-v-nauchno-obrazovatelnoy-informatsionnoy-sisteme.

83. *Krejčíř V.* Building the Czech Digital Mathematics Library upon DSpace System // In: Sojka P. (Ed) DML 2008. Towards Digital Mathematics Library. Brno: Masaryk University, 2008. P. 117–126. URL: https://dml.cz/bitstream/handle/10338.dmlcz/702539/DML_001-2008-1_14.pdf.

84. *MacGregor J., Stranack K., Willinsky J.* The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication // In: Bartling S., Friesike S. (Eds) Opening Science. The Evolving Guide on How the Internet is Changing

Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. P. 165–175. URL: https://doi.org/10.1007/978-3-319-00026-8_3.

85. Expressing Dublin Core metadata using XML. URL: https://www. dublincore. org/specifications/dublin-core/dc-xml/.

86. What is a MARC record and why is it important? URL: https://www. loc.gov/marc/umb/um01to06.html.

87. *MARC Standarts*. URL: http://www.loc.gov/marc/.

88. *Elizarov A.M., Lipachev E.K., Khaidarov S.M*. Programma avtomatizirovannogo formirovaniya metadanny`x v formate Rossijskogo indeksa nauchnogo citirovaniya dlya statej zhurnala "E`lektronny`e biblioteki". Svidetel`stvo o registracii programmy` dlya E`VM RU 2018612458, 16.02.2018. Zayavka № 2017663206 ot 19.12.2017. https://www.elibrary.ru/ item.asp?id=39291526.

89. *Gerasimov A.N., Elizarov A.M., Lipachev E.K.* Subsystem of Formation Metadata for Science Index Databases on Management Platform Electronic Scientific Journals // Russian Digital Libraries Journal. 2015. V. 18. No 1–2. P. 6–31. URL: https://elbib.ru/article/view/356/447.

90. *Akhmetov D.Yu., Elizarov A.M., Lipachev E.K.* Service-oriented Information System of “Russian Digital Libraries Journal” // Russian Digital Libraries Journal. 2016. V. 19. No 1. P. 2–39. URL: https://elbib.ru/article/view/377/468.

91. *Ley M.* DBLP – Some Lessons Learned // Proceedings of the VLDB Endowment. 2009. V. 2 (2). P. 1493–1500.

92. *Gafurova P.O.* Metody` normalizacii metadanny`x cifrovy`x matematicheskix bibliotek // V knige: Lomonosov-2019 Sbornik tezisov XXVI Mezhdunarodnoj nauchnoj konferencii studentov, aspirantov i molody`x ucheny`x; sekciya «Vy`chislitel`naya matematika i kibernetika». 2019. S. 162–164.

93. *Gafurova P.O., Elizarov A.M.,. Lipachev E.K.* Methods of Normalization of Metadata Digital Mathematical Collections // Ucheny`e zapiski ISGZ. 2019. T. 17. № 1. S. 141–148.

94. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 136–148.

95. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Methods of Formation and Normalization of Metadata in the Digital Mathematical Library //

Nauchnyj servis v seti Internet: trudy XXI Vserossijskoj nauchnoj konferencii. 2019. S. 234–244. https://doi.org/10.20948/abrau-2019-28. URL: http://keldysh.ru/abrau/2019/theses/28.pdf.

96. *Elizarov A.M., Zaitseva N.V., Lipachev E.K., Khaidarov S.M*. Programma avtomatizirovannogo formirovaniya metadanny`x dokumentov cifrovoj matematicheskoj biblioteki Lobachevskii DML. Svidetel`stvo o registracii programmy` dlya E`VM RU 2019611328, 24.01.2019. Zayavka № 2019610406 ot 15.01.2019. https://www.elibrary.ru/item.asp?id=39309871.

97. *Batyrshina R.R Elizarov A.M., Lipachev E.K* Organization of Digital Mathematics Library Collections by Semantic Analysis Methods // Nauchny`j servis v seti Internet: trudy` XXI Vserossijskoj nauchnoj konferencii (23–28 sentyabrya 2019 g., g. Novorossijsk). M.: IPM im. M.V. Keldy`sha, 2019. S. 85–90. https://doi.org/10.20948/abrau-2019-97. URL: http://keldysh.ru/abrau/2019/ theses/ 97.pdf.

98. *Elizarov A.M., Lipachev E.K.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proseedings. 2020. V. 2543. P. 354–360.

99. *Gafurova P.O., Lipachev E.K.* Methods for the Semantic Representation of Mathematical Collections of the Lobachevskii Digital Mathematics Library // Trudy` matematicheskogo centra im. N.I. Lobachevskogo. 2018. T. 56. S. 90–93.

СВЕДЕНИЯ ОБ АВТОРАХ

ГАФУРОВА Полина Олеговна – магистр математики, аспирант Высшей школы информационных технологий и информационных систем Казанского (Приволжского) федерального университета.

Polina GAFUROVA – Magister of Mathematics, Kazan (Volga Region) Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: pogafurova@gmail.com; ORCID: 0000-0002-1544-155X

ЕЛИЗАРОВ Александр Михайлович – доктор физико-математических наук, профессор, заслуженный деятель науки Республики Татарстан, профессор кафедры программной инженерии Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Alexander ELIZAROV – Doctor of Physics and Mathematics, Professor, Honored Worker of Science of the Republic of Tatarstan, Kazan Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: amelizarov@gmail.com; ORCID: 0000-0003-2546-6897

ЛИПАЧЁВ Евгений Константинович – кандидат физико-математических наук, доцент кафедры Интеллектуальных технологий поиска Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Evgeny Konstantinovich LIPACHEV – Candidate of Physics and Mathematics, Associate Professor, Kazan Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: elipachev@gmail.com; ORCID: 0000-0001-7789-2332

Материал поступил в редакцию 21 ноября 2019 года

УДК 004

НАУЧНЫЕ ПУБЛИКАЦИИ В РОССИИ. ЧТО НОВОГО

М. М. Горбунов-Посадов

Институт прикладной математики им. М.В. Келдыша Российской академии наук, г. Москва
gorbunov@keldysh.ru

Аннотация

Представлены события, происходившие в последнее время в мире российских научных публикаций. Наблюдается медленное сползание в сторону платного доступа части академических журналов, размещенных в открытом доступе в 2018 году. В Европейском союзе объявлен план массового перехода научных журналов к открытому доступу. Внедряются новые модели существования научной публикации. Отчетность по публикациям, затребованная Минобрнауки в 2019 году, не учитывает масштабы читательской аудитории статьи. Ни Минобрнауки, ни ВАК никак не поощряют размещение публикации в открытом доступе. В РИНЦ началась борьба с широко распространенной жульнической торговлей цитированиями статьи, однако ВАК эта деятельность не заинтересовала. Получил распространение внутренне противоречивый термин «автоплагиат», которым широко клеймят авторов и издания за множественные публикации.

***Ключевые слова**: открытый доступ, «план S», административная оценка статьи, сериальные издания, онлайновый читатель, индекс Хирша, РИНЦ, Диссернет, автоплагиат*

ВВЕДЕНИЕ

В 2018–2019 гг. в мире российских научных публикаций произошло немало событий. К сожалению, некоторые из них совсем не порадовали.

ОТКРЫТЫЙ ДОСТУП

Трудно складывается переход к открытому доступу. В начале 2018 г. обнадеживающие вести пришли из Российской академии наук. После жесткого замечания Счетной палаты большинство журналов РАН в онлайне стали доступны бесплатно и

даже без регистрации читателя. Однако вскоре это завоевание было во многом утрачено. Сначала при сохранении доступа для чтения на год была заблокирована возможность полного или частичного копирования текста. А затем понемногу журналы стали возвращаться к платному доступу. Так, например, сейчас полные тексты флагманского академического журнала «Вестник РАН» бесплатно доступны лишь на официальном сайте журнала, а на сайте eLibrary (РИНЦ) за скачивание полного текста одной статьи вновь, как и в 2017 году, требуют 220 руб. Нечто подобное происходит и с остальными академическими журналами.

В Европейском союзе энергично обсуждается «план S», согласно которому в ближайшие годы в открытом онлайне должна оказаться большая часть европейских научных публикаций. Впрочем, движение Европы к открытому доступу пока оставляет желать лучшего. Хотя концепция открытого доступа была впервые декларирована именно в Европе, за пятнадцать лет с 2004 по 2018 гг. охват публикаций открытым доступом увеличился там лишь с 15% до 20%. В то же время в России в открытый доступ уже сейчас попадает около 50% научных статей: одна Киберленинка охватывает таким доступом около четверти российских научных журналов.

Решительный разворот в сторону открытого доступа обещает наметившийся переход к новой модели существования научной публикации. Все чаще первая версия публикации, прошедшая рецензирование лишь в организации ее автора, сразу же оказывается в открытом доступе на сайте организации. И лишь затем статья направляется в научный журнал для получения дополнительной оценки ведущих специалистов. При получении положительной оценки, после возможной корректировки текста в соответствии с замечаниями рецензентов, статья получает своеобразный знак качества: «Принято к публикации авторитетным журналом». Эта информация размещается на заметном месте в файле статьи на сайте организации. Журнал вправе опубликовать такую статью у себя, но может и ограничиться размещением на своем сайте ссылки на исходное размещение статьи.

АДМИНИСТРАТИВНЫЕ РЕШЕНИЯ

2019 год начался с объявления очередных требований к количеству и составу статей, выпускаемых организацией, подчиненной Минобрнауки. Требования ока-

зались достаточно жесткими и не всегда согласующимися с представлениями авторов о ценности публикации.

В качестве примера применения этих требований рассмотрим две недавно вышедшие статьи автора этих строк. Первая — в «Троицком варианте» [1]: более 150 тыс. посещений, интереснейшее обсуждение, множество писем не только от коллег, но даже и от одноклассников и родственников. Другая, сходная по тематике статья в «Программировании» [2] проходит практически незамеченной: годичное эмбарго на открытый доступ, за полгода 2 (два!) скачивания по цене 220 руб. в РИНЦ. Казалось бы, оценка очевидна: впечатляющий успех в «Троицком варианте» и полный провал в «Программировании» — читателей «Программирования» оказалось в сто тысяч раз меньше.

Однако с точки зрения объявленных требований Минобрнауки полезность этих публикаций оценивается ровно наоборот. За статью в «Троицком варианте» ни автор, ни его организация зачетных баллов не получают: это издание не входит в Перечень ВАК, не индексируется ни в Web of Science (WoS), ни в РИНЦ. Даже в отчет о научно-популярных публикациях статья не может быть включена: многотысячный тираж печатной версии и сотни тысяч онлайновых читателей для министерства ничего не значат — ведь «Троицкий вариант» не имеет жестко затребованный министерством ISSN. Напротив, статья в «Программировании» оценена высшим баллом: «Программирование» индексируется в Web of Science, а важнее этого, как недавно стало известно, в науке нет ничего.

Вновь, на этот раз со стороны Минобрнауки, гонениям подверглись сериальные издания, не получающие в 2019 году зачетных баллов. А ведь именно за сериальными изданиями будущее. Периодичность изданий, требуемая в 2019 году министерством, объяснялась в свое время исключительно ограничениями равномерной загрузки полиграфических мощностей и почты. Сейчас, когда тиражи печатных версий научных журналов упали до сотни экземпляров, эти ограничения уже не работают. На первый план выходят очевидные слабости периодичности: портфель издания переполнен — вынуждены отклонять сильные статьи, портфель оскудел — публикуем слабые. Сериальные издания, напротив, публикуют все принятые редколлегией достойные статьи и только достойные статьи. Несколько лет назад ВАК наконец признал сериальные издания достой-

ными вхождения в Перечень, однако Минобрнауки пока не спешит прислушаться к приведенным выше аргументам.

Огорчает полное безразличие российского чиновника по отношению к самому массовому читателю — онлайновому. ВАК, с одной стороны, раз за разом множит жесткие требования к составу и срокам размещения в онлайне бесчисленных, во многом бюрократических атрибутов защиты, но, с другой стороны, признает полноценной зачетной публикацией диссертанта статью, полный текст которой к моменту защиты не попал в открытый доступ, т. е. был недоступен для массового читателя. В формате отчета 2019 года организации для Минобрнауки ни открытый доступ, ни интернет вообще не упоминаются. Напротив, авторов статей систематически призывают к публикациям в журналах, индексируемых в Web of Science, а большинство этих журналов в открытый доступ не попадает.

ИЗДЕРЖКИ ПУБЛИКАЦИОННОЙ ОТЧЕТНОСТИ

Несколько лет на видных местах популярных страниц интернета красовались баннеры с предложениями авторам статей о наращивании их показателей цитируемости, индекса Хирша. За появление в мало-мальски известном журнале библиографической ссылки на свою статью автору предлагалось заплатить жуликам 500 руб. Трудно определить, насколько популярной оказалась эта услуга, хотя обилие баннеров говорит о том, что такая коммерция процветала. В конце концов, масштабы липовых ссылок начали раздражать РИНЦ (eLibrary), и там решили произвести контрольную закупку липовых ссылок. Оплата ссылки производилась после предъявления заказчику номера журнала, где ссылка публиковалась, благодаря чему были выявлены более 50 журналов, не заботящихся о своей репутации, не контролирующих корректность библиографических списков публикуемых статей. Все эти журналы были решительно исключены из РИНЦ [3], что, разумеется, можно только приветствовать.

Однако интересна реакция ВАК на эти события. Часть исключенных журналов оказалась входящими в Перечень ВАК, и тут обнаружилось определенное противоречие. В требованиях к журналу, претендующему на включение в Перечень, записано, что журнал должен индексироваться в РИНЦ. Но теперь, после проведенной РИНЦ чистки, для изгнанных журналов это требование перестало удовлетворяться. ВАК находит неожиданных выход из возникшего противоре-

чия: вместо того, чтобы исключить (хотя бы на время) скомпрометировавшие себя журналы из своего Перечня, вопреки здравому смыслу исключает из своих правил требование об индексировании журнала в РИНЦ, заявляя при этом, что РИНЦ скомпрометировал себя проведением политики произвольного исключения изданий, включенных в Перечень ВАК.

Минобрнауки издает грозные распоряжения, и сотрудники бросаются любой ценой спасать административный рейтинг своей организации. Например, один из таких энтузиастов опубликовал за последнее время 626 своих работ в мусорных изданиях (160 – в журналах, 466 – в сборниках) [5]. Это, конечно, рекорд, однако подобных ему российских авторов с десятками бессмысленных мусорных публикаций насчитывается уже более десяти тысяч [5].

Не надо думать, что положение спасает сделанный в директивных документах акцент на издания из Web of Science (WoS). Профессор РГСУ И.Н. Медведев опубликовал в 2018 году 170 статей в индексируемом в WoS журнале "Research Journal of Pharmaceutical, Biological and Chemical Sciences", выпуская до сорока своих статей в одном номере журнала [6]. Правда, в 2019 году упомянутый журнал был исключен из WoS, однако И.Н. Медведев находит еще один сочувствующий ему WoS-журнал, так что и в 2019 году счет опубликованных им WoS-статей вновь пошел на десятки.

Повышенное внимание к публикационной активности ученого привело и к другим издержкам. Если еще недавно автора, издавшего свои две-три совпадающие или близкие по тексту публикации в разных изданиях, вполне спокойно воспринимали просто как желающего расширить свою читательскую аудиторию, то теперь повсюду расплодились этические кодексы, где такие действия клеймят безумным новомодным термином «автоплагиат», а и авторы, и журналы, допускающие такие множественные публикации, подвергаются всяческим гонениям.

ЗАКЛЮЧЕНИЕ

На поле научных публикаций обозначилось несколько мощных игроков — Минобрнауки, ВАК, РАН, РИНЦ, Диссернет — каждый из которых претендует на роль законодателя этических норм. Однако оказалось, что представления об этике публикаций у каждого из них сейчас весьма несхожи с соседними. Удастся

ли этим весьма несхожим структурам договориться, прийти к единому пониманию содержания и этики научной публикации — покажет время.

СПИСОК ЛИТЕРАТУРЫ

1. *Горбунов-Посадов М.М.* Цифровая наука в РАН // Троицкий вариант — наука. 2018. № 5. С. 14. URL: https://trv-science.ru/2018/03/13/cifrovaya-nauka-v-ran/

2. *Горбунов-Посадов М.М., Полилова Т.А.* Инструменты поддержки онлайновой научной публикации // Программирование. 2019. № 3. С. 38–42. URL: https://doi.org/10.1134/S013234741903004X

3. Об исключении журналов из РИНЦ. URL: https://elibrary.ru/ journals_excluded.asp

4. Новые подходы к нормативно-правовому регулированию системы аттестации научных кадров высшей квалификации в Российской Федерации. URL: https://vak.minobrnauki.gov.ru/uploader/loader?type=35&name=3397129001&f=3544

5. *Ростовцев А., Абалкина А.* Чемпионы мусорной науки // Троицкий вариант — наука. 2018. № 255. С. 1. URL: https://trv-science.ru/2018/06/05/chempiony-musornoj-nauki/

6. Research Journal of Pharmaceutical, Biological and Chemical Sciences. 2018. V. 9. Issue 5 (September – October). URL: http://rjpbcs.com/ 2018_9.5.html

RUSSIAN SCIENTIFIC PUBLICATION — 2019

M.M. Gorbunov-Posadov

Keldysh Institute of Applied Mathematics, Moscow

gorbunov@keldysh.ru

Abstract

The article presents the events that took place last year in the world of Russian scientific publications. There is a slow slide towards paid access of some academic journals turned in open access in 2018. The European Union has announced plan "S" for the mass transition of scientific journals to open access. New models of the scientific publication are introducing. Reporting on publications requested by the Ministry of education and science in 2019 does not take into account the size of the readership of the article. Neither the Ministry of education and science, nor the Higher Attestation Commission (HAC) does not encourage publication in the public domain. In Russian Science Citation Index began the fight against widespread fraudulent trade in references to the article, but the HAC is not interested in this activity. A proliferation of contradictory the term "self-plagiarism" has spread. This label is widely stigmatized authors and journals for repeated publications.

Keywords*: open access, plan "S" administrative assessment of article, serial publications, online reader, h-index, Dissernet, self-plagiarism*

REFERENCES

1. *Gorbunov-Posadov M.M.* Tsifrovaia nauka v RAN // Troitskii variant — nauka. 2018. № 5. S. 14. URL: https://trv-science.ru/2018/03/13/cifrovaya-nauka-v-ran/

2. *Gorbunov-Posadov M.M., Polilova T.A.* Instrumenty podderzhki onlainovoi nauchnoi publikatsii // Programmirovanie. 2019. № 3. S. 38–42. URL: https://doi.org/10.1134/S013234741903004X

3. Ob iskliuchenii zhurnalov iz RINTs. URL: https://elibrary.ru/journals_excluded.asp

4. Novye podkhody k normativno-pravovomu regulirovaniiu sistemy attestatsii nauchnykh kadrov vysshei kvalifikatsii v Rossiiskoi Federatsii. URL:

https://vak.minobrnauki.gov.ru/uploader/loader?type=35&name=3397129001&f=3544

5. *Rostovtsev A., Abalkina A.* Chempiony musornoi nauki // Troitskii variant — nauka. 2018. № 255. С. 1. URL: https://trv-science.ru/2018/06/05/chempiony-musornoj-nauki/

7. Research Journal of Pharmaceutical, Biological and Chemical Sciences. 2018. V. 9. Issue 5 (September – October). URL: http://rjpbcs.com/ 2018_9.5.html

СВЕДЕНИЯ ОБ АВТОРЕ

ГОРБУНОВ-ПОСАДОВ Михаил Михайлович – главный научный сотрудник ИПМ им. М.В.Келдыша РАН, д. ф.-м. н.

Mikhail Mikhailovich GORBUNOV-POSADOV – Keldysh Institute of Applied Mathematics, chief researcher

gorbunov@keldysh.ru

Материал поступил в редакцию 16 ноября 2019 года

УДК 004.658.6+004.82+004.8

ПОСТРОЕНИЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ НА ОСНОВЕ ЛОГИЧЕСКОЙ МОДЕЛИ ДАННЫХ

А. М. Гусенков[1], Н. Р. Бухараев[2], Е. В. Биряльцев[3]

[1,2]*Казанский (Приволжский) федеральный университет, Казань;*
[3]*Центр цифровых технологий Института прикладных исследований Академии наук Республики Татарстан, Казань;*

[1]gusenkov.a.m@gmail.com, [2]boukharay@gmail.com, [3]igenbir@yandex.ru

Аннотация

Представлена технология автоматизированного построения онтологии предметной области на основе информации, извлекаемой из комментариев реляционных баз данных ПАО «Татнефть». Технология основана на построении конвертора (компилятора), транслирующего логическую модель данных Epicentre Petrotechnical Open Software Corporation (POSC), представленную в виде ER-диаграмм и набора описаний на объектно-ориентированном языке EXPRESS, в язык описания онтологий OWL, рекомендованный консорциумом W3C. Описаны основные синтаксические и семантические аспекты преобразования.

***Ключевые слова**: онтология предметной области, реляционные базы данных, POSC, OWL.*

ВВЕДЕНИЕ

Создание онтологии предметной области является, как правило, крайне трудоемким процессом, требующим участия высококвалифицированных специалистов как в конкретной предметной области, так и в области компьютерной лингвистики. Одним из широко применяемых здесь методов является концептуализация понятийного аппарата [1]. При этом мы фактически строим объектную модель некой подобласти реального мира, определяя объекты, их атрибуты и взаимосвязи. Аналогичная техника выделения базовых сущностей и связей используется и для построения логических моделей при проектировании реляционных баз данных [2]. Методологическая близость приемов, используемых при разработке онтологий и логических моделей баз данных, дает основание пред-

положить возможность использования существующих логических моделей баз данных в качестве формализованного прототипа онтологии предметной области.

В статье, на примере разработки системы интеллектуального поиска для крупной нефтедобывающей компании, предложена методика автоматизированного построения онтологии предметной области на основе ее реляционной модели.

В качестве прототипа онтологии предметной области естественно выбрать относящиеся к данной области логические модели, имеющие статус отраслевого стандарта. Одной из них является модель данных Epicentre версии 3.0 нефтетехнической корпорации Petrotechnical Open Software Corporation (POSC [3]). Она представлена в виде ER-диаграмм [2], а также набора текстовых файлов на объектно-ориентированном языке EXPRESS (ISO 10303, part 11). Такое представление ориентировано на автоматическую генерацию структур баз данных по её логической модели, а также визуальное восприятие IT-специалистами.

Для описания онтологии был выбран язык OWL (Web Ontology Language) [4, 5], разработанный рабочей группой Semantic Web Activity и рекомендованный международным консорциумом W3C [6]. Реализация онтологии выполнена на диалекте языка OWL DL, соответствующем правилам дескриптивной логики, с перспективой дальнейшей разработки системы логического вывода утверждений о данной предметной области. В статье описана схема конвертации логической модели Epicentre в язык описания онтологий OWL, учитывающая, помимо общеотраслевых стандартов, специфику нефтегазодобывающей компании ПАО «Татнефть».

МОДЕЛЬ EPICENTRE

В модели данных Epicentre определено более 1000 реально существующих технических и бизнес-объектов, связанных с разведкой и добычей нефти. В терминологии POSC-моделирования данных эти объекты названы сущностями (entities). В модели определены характеристики объектов, называемые атрибутами сущностей (entity attributes). Наиболее важными из них являются атрибуты, определяющие взаимосвязи между сущностями.

Один из важных архитектурных принципов Epicentre основан на различии между объектами, свойствами или характеристиками объектов (properties), с одной стороны, и видами деятельности (activities), с другой. Это разделение поддерживается требованиями практики: возможностью свойств объекта иметь многократные версии или описания, а также тем, чтобы каждое свойство было однозначно связано со своим собственным определением (описанием) или историей обработки.

Логическая модель данных, предлагаемая Epicentre, представляет собой набор определений сущностей и связей между ними в виде выражений на языке Express и диаграмм «сущность–связь». Каждая сущность, представленная в модели, определяется такими параметрами, как набор атрибутов, локальные правила и цепочки супертипов. Атрибут – это перечень объектов, описывающих данную сущность. В свою очередь, атрибут имеет несколько параметров, таких, как имя атрибута, список опций (ключевой, обязательный, внешний и др.), и типы связей. С сущностями также связаны правила валидации сущности – расширенные ограничения целостности данных, определяющие возможные значения атрибутов и связей.

Описание сущностей модели делятся на уровни абстракции представления объектов. Описание предметной области начинается с очень высокого уровня абстракции, на котором нефтегазовая специфика практически отсутствует. Высокоуровневая модель сущностей строится с понятий E_and_P_data, под которым понимается любой информационный объект, и Data_collection, под которым понимается произвольный набор объектов. Модель Epicentre следует принципу моделирования, при котором конкретному экземпляру сущности (entity) – объекту – обеспечивается его существование. Поэтому каждый экземпляр представляет отдельный объект и не может быть версией существующего объекта. Следовательно, новые версии характеристик объекта должны быть указаны в другом месте. В Epicentre эти характеристики связаны с наличием атрибутов или свойств (properties) сущностей.

В модели повсюду применяется архитектурный принцип разделения объектов, свойств и деятельности. Это позволяет непосредственно использовать Epicentre для многих различных по характеру моделей данных. Таким образом, понятия, которые могут рассматриваться в некоторых моделях данных как про-

стые атрибуты сущностей, в Epicentre проявляются как сущности в их полном смысле.

Для обеспечения совместимости со спецификациями POSC множество сущностей в Epicentre расширяется за счет сущностей, имеющих стандартный набор экземпляров. Подобные сущности называются справочными (reference_entities).

В модели существует три типа справочных сущностей:

- POSC Fixed – имеет фиксированное количество экземпляров, определенных POSC;
- POSC Open – имеет фиксированные экземпляры, но также возможно создание дополнительных экземпляров, не определенных POSC;
- Local – нет фиксированных определенных POSC-экземпляров, возможно лишь определение собственных экземпляров.

Все справочные сущности имеют дополнительные характеристики, позволяющие задать источник и библиографию, связанную с источником содержащейся в данном экземпляре информации. Для обозначения справочных сущностей и их типов в схеме Epicentre используется отличительная приставка Ref_ к имени сущности.

Объектно-ориентированная концепция наследования класса – важная часть архитектуры Epicentre. Так как модель данных достаточно велика, концепция наследования классов обеспечивает эффективный способ организации всех сущностей в логически связанную структуру.

Другая фундаментальная часть архитектуры Epicentre основана на понимании того, что многие сущности характеризуются своим пространственным представлением. Каждый из подобных объектов деятельности может быть связан со своим местоположением через отношения с одним или несколькими аналогичными пространственными объектами.

Модель данных Epicentre задается на языке EXPRESS и использует базовые понятия этого языка:

- **Сущность** (entity);
- **Супертип** (supertype) и **подтип** (subtype);
- **Атрибут** (attribute), явный (explicit) и инверсный (inverse);

- **Определяемый** тип данных (defined data type);
- **Простой** тип (simple type);
- **Агрегатный** тип (aggregate type);
- **Ограничение согласованности**, «где» – правило (where rule);
- **Ограничение уникальности** (uniqueness rule);
- **Схема** (schema).

Описанием сущности является определение класса модели Epicentre, на основе которого создаются экземпляры класса или объекты.

Сущность может быть подтипом сущности-супертипа и наследовать от нее атрибуты, правила и ограничения уникальности. Спецификация супертипа – это способ задания свойств данного типа, общих для всех подтипов.

Супертип может быть абстрактным (ABSTRACT), что означает, что все экземпляры должны быть определены. Если сущность не является абстрактной, то экземпляры могут быть неопределенными.

Атрибут (свойство) – это конкретная характеристика сущности. Он может быть прямым (явным), инверсным или выведенным из описания некоторой сущности. Атрибут имеет имя и тип представления.

Явный атрибут – такой атрибут, который не ссылается на описание какого-либо другого атрибута в модели.

Инверсный атрибут служит для выражения обратного направления отношения принадлежности к сущности, возникающего при задании явного атрибута.

Определение схемы представляет собой контейнер, в котором содержатся определения всех сущностей, типов и ограничений, видимых в определенной схеме на языке EXPRESS (рис. 1).

```
ENTITY schema_definition;
        name : STRING;
        types : SET OF named_type;
        global_rules : SET OF global_rule;
        UNIQUE
        Url : name;
END_ENTITY;
```

Рис. 1. Пример определения схемы на языке EXPRESS

Здесь:

name: имя схемы;

types: набор сущностей и определенных типов, объявляемых в схеме;

global rules: набор глобальных ограничений, объявленных в схеме.

Формальные утверждения:

Url: имя схемы должно быть уникально.

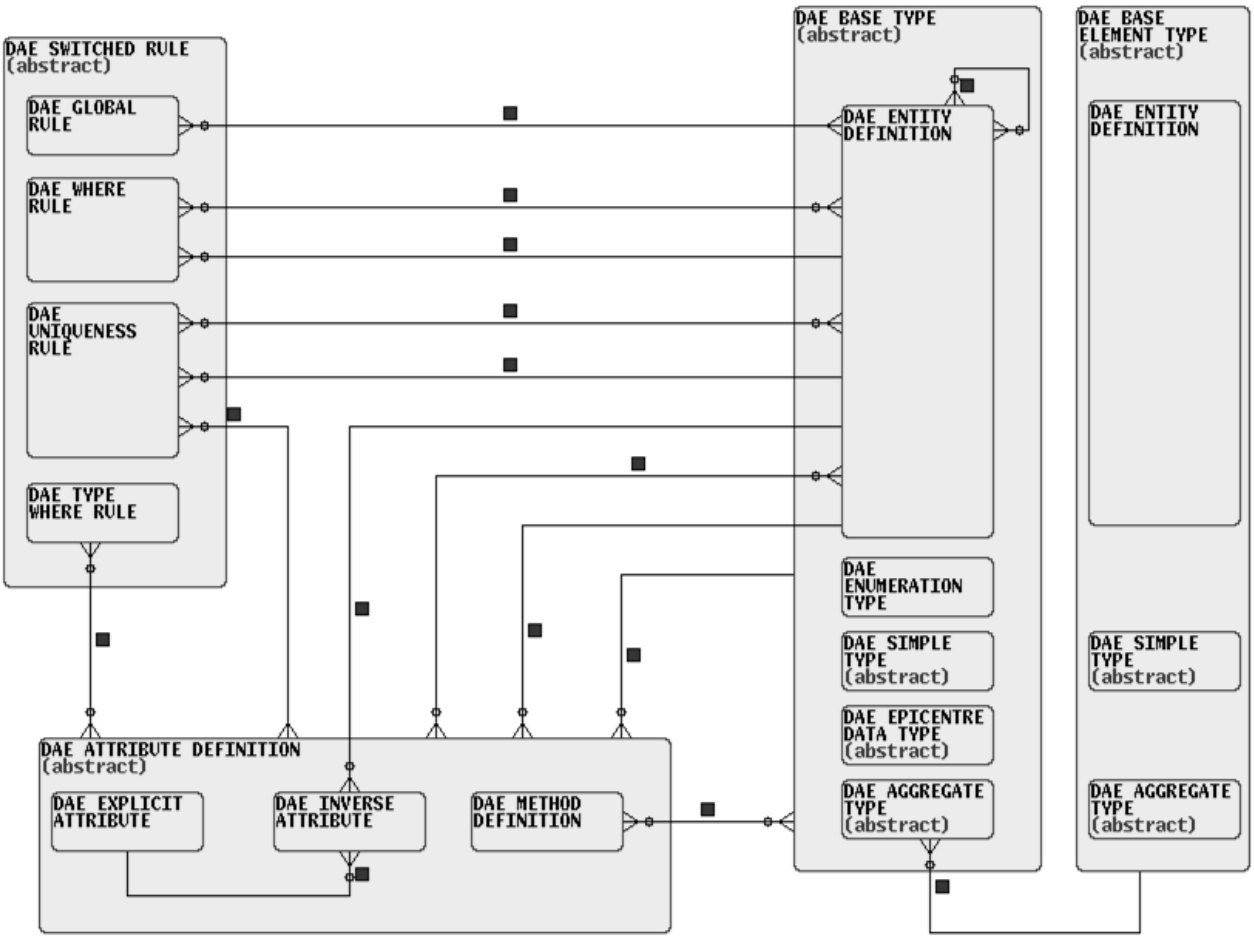


Рис. 2. Метамодель Epicentre

Ограничение уникальности указывает на комбинацию атрибутов, значения которых в совокупности однозначно идентифицируют конкретный экземпляр определяемой сущности.

Ограничение согласованности определяет ограничение, накладываемое на все экземпляры данной сущности.

Схема метамодели представлена на ER-диаграмме (рис. 2).

СТРУКТУРА OWL

Язык веб-онтологий OWL (Web Ontology Language) предоставляет возможности:

- формального определения классов и свойств этих классов;
- определения индивидов (объектов-экземпляров, представителей классов) и их свойств;
- уточнения определений классов и индивидов.

OWL максимально совместим с языками RDF [7] и RDF Schema [8]. Форматы XML [9] и RDF – часть стандарта OWL.

Основными элементами онтологии OWL являются классы, свойства, представители классов (индивиды или экземпляры) и отношения между этими представителями.

Класс – это именованная совокупность свойств, описывающих некий набор индивидов. Таким образом, классы должны соответствовать естественно определяемым наборам объектов рассматриваемой предметной области, а индивиды должны соответствовать реальным объектам, которые могут быть сгруппированы в эти классы.

Наиболее фундаментальные понятия предметной области должны соответствовать классам, которые находятся в корне различных таксономических деревьев. Каждый индивид в OWL является членом класса owl:Thing. Таким образом, каждый класс, определенный пользователем, автоматически является подклассом owl:Thing. Корневые классы, специфичные для данной предметной области, определяются простым объявлением именованного класса.

Фундаментальным таксономическим конструктором для классов является отношение «быть подклассом» rdfs:subClassOf, описывающее связь частного класса с более общим. Если X – подкласс Y, то каждый представитель X – также представитель Y. Отношение rdfs:subClassOf является транзитивным: если X – подкласс Y и Y – подкласс Z, то X – подкласс Z.

Пример: создадим определение для класса ACTIVITY (деятельность) из модели данных Epicentre. ACTIVITY наследует свойства от корневой сущности иерархии супертипов E_AND_P_DATA, представляющих данные геологической разведки и разработки месторождений.

```
<owl:Class rdf:ID=" ACTIVITY">
        <rdfs:subClassOf rdf:resource=" #e_and_p_data "/>
        ...
</owl:Class>
```

Для определения индивида достаточно объявить его членом какого-то класса.

Свойства позволяют утверждать общие факты о членах классов и особые факты об индивидах. Свойство – это бинарное отношение. Различают два типа свойств:

- свойства-значения – отношения между представителями классов и стандартными типами данных, определяемых XML- или RDF-схемой;
- свойства-объекты, определение которых ссылается на описание другого класса.

При определении свойства существует множество способов ограничить это отношение. Можно определить домен и диапазон. Свойство может быть определено как специализация (подсвойство) существующего свойства; возможны и более сложные ограничения. Свойства, как и классы, могут быть организованы в иерархию.

Существует возможность определить характеристики свойства как отношения, что обеспечивает мощный механизм их описания. Приведем некоторые характеристики свойств, важных для описания методики конвертации логической модели Epicentre в язык описания онтологий OWL.

TransitiveProperty

Если свойство P определено как транзитивное, то для любых x, y и z: P(x,y) и P(y,z) предполагают P(x,z).

SymmetricProperty

Если свойство P помечено как симметрическое, то для любых x и y: P(x,y), если P(y,x).

FunctionalProperty

Если свойство P помечено как функциональное, то для любых x, y и z: P(x,y) и P(x,z) предполагают y=z.

inverseOf

Если свойство P1 помечено как owl:inverseOf P2, то для всех x и y: P1(x,y), если P2(y,x).

Заметим, что синтаксис для owl:inverseOf берет в качестве аргумента название свойства. «A, если B» означает здесь (A предполагает B) и (B предполагает A).

InverseFunctionalProperty

Если свойство P помечено как обратно функциональное, то для всех x, y и z: P(y,x) и P(z,x) предполагает y=z.

В дополнение к обозначению характеристик свойств можно разными способами еще больше ограничить диапазон свойства в определенных контекстах. Это делается с помощью следующих ограничений свойств.

allValuesFrom

Данные ограничения являются локальными по отношению к содержащему их классу. Ограничение owl:allValuesFrom требует, чтобы для каждого представителя данного класса, который имеет данное свойство, все значения этого свойства являлись представителями класса, заданного в пункте owl:allValuesFrom.

someValuesFrom

Ограничение owl:someValuesFrom требует, чтобы для каждого представителя данного класса, который имеет данное свойство, хотя бы одно значение этого свойства являлось представителем класса, заданного в пункте owl:someValuesFrom.

Кардинальность

Параметр owl:cardinality позволяет указать количество элементов в связи. Например, свойство UNIQUE класса Name_entity определяет единственную связь (рис. 3).

В частности, в OWL DL owl:maxCardinality может использоваться для указания верхнего предела, а owl:minCardinality – для нижнего предела. В комбина-

ции друг с другом они могут использоваться для ограничения кардинальности свойства в пределах числового интервала.

```
<owl:Class rdf:ID=" Name_entity">
        <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty rdf:resource="#UNIQUE"/>
                        <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1
                        </owl:cardinality>
                </owl:Restriction>
        </rdfs:subClassOf>
</owl:Class>
```

Рис. 3. Пример определения кардинальности на языке OWL

КОНВЕРТАЦИЯ МОДЕЛИ EPICENTRE В ЯЗЫК OWL DL

При построении OWL-онтологии были использованы следующие основные подходы:

- любой сущности Epicentre соответствует простой именованный класс OWL-онтологии с сохранением в именах классов приставок, позволяющих идентифицировать сущности-свойства и сущности-справочники; все эти классы располагаются в корне таксономического дерева онтологии;
- степени связи сущностей (один-к-одному, один-ко-многим или многие-ко-многим) в OWL соответствует определение простых свойств-атрибутов, если связанная сущность не является типом данных, и свойств-значений – в противном случае; указание степени связи между классами реализовано с помощью понятия кардинальности в OWL.

В OWL отсутствуют структурные элементы, которые в полном объеме описывают определение уникальности объектов модели Epicentre. Поэтому в определение каждого класса на языке OWL добавлено новое предопределенное свойство, в котором перечислены все атрибуты, образующие уникальный ключ. Аналогичным же образом решена проблема сохранения условий ограничений.

Для каждой категории данных Epicentre построены отдельные классы OWL-онтологии. Построена формальная LR(1)-грамматика [10] модели Epicentre, на основе которой реализовано семантическое преобразование описанной на

языке EXPRESS модели Epicentre версии 3.0 в онтологию на языке OWL DL. Выполнена русификация описания сущностей и атрибутов модели Epicentre, а также соответствующих им классов и свойств на OWL. Построение онтологии реализовано на диалекте языка OWL DL, соответствующем правилам дескриптивной логики, что в дальнейшем позволит дополнить онтологию системой логического вывода.

Программная реализация конвертора модели Epicentre в язык OWL DL выполнена на языке Java с использованием генератора лексических анализаторов flex [11] и генератора синтаксических анализаторов CUP [12].

Для апробации интеграции баз данных на основе онтологий и проверки корректности построенных онтологий разработан программный интерфейс OakOwlProject 1.0, обеспечивающий навигацию и манипуляции с построенной онтологией. В качестве прототипа для реализации некоторых интерфейсных и функциональных возможностей среды OakOwlProject был выбран известный Java-проект с открытым исходным кодом Protégé [13], функционал которого был существенно расширен. Внешний вид интерфейса OakOwlProject в части его работы с онтологиями предметной области показан на рис. 4.

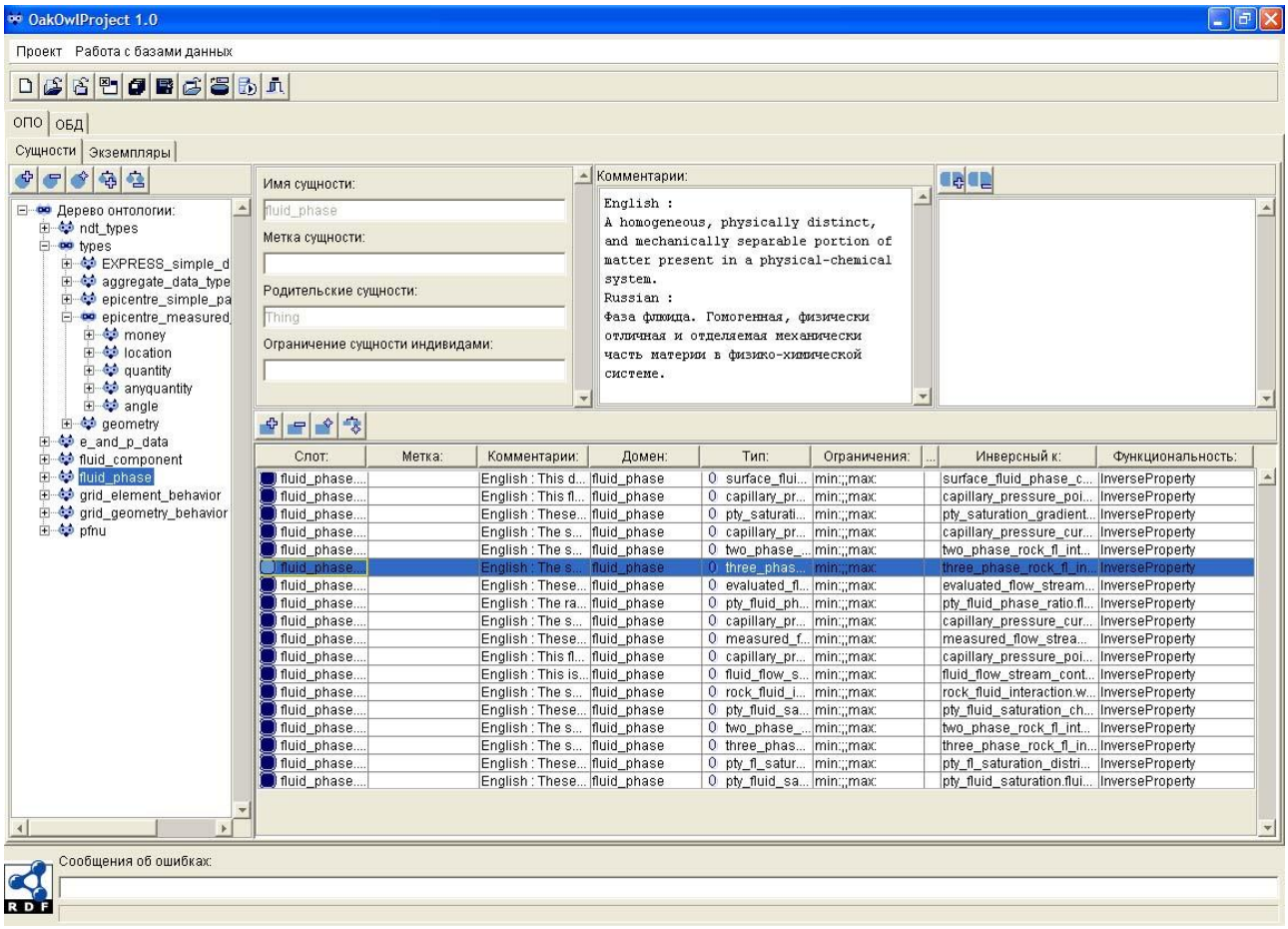


Рис. 4. Общий вид интерфейса пользователя системы OakOwlProject

Описание модели Epicentre на языке EXPRESS представляет собой последовательность описаний сущностей, каждая из которых имеет синтаксическую структуру, представленную на рис. 5.

```
ENTITY name_entity
        ABSTRACT SUPERTYPE OF (
                ONEOF(name_entity1, ..., name_entityN)
        )
        SUBTYPE OF (sub_name1, ..., sub_nameN);
        name_attribute: OPTIONAL SET[0 :?] OF type_name
        ……..
        INVERSE
                invers_attribute_name: SET[0 :?] OF invers_name_entity
                                                FOR entity_invers_type
        ……..
        UNIQUE
        si: name_attribute1, ..., name_attributeN
        WHERE условие
        ……..
END_ENTITY;
```

Рис. 5. Описаний сущностей на языке EXPRESS

В данной синтаксической конструкции используются следующие обозначения:

name_entity – имя сущности; может иметь приставку Pty_ или Ref, указывающую на сущность–свойство или справочную сущность соответственно;

name_entity1, ... , name_entityN – список сущностей;

name_attribute – имя прямого атрибута, может иметь приставку Ref_;

sub_name – имя родительской сущности, может быть обычной сущностью (приставка отсутствует) или сущностью-свойством (имеется приставка Pty_);

sub_name1, ... , sub_nameN – список сущностей sub_name;

type_name – имя типа прямого атрибута, может быть справочной сущностью (Ref_), обычной сущностью (приставка отсутствует) или именованным типом данных (приставка Ndt_);

inverse_attribute_name – имя инверсного атрибута, может быть обычной сущностью (приставка отсутствует) или сущностью-свойством (имеется приставка Pty_);

inverse_name_entity – имя сущности, которая связана обратным отношением с заданной сущностью, может быть справочной сущностью (Ref_), обычной сущностью (приставка отсутствует) или сущностью-свойством (Pty_);

entity_inverse_type – имя сущности, которая является типом прямого атрибута, соответствующего данному инверсному атрибуту, может быть обычной или справочной сущностью;

name_attribute1, ... , name_attributeN – список атрибутов.

Смысл ключевых слов, набранных прописными буквами, будет описан ниже.

На рис. 6 представлен способ конвертации сущностей в классы OWL.

```
ENTITY name_entity          =>   <owl:Class rdf:ID=" name_entity ">
...                              ...
END_ENTITY;                      </owl:Class
```

Рис. 6. Конвертация сущностей в классы OWL

В Epicentre базовыми классами являются: E_AND_P_DATA, FLUID_COMPONENT, FLUID_PHASE, GRID_ELEMENT_BEHAVIOR, GRID_GEOMETRY_BEHAVIOR, PFNU. Автоматически данные классы будут являться подклассами класса owl:Thing и располагаться в корне таксономического дерева. В оболочке OakOwlProject_1.0, например, базовые классы находятся в корне иерархии, показанной на рис. 7.

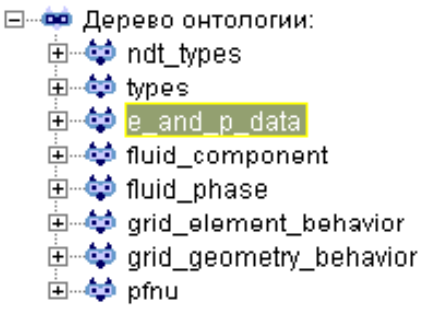


Рис. 7. Базовые классы модели

Конвертация отношений наследования показана на рис. 8. Здесь в конструкции SUBTYPE OF перечисляется список родительских сущностей для данной сущности. Данному отношению в OWL однозначно соответствует синтаксическая конструкция subclass.

```
SUBTYPE OF (                        <owl:Class rdf:ID="name_entity">
    sub_name1, ..., sub_nameN =>    <rdfs:subClassOf rdf:resource="# sub_name1" />
);                                          ...
                                            <rdfs:subClassOf rdf:resource="# sub_nameN" />
                                            ...
                                    </owl:Class>
```

Рис. 8. Конвертация отношений наследования классов

Здесь в конструкции SUBTYPE OF перечисляется список родительских сущностей для заданной сущности. Данному отношению в OWL однозначно соответствует синтаксическая конструкция subclass.

Например, сущности aircraft:

```
ENTITY aircraft
        SUBTYPE OF (  general_facility );
        UNIQUE
        si: identifier,
        identifying_facility,
        ref_existence_kind;
END_ENTITY;
```

соответствует следующий класс OWL:

```
        <owl:Class rdf:ID="aircraft">
        <rdfs:subClassOf rdf:resource="#general_facility" />
        </owl:Class>
```

В оболочке OakOwlProject_1.0 это отношение отображается в поле «Родительские сущности» (рис. 9).

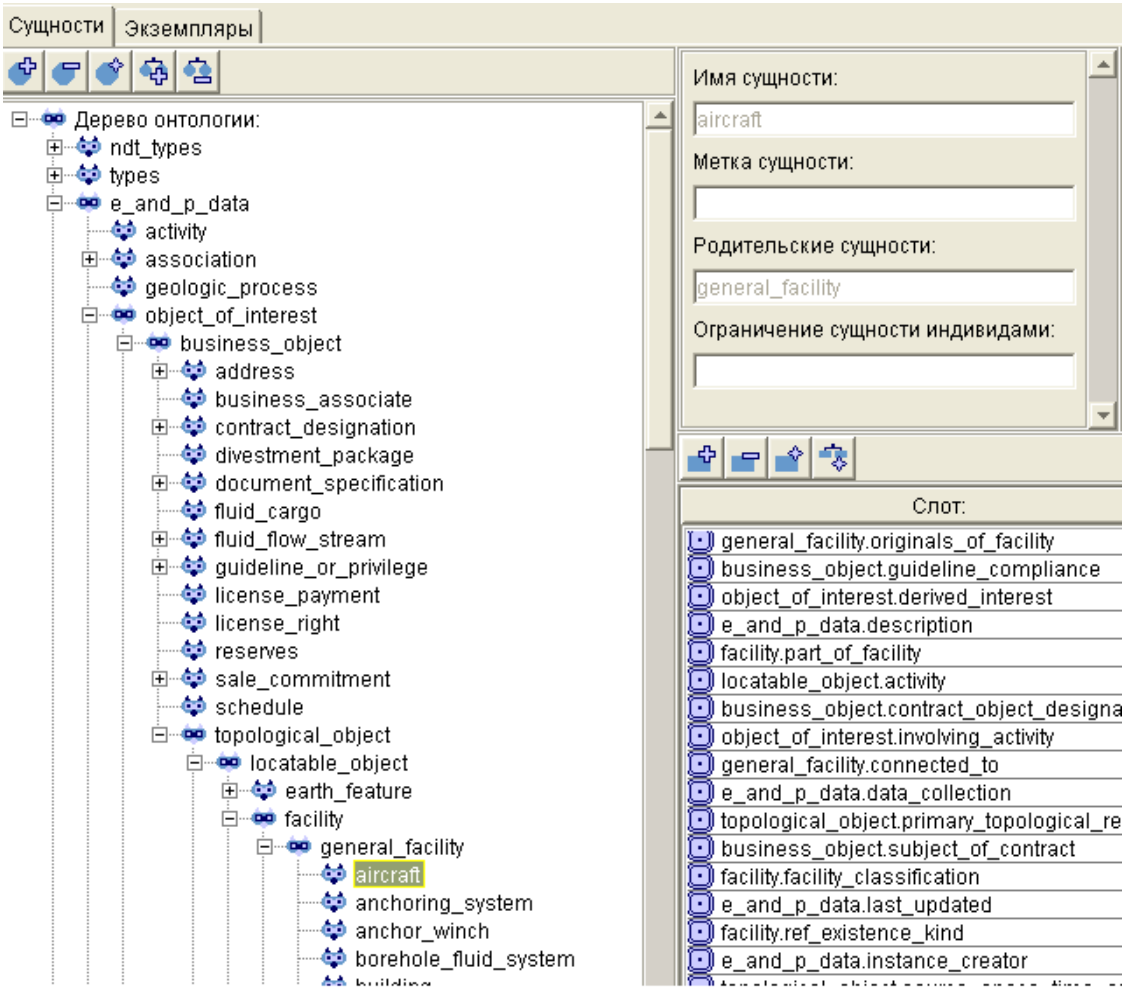


Рис. 9. Отношение наследования

Следующая синтаксическая конструкция языка EXPRESS перечисляет список сущностей, для которых данная сущность является родительской. Ключевое слово ABSTRACT указывается только для абстрактных сущностей (рис. 10).

ABSTRACT SUPERTYPE OF (ONEOF(

name_entity1, ..., name_entityN

Рис. 10. Список сущностей, для которых данная сущность является родительской

В синтаксисе языка OWL отсутствуют средства явного указания классов, которые наследуются из данного класса. Однако в онтологии OWL наследование является обязательным, и оно присутствует в таксономическом дереве онтоло-

гии. Например, в оболочке OakOwlProject_1.0 для абстрактной сущности GRID_ELEMENT_BEHAVIOR, раскрывая список, можно видеть всех ее прямых потомков (рис. 11).

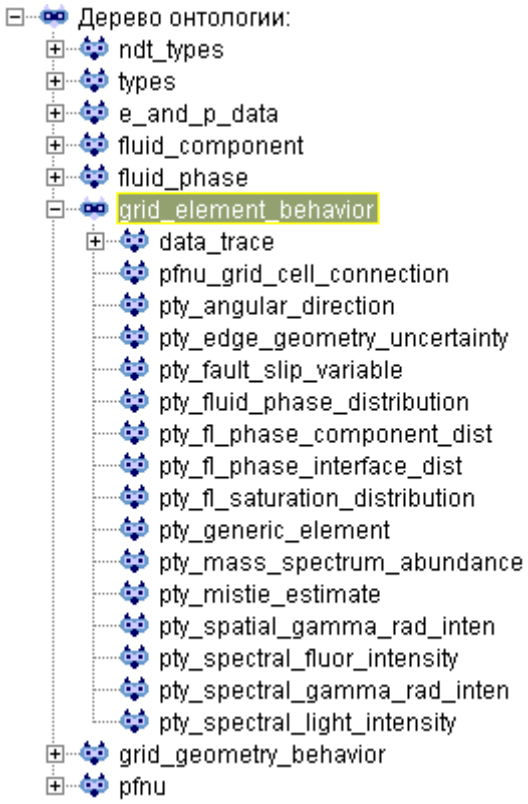


Рис. 11. Визуализация прямых потомков

После определения названия сущности и ее места в иерархии в модели Epicentre приводится список атрибутов, как прямых, так и инверсных. Определению прямого атрибута сущности соответствует синтаксическая конструкция на рис. 12.

name_attribute: OPTIONAL SET[0 :?] OF type_name

Рис. 12. Определение прямого атрибута сущности

Синтаксические конструкции OPTIONAL и SET[0:?] OF могут отсутствовать. Ключевое слово OPTIONAL определяет необязательность наличия значения атрибута, а конструкция SET[0:?] OF определяет степень связи (один-к-одному, один-ко-многим, многие-ко-многим), если атрибут является агрегатным. Такому блоку в OWL соответствует определение простых свойств-атрибутов, если свя-

занная сущность не является типом данных, и свойств-значений в противном случае. Таким образом, такой блок может быть представлен в OWL синтаксической конструкцией, показанной на рис. 13.

```
<owl:ObjectProperty rdf:ID="name_attribute">
        <rdfs:domain rdf:resource="#name_entity"/>
        <rdfs:range rdf:resource="#type_name"/>
</owl:ObjectProperty>
```

Рис. 13. Определение в OWL свойств-атрибутов и свойств-значений

Здесь имя свойства будет соответствовать имени атрибута, а отношение будет связывать исходную сущность и некоторый класс. Проблема повторения имен (например, названия атрибута-свойства и сущности-класса, с которой существует связь), решается введением точечной нотации, то есть свойство будет называться в онтологии name_entity.name_attribute. Если атрибут имеет значения некоторого типа данных, то по смыслу должно быть свойство-значение, но, так как в Epicentre типы имеют более сложную структуру, фактически получаем те же свойства-объекты.

Указание степени связи между классами, а также необязательность (OPTIONAL) можно реализовать с помощью понятия кардинальности в OWL.

```
ENTITY well_test_open_period_recovery
        SUBTYPE OF ( well_test_recovery );
        time_to_surface: OPTIONAL ndt_time;
END_ENTITY;
```

Рис. 14. Ограничение на свойство на языке EXPRESS

Так, OPTIONAL означает, что связь может быть и 0, поэтому в описании класса добавляется ограничение на свойство. Например, для сущности well_test_open_period_recovery (рис. 14) описание в OWL будет следующим (см. рис. 15). Определение самого свойства показано на рис. 16.

```
<owl:Class rdf:ID="well_test_open_period_recovery">
        <rdfs:subClassOf rdf:resource="#well_test_recovery"/>
        <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty>
                                <owl:ObjectProperty rdf:resource=
                                "#well_test_open_period_recovery.time_to_surface"/>
                        </owl:onProperty>
                        <owl:minCardinality>0</owl:minCardinality>
                        <owl:maxCardinality>1</owl:maxCardinality>
                        <owl:cardinality>0</owl:cardinality>
                </owl:Restriction>
        </rdfs:subClassOf>
</owl:Class>
```

Рис. 15. Реализация степени связи на OWL

```
<owl:ObjectProperty>
        <rdf:about>#well_test_open_period_recovery.time_to_surface
        </rdf:about>
        <rdfs:domain>
                <owl:Class>
                        <rdf:about>#well_test_open_period_recovery</rdf:about>
                </owl:Class>
        </rdfs:domain>
        <rdfs:range>
                <owl:Class>
                        <rdf:about>#ndt_time</rdf:about>
                </owl:Class>
        </rdfs:range>
</owl:ObjectProperty>
```

Рис. 16. Определение свойства на OWL

Определению инверсного атрибута сущности соответствует следующая синтаксическая конструкция (рис. 17):

```
INVERSE
invers_attribute_name: SET[0 :?] OF invers_name_entity
        FOR entity_invers_type;
```

Рис. 17. Определение инверсного атрибута

Данной синтаксической конструкции в OWL можно сопоставить схему, как и в отношении прямого атрибута, только с некоторыми модификациями. Аналогичным образом реализуются и ограничения кардинальности (рис. 18).

```
<owl:ObjectProperty rdf:ID="invers_attribute_name">
        <rdfs:domain rdf:resource="#invers_name_entity"/>
        <rdfs:range rdf:resource="#entity_invers_type"/>
</owl:ObjectProperty>
```

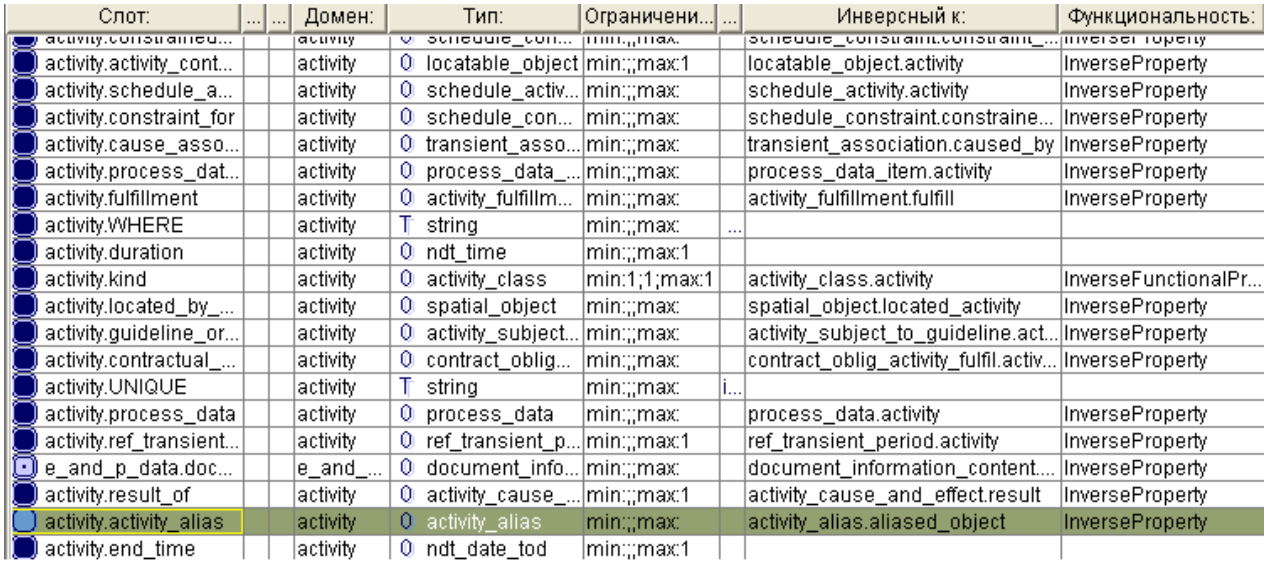
Рис. 18. Ограничения кардинальности

Так, например, у сущности activity есть атрибут activity_alias, который является инверсным к aliased_object, поэтому необходимо обозначить инверсность путем добавления в определение свойства Type: InverseProperty, а также указать, по отношению к какому свойству данное свойство является обратным (рис. 19).

```
<owl:ObjectProperty>
        <rdf:about>#activity.activity_alias</rdf:about>
        <owl:type>
                <rdf:about>#InverseProperty</rdf:about>
        </owl:type>
        <owl:inverseOf>
                <owl:ObjectProperty>
                        <rdf:about>#activity_alias.aliased_object</rdf:about>
                </owl:ObjectProperty>
        </owl:inverseOf>
</owl:ObjectProperty>
```

Рис. 19. Определение инверсного атрибута на OWL

В оболочке OakOwlProject 1.0 все атрибуты сущностей прописаны в поле «Слоты». Наличие переопределений (например, указание кардинальности либо конкретных значений) явно указывается (рис. 20).



Слот:	...	...	Домен:	Тип:	Ограничени...	...	Инверсный к:	Функциональность:
activity.activity_cont...			activity	O locatable_object	min:;;max:1		locatable_object.activity	InverseProperty
activity.schedule_a...			activity	O schedule_activ...	min:;;max:		schedule_activity.activity	InverseProperty
activity.constraint_for			activity	O schedule_con...	min:;;max:		schedule_constraint.constraine...	InverseProperty
activity.cause_asso...			activity	O transient_asso...	min:;;max:		transient_association.caused_by	InverseProperty
activity.process_dat...			activity	O process_data_...	min:;;max:		process_data_item.activity	InverseProperty
activity.fulfillment			activity	O activity_fulfillm...	min:;;max:		activity_fulfillment.fulfill	InverseProperty
activity.WHERE			activity	T string	min:;;max:	...		
activity.duration			activity	O ndt_time	min:;;max:1			
activity.kind			activity	O activity_class	min:1;1;max:1		activity_class.activity	InverseFunctionalPr...
activity.located_by_...			activity	O spatial_object	min:;;max:		spatial_object.located_activity	InverseProperty
activity.guideline_or...			activity	O activity_subject...	min:;;max:		activity_subject_to_guideline.act...	InverseProperty
activity.contractual_...			activity	O contract_oblig...	min:;;max:		contract_oblig_activity_fulfil.activ...	InverseProperty
activity.UNIQUE			activity	T string	min:;;max:	i...		
activity.process_data			activity	O process_data	min:;;max:		process_data.activity	InverseProperty
activity.ref_transient...			activity	O ref_transient_p...	min:;;max:1		ref_transient_period.activity	InverseProperty
e_and_p_data.doc...			e_and_...	O document_info...	min:;;max:		document_information_content....	InverseProperty
activity.result_of			activity	O activity_cause_...	min:;;max:1		activity_cause_and_effect.result	InverseProperty
activity.activity_alias			activity	O activity_alias	min:;;max:		activity_alias.aliased_object	InverseProperty
activity.end_time			activity	O ndt_date_tod	min:;;max:1			

Рис. 20. Визуализация атрибутов

Определению уникальности и ограничений на атрибуты сущности соответствует следующая синтаксическая конструкция (рис. 21):

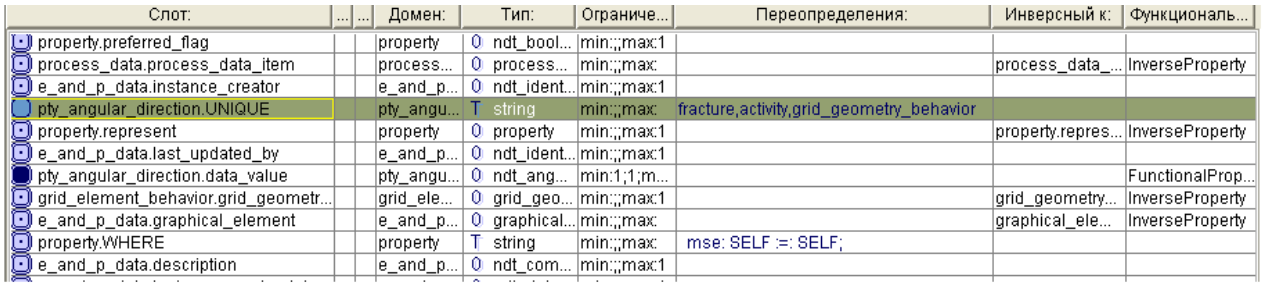
UNIQUE
si: name_attribute1, ..., name_attributeN;
WHERE *условие*

Рис. 21. Определение уникальности и ограничений на атрибуты сущности

Здесь после ключевого слова UNIQUE приведен список атрибутов, образующих уникальный ключ для данной сущности.

В OWL отсутствуют структурные элементы, которые в полном объеме описывают определение уникальности Epicentre. Поэтому в определение каждого класса на языке OWL добавлено новое свойство с именем name_entity.UNIQUE, в котором перечислены все ключевые атрибуты. Аналогичным образом решается проблема ограничений согласованности WHERE: в определение каждого класса на языке OWL добавлено новое свойство с именем name_entity.WHERE, в котором записаны соответствующие условия ограничений.

В оболочке OakOwlProject_1.0 свойства UNIQUE и WHERE, так же, как и атрибуты сущности, относятся к слотам, где в поле «Переопределения» указаны конкретные значения ключей и ограничений (рис. 22).



Слот:	...	...	Домен:	Тип:	Ограниче...	Переопределения:	Инверсный к:	Функциональ...
property.preferred_flag			property	0 ndt_bool...	min:;;max:1			
process_data.process_data_item			process...	0 process...	min:;;max:		process_data_...	InverseProperty
e_and_p_data.instance_creator			e_and_p...	0 ndt_ident...	min:;;max:1			
pty_angular_direction.UNIQUE			pty_angu...	T string	min:;;max:	fracture,activity,grid_geometry_behavior		
property.represent			property	0 property	min:;;max:1		property.repres...	InverseProperty
e_and_p_data.last_updated_by			e_and_p...	0 ndt_ident...	min:;;max:1			
pty_angular_direction.data_value			pty_angu...	0 ndt_ang...	min:1;1;m...			FunctionalProp...
grid_element_behavior.grid_geometr...			grid_ele...	0 grid_geo...	min:;;max:1		grid_geometry...	InverseProperty
e_and_p_data.graphical_element			e_and_p...	0 graphical...	min:;;max:		graphical_ele...	InverseProperty
property.WHERE			property	T string	min:;;max:	mse: SELF :=: SELF;		
e_and_p_data.description			e_and_p...	0 ndt_com...	min:;;max:1			

Рис. 22. Визуализация свойств UNIQUE и WHERE

Типы данных в модели Epicentre классифицированы на следующие категории:

- **Simple Data Types** – простые;
- **Simple Pattern Types** – простые структурные;
- **Measured Quantity Types** – метрические численные;
- **Geometry Types** – геометрические.

Для каждой категории данных в онтологии OWL построены отдельные классы, в свойствах которых использовались встроенные типы данных языка OWL.

ЗАКЛЮЧЕНИЕ

Для синтаксического описания модели Epicentre 3.0 корпорации POSC построена формальная LR(1) грамматика, которая использовалась в качестве входного файла генератора синтаксических анализаторов Java Cup. На основе этой грамматики и схемы конвертации модели Epicentre средствами языка Java построена онтология предметной области природно-технических объектов на языке OWL. Не все конструкции модели Epicentre могут быть синтаксически выражены средствами языка OWL. Для конвертации такой информации были определены специальные классы и зарезервированные свойства OWL, которые могут быть использованы семантически адекватно модели Epicentre.

Общий объем LR(1) грамматики с встроенной в нее семантикой конвертации составил порядка 30 страниц. Файл с описанием модели Epicentre на языке EXPRESS содержит около 500 страниц. Полученная модель на языке OWL имеет объем около 3500 страниц текста. Таким образом, в данной работе полностью, без потери информации, осуществлено преобразование модели Epicentre в OWL-

онтологию. Визуализация построенной онтологии предметной области позволяет наглядно убедиться в ее корректности.

Корпоративное веб-приложение ПАО «Татнефть», основанное на интеграции реляционных баз данных в виде: универсальной онтологии реляционных баз данных, онтологии предметной области, лингвистического тезауруса, и генерирующего SQL-запросы, задаваемые на русском языке, – описано в работах [14–18]. Способ реализации извлечения информации из комментариев реляционных баз данных существенно использует методы компьютерной лингвистики; подробно соответствующий алгоритм извлечения знаний изложен в этих же работах.

Методы оптимизации алгоритма построения SQL-запросов, наиболее ресурсоемкая часть которого связана с необходимостью перебора соединений таблиц, рассмотрены в работе [19].

Благодарности

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 1.2368.2018/ПЧ, а также при финансовой поддержке Российского фонда фундаментальных исследований, проект 18-07-00964.

СПИСОК ЛИТЕРАТУРЫ

1. *Гаврилова Т.А., Хорошевский Т.А.* Базы знаний интеллектуальных систем. СПб.: Питер, 2001. 384 с.

2. *Дейт К.Д.* Введение в системы баз данных. М.: Изд. Дом «Вильямс», 2001. 72 с.

3. *Epicentre v3.0.* URL: http://www.energistics.org/energistics-standards-directory/epicentre-archive.

4. *OWL Web Ontology Language.* URL: https://www.w3.org/TR/2004/REC-owl-features-20040210/

5. Towards the Semantic Web: Ontology-Driven Knowledge Management. Chicester, UK: John Wiley & Sons, 2003. 312 p.

6. *The World Wide Web Consortium (W3C).* URL: http://www.w3c.org.

7. *RDF 1.1 Concepts and Abstract Syntax*. URL: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

8. *RDF Schema 1.1.* URL: https://www.w3.org/TR/rdf-schema/

9. *Extensible Markup Language (XML).* URL: https://www.w3.org/XML/.

10. *Льюис Ф., Розенкранц Д., Стирнз Р*. Теоретические основы проектирования компиляторов. М.: Мир, 1979. 654 с.

11. *Allmon B.J., Anderson J.* Flex on Java. Manning Publications Co. Greenwich, CT, USA, ISBN: 1933988797, 2010. 264 p.

12. *CUP Parser Generator for Java.* URL: https://www.cs.princeton.edu/~appel/modern/java/CUP/

13. *Protégé.* URL: http://protege.stanford.edu/.

14. *Birialtsev E., Bukharaev N., Gusenkov A.* Intelligent search in Big Data // Journal of Physics: Conference Series. V. 913, conference 1. Published online: 25 October 2017.

15. *Гусенков А.М.* Интеллектуальный поиск сложных объектов в массивах больших данных // Электронные библиотеки. 2016. Т. 19. № 1. С. 3–39.

16. *Гусенков А., Биряльцев Е., Жибрик О.* Интеллектуальный поиск в структурированных массивах информации. LAP LAMBERT Academic Publishing. Deutschland: OmniScriptum Marceting DEU GmbH, ISBN 978-3-659-76919-1, 2015. 129 с.

17. *Гусенков А.М., Биряльцев Е.В.* Интеграция реляционных баз данных на основе онтологий // Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2007. Т. 149. Кн. 2. С. 13–34.

18. *Gusenkov A., Bukharaev N., Birialtsev E.* On ontology based data integration: problems and solutions // Journal of Physics: Conference Series. V. 1203, conf. 1. 012059. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1203/1/012059/meta

19. *Gusenkov A., Bukharaev N.* On Semantic Search Algorithm Optimization // New Knowledge in Information Systems and Technologies. WorldCIST'19. Advances in Intelligent Systems and Computing, V. 930. Springer, Cham, 2019. URL: https://link.springer.com/chapter/10.1007/978-3-030-16181-1_45.

BUILDING SUBJECT DOMAIN ONTOLOGY ON THE BASE OF A LOGICAL DATA MOD

A. M. Gusenkov[1], N. R. Bukharaev[2], E. V. Biryaltsev[3]

[1,2]*Kazan Federal University, Kazan, Russia;*

[3]*Center of Digital Technologies of the Institute of Applied Research of the Academy of Sciences of the Republic of Tatarstan, Kazan, Russia*

[1]gusenkov.a.m@gmail.com, [2]boukharay@gmail.com, [3]igenbir@yandex.ru

Abstract

The technology of automated construction of the subject domain ontology, based on information extracted from the comments of the TATNEFT oil company relational databases, is considered. The technology is based on building a converter (compiler) translating the logical data model of Epicenter Petrotechnical Open Software Corporation (POSC), presented in the form of ER diagrams and a set of the EXPRESS object-oriented language descriptions, into the OWL ontology description language, recommended by the W3C consortium. The basic syntactic and semantic aspects of the transformation are described.

Keywords*: subject domain ontology, relational databases, POSC, OWL*

REFERENCES

1. *Gavrilova T.A., Khoroshevskii T.A.* Bazy znanii intellektualnykh system. SPb.: Piter, 2001. 384 p.

2. *Date C.J. Deit K.D.* Vvedenie v sistemy baz dannykh. M.: Izd. Dom «Viliams», 2001. 72 p.

3. *Epicentre v3.0.* URL: http://www.energistics.org/energistics-standards-directory/epicentre-archive.

4. *OWL Web Ontology Language.* URL: https://www.w3.org/TR/2004/REC-owl-features-20040210/.

5. Towards the Semantic Web: Ontology-Driven Knowledge Management. Chicester, UK: John Wiley & Sons, 2003. 312 p.

6. *The World Wide Web Consortium (W3C).* URL: http://www.w3c.org

7. *RDF 1.1 Concepts and Abstract Syntax*. URL: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

8. *RDF Schema 1.1.* URL: https://www.w3.org/TR/rdf-schema/

9. *Extensible Markup Language (XML).* URL: https://www.w3.org/XML/.

10. *Lewis P., Rosenkrantz D., Stearns R.* Teoreticheskie osnovy proektirovaniia kompiliatorov. M.: Mir, 1979. 654 p.

11. *Allmon B.J., Anderson J.* Flex on Java. Manning Publications Co. Greenwich, CT, USA, ISBN: 1933988797, 2010. 264 p.

12. *CUP Parser Generator for Java.* URL: https://www.cs.princeton.edu/~appel/modern/java/CUP/

13. *Protégé.* URL: http://protege.stanford.edu/

14. *Birialtsev E., Bukharaev N., Gusenkov A.* Intelligent search in Big Data // Journal of Physics: Conference Series, V. 913, conf. 1. Published online: 25 October 2017.

15. *Gusenkov A.M.* Intellektualnyi poisk slozhnykh obieektov v massivakh bolshikh dannykh // Elektronnye biblioteki. 2016. T. 19. No 1. S. 3–39.

16. *Gusenkov A., Biryaltsev E., Zhibrik O.* Intellektualnyi poisk v strukturirovannykh massivakh informatsii. LAP LAMBERT Academic Publishing. Deutschland: OmniScriptum Marceting DEU GmbH, ISBN 978-3-659-76919-1, 2015. 129 p.

17. *Gusenkov A.M., Biryaltsev E.V.* Integratsiia reliatsionnykh baz dannykh na osnove ontologii // Uchenye zapiski Kazanskogo gosudarstvennogo universiteta. Seriia Fiziko-matematicheskie nauki. 2007. T. 149, book 2. S. 13–34.

18. *Gusenkov A., Bukharaev N., Birialtsev E.* On ontology based data integration: problems and solutions // Journal of Physics: Conference Series, V. 1203, conf. 1. 012059. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1203/1/012059/meta

19. *Gusenkov A., Bukharaev N.* On Semantic Search Algorithm Optimization // New Knowledge in Information Systems and Technologies. WorldCIST'19. Advances in Intelligent Systems and Computing, V. 930. Springer, Cham, 2019. URL: https://link.springer.com/chapter/10.1007/978-3-030-16181-1_45.

СВЕДЕНИЯ ОБ АВТОРАХ

ГУСЕНКОВ Александр Михайлович – к. т. н., доцент Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета. Области научных интересов: технологии извлечения знаний, обработка естественных языков, большие данные, интеллектуальный анализ данных.

Alexander M. GUSENKOV – assistant professor, Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Ph.D. Current scientific interests: knowledge extraction technologies, Natural Language Processing, big data, data mining.

email: gusenkov.a.m@gmail.com

БУХАРАЕВ Наиль Раисович – к. ф.-м. н., доцент Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета. Сфера научных интересов: методология IT-образования, извлечение знаний, большие данные.

Naille R. BUKHARAEV – assistant professor, Institute of Computational Mathematics and Information Technologies, Kazan Federal University. Ph.D. Current scientific interests: IT education methodology, knowledge extraction technologies, big data.

email: boukharay@gmail.com

БИРЯЛЬЦЕВ Евгений Васильевич – к. т. н., специалист в области специализированных информационных систем, автор более 50 публикаций, в том числе 5 свидетельств о регистрации программ, 3 изобретений. Заведующий Центром цифровых технологий Института прикладных исследований Академии наук Республики Татарстан.

Evgeny V. BIRYALTSEV – specialist in the field of specialized information systems, Ph. D., author of more than 50 publications, including 5 certificates of registration of programs, 3 inventions. Head of the Center of digital technologies of the Institute of applied research of the Academy of Sciences of the Republic of Tatarstan.

email: igenbir@yandex.ru

Материал поступил в редакцию 17 ноября 2019 года

УДК 004.62

ВИЗУАЛИЗАЦИЯ ЦИФРОВЫХ 3D-ОБЪЕКТОВ ПРИ ФОРМИРОВАНИИ ВИРТУАЛЬНЫХ ВЫСТАВОК

Н. Е. Каленов, С. А. Кириллов, И. Н. Соболевская, А. Н. Сотников

Межведомственный Суперкомпьютерный Центр РАН – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Москва

ansotnikov@jscc.ru, ins@jscc.ru, skirillov@jscc.ru, nkalenov@jscc.ru

Аннотация

Представлены подходы к решению задачи создания реалистичных интерактивных 3D веб-коллекций музейных экспонатов. Рассмотрено представление 3D-моделей объектов на основе ориентированных полигональных структур. Описан метод создания виртуальной коллекции 3D-моделей по технологии интерактивной анимации. Также показано, как на основе отдельных кадров экспозиции с помощью методов фотограмметрии строится высококачественная 3D-модель. Приведены результаты расчетов для построения 3D-моделей реальных музейных экспонатов. Для создания 3D-моделей с целью предоставления их широкому кругу пользователей через интернет использована технология интерактивной анимации. Приведены различия между представлениями цифровых 3D-моделей. Описана технология создания цифровых 3D-моделей объектов из фондов Государственного биологического музея им. К.А. Тимирязева и формирования на их основе средствами электронной библиотеки «Научное наследие России» виртуальной выставки, посвященной научной деятельности М.М. Герасимова и его антропологическим реконструкциям. Выставка наглядно продемонстрирована возможности интеграции информационных ресурсов средствами электронной библиотеки. Формат виртуальных выставок позволил объединить ресурсы партнеров для предоставления широкому кругу пользователей коллекций, хранящихся в музейных, архивных и библиотечных фондах.

***Ключевые слова**: фотограмметрия, 3D-моделирование, интерактивная мультипликация, веб-дизайн, полигональное моделирование*

ВВЕДЕНИЕ

Одним из способов представления междисциплинарных коллекций в распределенной среде электронной библиотеки является формирование виртуальной выставки. Виртуальная выставка – это мультимедийный информационный ресурс, демонстрирующий пользователям разнородную информацию (цифровые копии печатной продукции, архивных документов, музейных предметов и т. п.), объединенную по заданным признакам. Наряду с представлением материалов различных типов в процессе формирования цифровых естественнонаучных коллекций возникает необходимость в мультимедийных объектах, в частности, цифровых 3D-моделях музейных предметов и объектах виртуальной реальности [1]. Формируемые виртуальные выставки могут быть представлены не только в интернете, но и стать частью или даже основным элементом реальной музейной экспозиции.

В последние годы широко используются методы визуализации, основанные на фотограмметрических снимках, полученных в заданном диапазоне длин волн электромагнитного излучения, и дальнейших построениях трехмерных структур из последовательностей двумерных изображений, которые могут быть объединены с локальными сигналами движения. Этот метод называется Structure from motion (SfM) [2–4]. В биологическом зрительном восприятии SfM представляет собой «аппарат», посредством которого люди (и другие живые существа) могут восстанавливать трехмерную структуру из проецируемого на сетчатке глаза 2D-поля движения движущегося объекта или сцены. Этот метод применяется в области компьютерного моделирования, связанного с моделированием зрительного восприятия. Однако существуют такие граничные условия, при которых этот метод «не работает», например, при наличии стеклянных и отражающих поверхностей.

Для создания трехмерных моделей и элементов виртуальной выставки применяются разные программные и технологические решения, в частности, технологии лазерного и оптического 3D-сканирования, компьютерного моделирования, фотограмметрии, анимации 3600 и др. [5].

1. ТЕХНОЛОГИИ ЛАЗЕРНОГО И ОПТИЧЕСКОГО 3D СКАНИРОВАНИЯ

Технологии лазерного и оптического 3D-сканирования позволяют создать цифровую копию предмета, например, музейного, с помощью 3D-сканера [6].

Оптические 3D-сканеры используют технологию структурированного света (рис. 1). На сканируемый объект направляется проекция световой сетки. Анализ деформации световых линий сетки и позволяет вычислить форму поверхности сканируемого объекта [7].

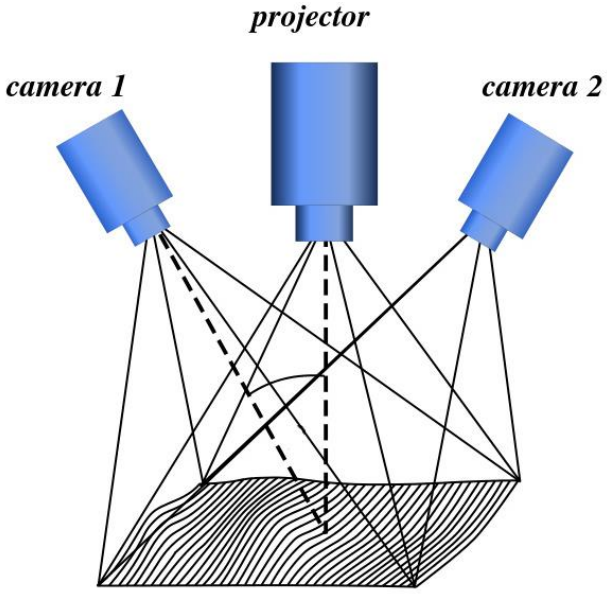


Рис. 1. Схема работы оптического 3D-сканера

Оптические 3D-сканеры используются для быстрой оцифровки различных мелких и средних предметов, так как одновременно могут оцифровать множество точек или все «поле зрения» сканера. А использование в качестве источника света специальных ламп белого или синего цвета позволяет выполнять оцифровку геометрии и захват текстуры при низком освещении. В нашем проекте мы использовали ручные оптические сканеры фирм Artec и Creaform. Были получены 3D-модели музейных предметов различной природы: гипсовых скульптур, посуды (в том числе стеклянной), музейных предметов растительного и животного вида.

На рис. 2 показана 3D-модель чучела змеи, находящегося в запасниках Государственного биологического музея им. К.А. Тимирязева. Это чучело было вы-

брано в качестве опытного образца для создания 3D-модели объекта, поверхность которого состоит из глянцевых светопоглощающих материалов [8]. На рис. 3 показана 3D-модель чучела утки, также находящегося в запасниках Государственного биологического музея им. К.А. Тимирязева. Оно было выбрано в качестве опытного образца для создания 3D-модели объекта, поверхность которого состоит из светопропускающих материалов (в данном случае – перьев).

Рис. 2. Чучело змеи

Рис. 3. Чучело утки

В целом мы получили качественные, законченные полигональные модели, которые можно использовать для формирования реалистичных интерактивных 3D-коллекций музейных экспонатов. Однако часть 3D-моделей музейных предметов имела много шумов и погрешностей, вызванных прозрачностью или блес-

ком материала. В частности, не очень реалистично выглядит перьевой и шерстяной покровы у чучел птиц и млекопитающих. Поэтому было принято решение провести дополнительное сканирование лазерным 3D-сканером.

Лазерные 3D-сканеры (рис. 4) обеспечивают наибольшие точность и детализацию при оцифровке объектов, они оборудованы специализированным лазером, который относят ко II классу. Лазер данного типа достаточно безопасен для человеческого зрения. Особенностью использования данного типа сканеров является применение специальных маркеров, которые крепятся в непосредственной близости от объекта или непосредственно на объекте сканирования. Это необходимо для точной пространственной привязки лазера и сканируемого объекта.

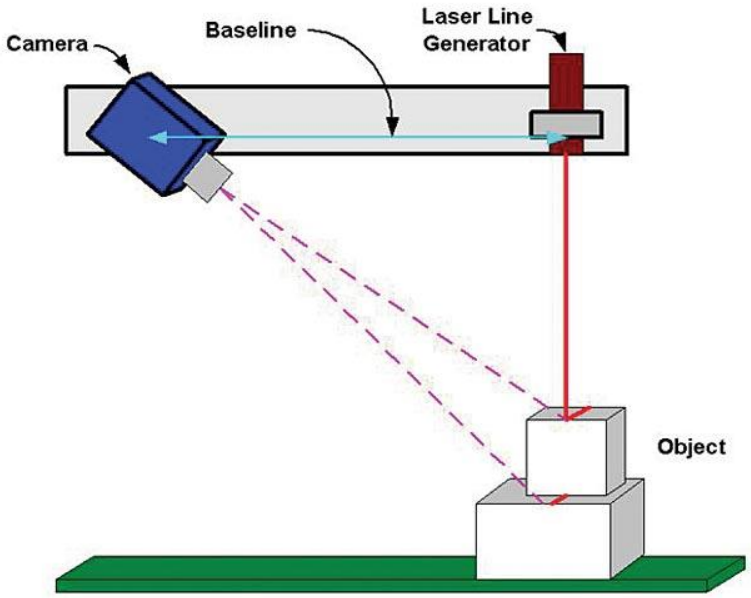


Рис. 4. Принцип работы лазерного сканера

В работе по формированию коллекций музейных объектов использовался лазерный 3D-сканер Creaform HandySCAN 700. Данный сканер оборудован 7 высокоточными лазерами и системой динамической привязки TRUaccuracy [9], что обеспечивает точность сканирования до 0,03 мм с разрешением 0,05 мм.

Основным недостатком сканера является его неспособность передавать цвет текстуры объекта, однако он обеспечивает вполне удовлетворительные результаты при решении задач по построению максимально детальной поверхности сканируемого объекта. Использование данного типа сканирующего устройства

обеспечило высокое качество 3D-моделей гипсовых антропологических реконструкций, выполненных М.М. Герасимовым (http://acadlib.ru/index.php/ pages). Опыт 3D-оцифровки показал, что некоторые музейные предметы оказались слишком сложными для обработки существующими на сегодняшний день 3D-сканерами. Так, результаты сканирования шерстяного покрова чучела мыши-полёвки оказались неудовлетворительными, для их улучшения потребовалась бы обработка «проблемных зон» (светопоглощающей шерсти или светоотражающих глаз) специальным составом, что, возможно, нанесло бы ущерб сканируемым объектам. Одно из основных требований, предъявляемых к оцифровке музейных объектов – обеспечение максимальной сохранности объекта при сканировании, решение данной проблемы требует развития и применения технологий компьютерного моделирования и/или интерактивной анимации 360^0.

2. ТЕХНОЛОГИИ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

Технологии компьютерного моделирования позволяют решать задачи по визуализации, например, безвозвратно утраченных или поврежденных музейных предметов, производить детализацию реконструируемых объектов, восстанавливать обстановку исторических помещений и т. д. В нашем проекте эта технология используется при редактировании полученных методом сканирования 3D-моделей, обладающих проблемными областями. В таких случаях оператору в программе 3D-моделирования приходится заново создавать отдельные элементы 3D-модели.

Рис. 5. Бюст австралопитека с «дефектом»

Рис. 6. Бюст австралопитека без «дефекта»

На рис. 5 показана цифровая копия бюста австралопитека, представляющего собой антропологическую реконструкцию, выполненную М.М. Герасимовым. Красная стрелка указывает на дефект объекта. На рис. 6 показан этот же объект, но после ручной цифровой «реставрации».

3. ТЕХНОЛОГИИ ФОТОГРАММЕТРИИ

Технология фотограмметрии [10, 11] позволяет построить высококачественную 3D-модель. Эта технология активно разрабатывается с 1970-х годов и изначально применялась для построения карт рельефа по аэрофотоснимкам. Фотограмметрия использует способы и приёмы различных дисциплин, в основном, заимствованных из оптики и проективной геометрии. В простейшем случае пространственные координаты точек объекта определяются путём измерений, выполняемых по двум или более фотографиям, снятым из разных положений. Основной задачей в этом случае является определение общих точек на двух соседних изображениях. После создания массива общих точек формируется набор прямых, проходящих через каждую общую точку и местоположение фотоаппарата (точки съемки). Пересечение этих прямых и определяет расположение точки на поверхности исходного объекта в пространстве. Более сложные алгоритмы могут использовать другую, известную заранее информацию об объекте, например, симметрию элементов объекта, что в некоторых случаях позволяет реконструировать пространственные координаты точек объекта по ограниченному количеству фотоизображений.

4. ТЕХНОЛОГИЯ ИНТЕРАКТИВНОЙ АНИМАЦИИ 360^0

Для создания виртуальной коллекции 3D-моделей с целью предоставления ее широкому кругу пользователей может быть применена технология интерактивной анимации [12]. Эта технология не предполагает построения полигональной 3D-модели, а основана на программной смене фиксированного набора кадров с помощью специализированных интерактивных программ отображения, имитирующих вращение объекта. Важно, что на основе такого же набора данных (отдельных кадров экспозиции) может быть построена и высококачественная 3D-модель с помощью методов фотограмметрии.

Для проведения таких работ применяется, в частности, комплекс 3D-оцифровки на основе поворотной платформы Recam T-50, управляющей программы 3D-Maker [13] и цифрового фотоаппарата Canon EOS600D, который позволяет выполнять в автоматическом режиме съемку музейных предметов высотой до 150 см и весом до 50 кг.

На рис. 7 представлена последовательность кадров, на основе программной смене которых создается отображение цифровой 3D-модели.

Рис. 7. Последовательность кадров, на основе программной смены которых создается отображение цифровой 3D-модели

ЗАКЛЮЧЕНИЕ

Развитие методов формирования 3D-моделей в направлении получения реалистичного представления коллекций различных предметов открывает возможности для формирования 3D-коллекций музейных объектов высокого качества, как для обеспечения сохранности оригиналов, так и для расширения доступности высококачественных цифровых копий музейных экспонатов [14, 15].

Полученные результаты легли в основу технологии формирования коллекций 3D-моделей объектов из фондов Государственного биологического музея им. К.А. Тимирязева и формирования на их основе виртуальных выставок средствами электронной библиотеки «Научное наследие России» [8], в частности, виртуальной выставки, посвященной научной деятельности М.М. Герасимова и его антропологическим реконструкциям, доступной по адресу http://acadlib.ru/.

Работа выполнена в МСЦ РАН – филиале ФГУ ФНЦ НИИСИ РАН в рамках государственного задания № 0065-2019-0014.

СПИСОК ЛИТЕРАТУРЫ

1. *Ляшков А.А., Панчук К.Л., Варепо Л.Г.* Особенность отображения гиперповерхности четырехмерного пространства.// М.: Геометрия и графика, 2017.

2. *Wróżyński R., Pyszny K., Sojka M., Przybyła C., Murat-Błażejewska S.* Ground volume assessment using 'Structure from Motion' photogrammetry with a smartphone and a compact camera // Open Geosciences. 2017. V. 9. P. 281–294.

3. Scopigno R. Digital fabrication techniques for cultural heritage: a survey // Comput. Graph. Forum. 2017. V. 36. P. 6–21.

4. *Garstki K.*// Virtual representation: the production of 3D digital artifacts // J. Archaeol. Method Theory. 2017. V. 24. P. 726–750.

5. *Gonizzi Barsanti S., Guidi G.* 3D digitization of museum content within the 3D icons project //ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2013. P. 151–156.

6. *Gonizzi Barsanti S., Guidi G.* 3D digitization of museum content within the 3D icons project //ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2013. P. 151–156.

7. *Прасолов В.В., Тихомиров В.Н.* Геометрия. М.: МЦНМО, 2007.

8. *Кириллов С.А., Соболевская И.Н., Сотников А.Н.* Использование мультимедийных технологий при формировании виртуального естественнонаучного музейного пространства // Информационное обеспечение науки: новые технологии. М.: 2017. С. 201–207.

9. *Wohlfeil J., Strackenbrock B., Kossyk I.* Automated high resolution 3d reconstruction of cultural heritage using multi-scale sensor systems and semi-global matching.// International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences. 2013. V. 40-4. P. 37–43.

10. *Лобанов А.Н. Фотограмметрия. М*.: «Недра», 1984.

11. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация (Practical optimization). М.: Мир, 1985.

12. *Сотников А.Н., Соболевская И.Н., Кириллов С.А., Чередниченко И.Н.* Технологии визуализации 3d web-коллекций // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 438–447.

13. *Kratochvil J., Sadilek M., Musil V., Pagac M., Stancekova D*. The effectiveness of strategies printing printer easy 3d maker// Advances in Science and Technology-research Jurnal, 2018. V. 12. P. 197–205.

14. *Hernando A., Bobadilla J., Ortega F., Gutiérrez A.* Method to interactivelyvisualize and navigate related information // Expert Systems with Applications. 2018.

15. *Schulz T., Juttler B.* Envelope Computation by Approximate Implicitization// Industrial Geometry. 2010. 20 p.

DIGITAL 3D-OBJECTS VISUALIZATION IN FORMING VIRTUAL EXHIBITIONS

N. E. Kalenov, S. A. Kirillov, I. N. Sobolevskaya, A. N. Sotnikov

Joint Supercomputer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences"

ansotnikov@jscc.ru, ins@jscc.ru, skirillov@jscc.ru, nkalenov@jscc.ru

Abstract

The paper is presents approaches to solving the problem of creating realistic interactive 3D web-collections of museum exhibits. The presentation of 3D-models of objects based on oriented polygonal structures is considered. The method of creating a virtual collection of 3D-models using interactive animation technology is described. It is also shown how a full-fledged 3D-model is constructed on the basis of individual exposure frames using photogrammetry methods. The paper assesses the computational complexity of constructing realistic 3D-models. For the creation of 3D-models in order to provide them to a wide range of users via the Internet, the so-called interactive animation technology is used. The paper presents the differences between the representations of full-fledged 3D-models and 3D-models presented in the form of interactive multiplication. The technology of creating 3D-models of objects from the funds of the State Biological Museum named K.A Timiryazev and the formation on their basis of the digital library “Scientific Heritage of Russia” of a virtual exhibition dedicated to the scientific activities of M.M. Gerasimov and his anthropological reconstructions, and vividly demonstrating the possibility of integrating information resources by means of an electronic library. The format of virtual exhibitions allows you to combine the resources of partners to provide a wide range of users with collections stored in museum, archival and library collections.

Keywords*: photogrammetry, 3D-modeling, interactive animation, web-design, polygonal modeling.*

REFERENCES

1. *Lyashkov A.A., Panchuk K.L., Varepo L.G.* Osobennost' otobrazheniya giperpoverkhnosti chetyrekhmernogo prostranstva. M.: Geometriya i grafika, 2017. S. 3–10.

2. *Wróżyński R., Pyszny K., Sojka M., Przybyła C., Murat-Błażejewska S.* Ground volume assessment using 'Structure from Motion' photogrammetry with a smartphone and a compact camera // Open Geosciences. 2017. V. 9. P. 281–294.

3. Scopigno R. Digital fabrication techniques for cultural heritage: a survey // Comput. Graph. Forum. 2017. V. 36. P. 6–21.

4. *Garstki K.*// Virtual representation: the production of 3D digital artifacts // J. Archaeol. Method Theory. 2017. V. 24. P. 726–750.

5. *Gonizzi Barsanti S., Guidi G.* 3D digitization of museum content within the 3D icons project //ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2013. P. 151–156.

6. *Gonizzi Barsanti S., Guidi G.* 3D digitization of museum content within the 3D icons project //ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2013. P. 151–156.

7. *Prasolov V.V., Tikhomirov V.N.* Geometriya. M.: MTSNMO, 2007.

8. *Kirillov S.A., Sobolevskaya I.N., Sotnikov A.N.* Ispol'zovaniye mul'timediynykh tekhnologiy pri formirovanii virtual'nogo yestestvennonauchnogo muzeynogo prostranstva// Informatsionnoye obespecheniye nauki: novyye tekhnologii. M., 2017. S. 201–207.

9. *Wohlfeil J., Strackenbrock B., Kossyk I.* Automated high resolution 3d reconstruction of cultural heritage using multi-scale sensor systems and semi-global matching.// International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences. 2013. V. 40-4. P. 37–43.

10. *Lobanov A.N.* Fotogrammetriya*. M.: «*Nedra*»,* 1984.

11. *Gill F., Myurrej U., Rajt M. Prakticheskaya optimizaciya (Practical optimization). M.: «Mir», 1985.*

12. *Sotnikov A.N., Sobolevskaya I.N., Kirillov S.A., CHerednichenko I.N.* Tekhnologii vizualizacii 3d web-kollekcij// Nauchnyj servis v seti Internet: trudy XX Vserossijskoj nauchnoj konferencii (17—22 sentyabrya 2018 g., g. Novorossijsk). M.: IPM im. M.V. Keldysha, 2018. S. 438—447.

13. *Kratochvil J., Sadilek M., Musil V., Pagac M., Stancekova D*. The effectiveness of strategies printing printer easy 3d maker// Advances in Science and Technology-research Jurnal, 2018. V. 12. P. 197–205.

14. *Hernando A., Bobadilla J., Ortega F., Gutiérrez A.* Method to interactivelyvisualize and navigate related information // Expert Systems with Applications. 2018.

15. *Schulz T., Juttler B.* Envelope Computation by Approximate Implicitization// Industrial Geometry. 2010. 20 p.

СВЕДЕНИЯ ОБ АВТОРАХ

КАЛЕНОВ Николай Евгеньевич – главный научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. т. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема.

Nikolay Evgenvich KALENOV – Chief Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing.

email: nkalenov@jscc.ru

КИРИЛЛОВ Сергей Александрович – зав. сектором Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук».

KIRILLOV Sergey Alexandrovich – Head. sector of the Interdepartmental Supercomputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”.

email: skirillov@jscc.ru

СОБОЛЕВСКАЯ Ирина Николаевна – старший научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», к. ф.-м. н. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; 3D-моделирование.

Irina Nikolaevna SOBOLEVSKAYA – senior scientist researcher of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; 3D modeling.

email: ins@jscc.ru

СОТНИКОВ Александр Николаевич – зам. директора по научной работе Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. ф.-м. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; нейронные и семантические сети.

Aleksandr Nikolaevich SOTNIKOV – Deputy director for science of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; semantic and nerve nets.

email: ansotnikov@jscc.ru

Материал поступил в редакцию 16 ноября 2019 года

УДК 004.62

ФОРМАЛИЗАЦИЯ ПРОЦЕССОВ ФОРМИРОВАНИЯ ПОЛЬЗОВАТЕЛЬСКИХ КОЛЛЕКЦИЙ В ЦИФРОВОМ ПРОСТРАНСТВЕ НАУЧНЫХ ЗНАНИЙ

Н. Е. Каленов, И. Н. Соболевская, А. Н. Сотников

Межведомственный Суперкомпьютерный Центр РАН (МСЦ РАН) – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Москва

nkalenov@jscc.ru, ins@jscc.ru, ansotnikov@jscc.ru

Аннотация

Исследована задача формирования цифрового пространства научных знаний (ЦПНЗ). Рассмотрено отличие этого понятия от общего понятия пространства знаний. ЦПНЗ представлено как множество, содержащее объекты, верифицированные мировым научным сообществом. Формой структурированного представления цифрового пространства знаний является семантическая сеть, основной принцип организации которой основан на системе классификации объектов и последующем построении их иерархии, в частности, по принципу наследования. Введена классификация объектов, составляющих контент ЦПНЗ. Предложена модель ЦПНЗ как совокупности непересекающихся множеств, содержащих цифровые образы реальных объектов и их характеристики, обеспечивающие отбор и визуализацию объектов в соответствии с многоаспектными пользовательскими запросами. Определено понятие пользовательской коллекции, предложена иерархическая классификация типов пользовательских коллекций. Использование понятий теории множеств при построении ЦПНЗ позволяет разбивать информацию по уровням детализации и формализовать алгоритмы обработки пользовательских запросов, что проиллюстрировано конкретными примерами.

Ключевые слова*: семантическая сеть, информационное пространство, научные знания, электронная библиотека, уровни детализации, иерархия информационных объектов.*

ВВЕДЕНИЕ

Информация играет центральную роль во многих сферах нашей жизни. Развитие информационных и вычислительных технологий расширило возможности для сбора, анализа, распространения, обработки и использования научной информации.

Современные потребности в профессиональной информации требуют развития пространства знаний, представляющего собой цифровую среду, в которую интегрированы информационные ресурсы и сервисы из разных областей науки, культуры и образования. Частью общего пространства знаний является цифровое пространство научных знаний (ЦПНЗ), отличающееся от других составляющих общего пространства (в частности, такого, как Википедия) тем, что информационные объекты, представленные в ЦПНЗ, верифицированы мировым научным сообществом и отделены от информационных объектов, которые носят идеологический, религиозный и другой спорный с научной точки зрения характер [1].

Поток запросов в ЦПНЗ часто является непрерывным, быстро меняющимся во времени, не всегда предсказуемым и неограниченным по форме запроса. Программное обеспечение, обрабатывающее такие запросы, не может позволить себе хранить и «пересматривать» параметры запроса, часто требующего быстрого ответа в режиме реального времени. Требования точности поиска данных в ЦПНЗ (в отличие от общих поисковых машин интернета) обусловливают необходимость разработки специальных методов обработки поисковых запросов с обеспечением достаточно точного отображения текста запроса на пространство метаданных, описывающих те или иные объекты ЦПНЗ. Метаданные ЦПНЗ, в свою очередь, включают не только наборы ключевых слов, но и более сложные структуры, например, иерархические классификационные системы.

Формой структурированного представления цифрового пространства знаний является семантическая сеть, основной принцип организации которой основан на системе классификации объектов и последующем построении их иерархии, в частности, по принципу наследования: «макроэкономика» – раздел «экономики», «поэтический сборник» – издание и т. п. В соответствии с этим принципом объекты классифицируются на некоторое число категорий или классов на основании их общих свойств.

Большинство цифровых коллекций данных представляет собой разнородную информационную сеть, связывающую объекты различного типа. Например, электронная публикация (тип объекта – «книга»), помимо простого текста, содержит дополнительную информацию, такую, как автор публикации (тип объекта – «персона»), год издания, издательство, место издания и т. п. В свою очередь, объект «персона», кроме последовательности символов, задающих фамилию, связан с биографией, областью научных интересов («тематика объекта») и т. п. Таким образом, от объекта «публикация» может быть установлена связь с другим «объектом» («автор»), текстом этой публикации, «тематикой объекта» и т. п. В общем случае ЦПНЗ должно поддерживать различные типы связей между его элементами – как внутри одного класса объектов (в частности, рекурсивную иерархическую связь), так и между объектами различных классов.

Вопросу построения тематических иерархий, иерархии понятий, объектных моделей и т. д., обеспечивающих иерархическую организацию данных на разных уровнях детализации и имеющих такие приложения, как задачи веб-поиска и просмотра, посвящено значительное количество исследований [2,3, 4].

В [5] описан алгоритм NetClus, который позволяет устанавливать связи между многотипными объектами для создания высококачественных сетевых кластеров. Алгоритм NetClus позволяет переупорядочивать объекты атрибутов в каждом вновь определенном сетевом кластере.

В работе [6] нами предложена иерархия представления объектов в среде электронной библиотеки, которую можно рассматривать как прообраз ЦПНЗ.

В данной работе мы рассматриваем ЦПНЗ в аспекте теории множеств, что позволяет подойти к вопросам построения пространства и работы с ним с новой точки зрения.

1. СТРУКТУРА ЦПНЗ

Пусть Ω – ЦПНЗ, содержащее все множество элементов цифрового научного пространства, размещенных в некотором (возможно, распределенном) хранилище. Оно включает, в свою очередь, *n* подпространств, каждое из которых относится к определенной области науки. Каждое подпространство $\boldsymbol{\Omega}_{\mathrm{i}}$ состоит из двух множеств. Первое из них (обозначим его A_{i}) состоит из пронумерованных неко-

торым образом цифровых образов объектов реального мира (оцифрованные публикации, архивные документы, фотографии и пр.) и объектов, созданных исключительно в цифровой среде (электронные публикации, 3D-модели, мультимедийные материалы и т. п.). Нумерация должна однозначно идентифицировать объект и обеспечивать возможность его извлечения из хранилища. Второе множество (обозначим его B_{i}) включает метаданные, содержащие многоаспектные характеристики объектов первого множества, обеспечивающие их выборку по запросам к ЦПНЗ и представление пользователям.

Множество A_{i} состоит из элементов a_{ij}, где *j=1…N* (*N* – общее количество объектов, отраженных в Ω_{i}). В качестве этих элементов выступают объекты следующих видов:

- текстовые файлы (распознанные оцифрованные печатные или рукописные документы) или документы, изначально сформированные в электронном виде;
- статические изображения (нераспознанные оцифрованные документы, оцифрованные или изначально сформированные в цифровом виде фотографии);
- цифровые или оцифрованные аудиозаписи;
- цифровые или оцифрованные видео/киноматериалы;
- 3D-модели различных предметов;
- мультимедийные инсталляции (цифровые модели природных процессов и технических устройств, учебные материалы, виртуальные экскурсии и т. п.).

Если элементы множества A_{i} представлены простой совокупностью пар «объект – его номер», то множество B_{i}, в общем случае, представляет собой достаточно сложную фасетно-иерархическую структуру. Каждый его элемент представлен не только конкретным значением и ссылкой на элемент множества a_{ij} (что имеет место в традиционных библиографических информационно-поисковых системах), но может включать указание на связи с другими элементами. Таким образом, под элементами множества B_{i} будем понимать структуру, включающую смысловое значение характеристики объекта, указания на один или несколько элементов множества A_{i}, к которому относится данная характеристика, и

указание на связи с другими структурами, являющимися также элементами множества B_{i}.

В качестве составляющих элементов множества B_{i} могут выступать индексы классификационных систем (таких, как ГРНТИ, УДК и пр.), отражающих тематику документов, индивидуальные характеристики персоны (фамилия и имя, дата рождения и т. п.), наименования событий, их текстовые описания, временные и географические характеристики объектов и др.

Для обеспечения точности поиска объектов в ЦПНЗ множество B_{i} должно включать ряд непересекающихся подмножеств, характеризующих различные аспекты информации об элементах множества A_{i}. Очевидно, что таких разбиений может быть бесконечно много, но ограничимся рассмотрением «интуитивно-минимального» набора данных (но охватывающего широкий спектр характеристик объектов), включающего классы типа «что (кто), где, когда» (подмножество $B_{\mathrm{i}3}$,), дополненного классом «тематика» (подмножество $B_{\mathrm{i}4}$), формальными характеристиками, специфичными для ЦПНЗ, выделенными в подмножества $B_{\mathrm{i}1}$ (виды объектов, перечисленные выше) и $B_{\mathrm{i}2}$ (условия предоставления пользователям тех или иных объектов множества A_{i}).

Подмножество $B_{\mathrm{i}1}$ множества B_{i} состоит из 6-ти элементов, совпадающих с перечисленными выше:

b_{i11} – текстовый вид с возможностью поиска фрагмента текста;

$b_{\mathrm{i}12}$ – статическое изображение;

$b_{\mathrm{i}13}$ – 3D-объект;

$b_{\mathrm{i}14}$ – аудиодокумент;

$b_{\mathrm{i}15}$ – видеодукумент;

$b_{\mathrm{i}16}$ – – мультимедийный объект.

Подмножество $B_{\mathrm{i}2}$ множества B_{i} состоит из элементов, определяющих условия предоставления цифрового объекта пользователю. Введение данного подмножества обусловлено различными законодательными требованиями к публичному представлению объекта. Элементы множества B_2:

$b_{\mathrm{i}21}$ – объект находится в свободном доступе;

$b_{\mathrm{i}22}$ – объект находится в доступе, бесплатном для определенной группы

пользователей (например, оплаченная подписка на полнотекстовые научные издания для сотрудников некоторого учреждения) и недоступном для остальных пользователей;

$b_{\mathrm{i}23}$ – объект находится в ограниченном доступе, бесплатном для определенной группы и коммерческом для остальных пользователей (например, цифровая модель музейного экспоната может быть доступна для бесплатного просмотра посетителям музея, а удаленный просмотр предусматривает определенную плату;

$b_{\mathrm{i}24}$ – объект находится в коммерческом доступе, т. е. пользователю необходимо оплатить доступ к данному ресурсу.

Подмножество B_{i3} множества B_{i} включает: обозначение **типа** объекта (персона, организация, публикация, архивный документ, минерал, теорема и т. п., в зависимости от направления науки); основные характеристики объекта, необходимые для его идентификации при поиске («**что**» или «**кто**», «**где**», «**когда**»); **условия визуализации**. В качестве обязательного элемента множества B_{i3}, относящегося к классу «**что (кто)**», выступает название конкретного объекта (имя персоны), которое может быть дополнено элементами, конкретизирующими вид объекта внутри данного типа (например, для типа «публикация» это могут быть варианты: «научная монография», «научная статья», «поэтический сборник», учебник и т. д.), а также неструктурированными пояснениями, содержащими ту или иную информацию об объекте. Это может быть биография ученого, аннотация публикации, описание музейного предмета и т. п.). Например, коллекция фотографий Москвы 1930-х годов может быть дополнена развернутой статьей об архитектуре города того времени, представленной в виде гипертекста.

В качестве элемента класса «**где**» множества B_{i3} могут выступать различные реалии, связанные как непосредственно с географической принадлежностью объекта (например, для персоны – место рождения, для музейного объекта – место его первоначального обнаружения, для события – страна или город, где оно произошло, и т. п.), так и с организацией, описанной, в свою очередь, своими метаданными (например, места работы персоны, место хранения музейного предмета, издательство для печатного документа и т. п.).

В качестве элемента класса «**когда**» множества B_3 может выступать, напри-

мер, год публикации печатного издания, год рождения персоны, дата запуска космического корабля и т. п.

Класс **«условия визуализации»** содержит информацию о группах пользователей, которым может предоставляться данный объект без каких-либо условий, и условия предоставления объекта другим группам пользователей

Отметим, что подмножества B_{i1}, B_{i2} и B_{i3} множества B_{i} не пересекаются между собой.

Подмножество B_{i4} содержит элементы класса «тематика», оно может иметь достаточно сложную структуру, содержащую индексы и наименования элементов различных классификационных систем – строго иерархическую типа ГРНТИ [7], фасетную – типа УДК [8] и т. п.), ключевые термины, в том числе, оформленные в виде тезаурусов.

2. МОДЕЛЬ ОБРАБОТКИ ЗАПРОСОВ В ЦПНЗ

Для упрощения дальнейшего изложения будем предполагать, что мы рассматриваем подпространство ЦПНЗ, относящееся к одной из научных областей (если не будет сказано иного), и опустим нижний индекс при рассмотрении множеств A_{i} и. B_{i} Соответственно, при обозначении подмножеств этих множеств перейдем от двойных индексов к одинарным. Обозначим множество запросов пользователей, в соответствии с которыми из ЦПНЗ отбираются интересующие их документы, через F. Его составляющие (f_s) могут содержать элементы естественного языка, рубрики тех или иных классификационных систем, химические формулы, математические выражения и т. п., связанные булевыми операторами. Это множество, в отличие от конечных множеств A и B, содержит бесконечное число элементов ($F = \bigcup_s^{\infty} f_s$). Его элементами являются как разовые запросы отдельных пользователей, так и постоянные запросы, формируемые в рамках систем избирательного распространения информации [10, 11], а также запросы, в соответствии с которыми в ЦПНЗ формируются те или иные коллекции документов.

Если некоторый запрос f_s удовлетворяет условию $f_s = \bigcup_n^N f_{sn}$, где $f_{sn} \in B$ (т. е. f_{sn} представляет собой логическое выражение, включающее элементы одного из подмножеств B_{i}), то задача выбора и визуализации объектов из ЦПНЗ сводится к сравнению составляющих запроса f_{sn} с элементами подмножеств B_{i} (i=1,3,4) (назовем это линейным поиском). Результатом сравнения являются

адреса соответствующих элементов множества A, по которым эти элементы извлекаются из хранилища и предоставляются пользователю в соответствии с условиями, отраженными в подмножестве B_2.

Однако на практике запросы пользователя зачастую либо пересекаются с множеством *B* лишь частично, либо вообще не пересекаются. Это приводит к тому, что результат линейного поиска содержит лишь часть объектов ЦПНЗ, необходимых пользователю, либо не содержит их вообще, хотя во множестве A они имеются. Например, пользователю необходима подборка произведений поэтов «Серебряного века», отраженных в электронной библиотеке (ЭБ) публикаций XX века. При формировании элементов ЭБ, с большой долей вероятности, «Серебряный век» не указывался в качестве временной характеристики, его также нет ни в одной из классификационных систем, которые могли использоваться при формировании ЭБ. Соответственно, если обработать запрос в терминах «поэты Серебряного века», его результатом будет пустое подмножество элементов множества A. В то же время, очевидно, что среди элементов множества A есть объекты, соответствующие требованиям, предъявляемым к данной коллекции. Для их обнаружения необходимо построить отображение пользовательского запроса на множество B и далее реализовать линейный поиск по запросу, включающему соответствующие элементы подмножеств B_1 и B_3.

3. ФОРМИРОВАНИЕ ПОЛЬЗОВАТЕЛЬСКИХ КОЛЛЕКЦИЙ

Под пользовательской коллекцией будем понимать коллекцию элементов ЦПНЗ или электронной библиотеки, соответствующих запросу, сформулированному на естественном языке.

Следуя принципам формализации общего подхода к формированию пользовательских коллекций, построим иерархию их представления в ЦПНЗ.

На первом уровне этой иерархии располагаются элементы множества A, отобранные в соответствии с запросом, отображаемым на подмножество B_3 (предметные коллекции); на втором – элементы множества A, соответствующие запросу, отображаемому на подмножества B_1 и B_3 (предметно-видовые коллекции); третий уровень (тематико-видовые коллекции) формируется по запросам, отражаемым на подмножества B_1 и B_4; на четвертом уровне располагаются тематические коллекции, соответствующие запросам, отражаемым на подмножество

B_4. Наконец, на самом верхнем уровне иерархии располагаются междисциплинарные коллекции. Запросы на формирование таких коллекций не могут быть отражены с необходимой полнотой на одном множестве B, соответствующем той или иной области науки. Для получения полной коллекции объектов, соответствующих такому запросу, необходимо рассматривать несколько подпространств ЦПНЗ и строить отражение запроса на соответствующие подмножества нескольких множеств B_i.

В качестве примера коллекции первого уровня можно привести подборку всех материалов, касающихся конкретного ученого. Например, на запрос «М.В. Ломоносов» будут выданы произведения М.В. Ломоносова; связанные с ним публикации; МГУ им. М.В. Ломоносова; хребет Ломоносова и т. д.

Коллекция опубликованных трудов М.В. Ломоносова относится ко второму уровню иерархии. Отражение такого запроса на подмножество B_1 предписывает отбирать объекты вида b_{11} и b_{12}. Отражение запроса на подмножество B_3 позволяет осуществить линейный поиск по условию: «выбрать тип объекта «персона» с фамилией Ломоносов, инициалами М. В. (класс характеристик **«Кто»**) и тип объекта «публикация», в авторах которых (класс характеристик **«Кто»**) указаны выбранные персоны. В результате будет получен ряд полных текстов публикаций, автором которых является М.В. Ломоносов. Однако, в силу того, что в ЭБ или в ЦПНЗ могут быть отражены несколько персон с «именем» М.В. Ломоносов, сформированная коллекция будет содержать все их публикации. Если при запросе на формирование коллекции подразумевался конкретный Михаил Васильевич Ломоносов, родившийся в 1711 году, полученный результат будет некорректным. Чтобы получить требуемую коллекцию, отражение запроса на множество B_3 должно содержать год рождение персоны (класс характеристик **когда**).

В качестве примера пользовательской коллекции третьего уровня можно привести объекты, извлеченные из ЦПНЗ по запросу «3D-модели антропологических объектов». Для получения такой коллекции этот запрос необходимо отразить на подмножества B_1 и B_4. Первое предписывает отбирать объекты вида b_{13}, а для получения второго необходимо анализировать конкретные классификационные системы, используемые в ЭБ (ЦПНЗ).

К пользовательской коллекции верхнего уровня можно отнести такие коллекции, как «материалы, связанные с освоением космического пространства», в

которые должны быть включены объекты, относящиеся к физике, механике, технике, химии, астрономии, возможно, к философии, политологии и т. п.

Создание такой иерархии позволяет оптимизировать процесс формирования и сопровождения информационных фондов электронных библиотек, а также позволяет пользователю выбрать из всего множества взаимосвязанных ресурсов электронной библиотеки те информационные объекты, которые объединены одним или несколькими признаками.

В соответствии с моделью алгоритм формирования пользовательской коллекции включает следующие этапы:

1. Анализ соответствия терминов, включенных в запрос f_s, элементам множества B, определение уровня иерархии, которому соответствует пользовательская коллекция.
2. Разбиение терминов запроса на два подмножества: подмножество, содержащее в явном виде элементы множества B (например, вид объекта, тип объекта и т. п.) $- f_{s1}$, и подмножество, не содержащее в явном виде элементы $B\ - f_{s2}$.
3. Реализация алгоритма отображения элементов подмножества f_{s2} на множество B; формирование запроса $f_s^{\ b} \in B$.
4. Линейный поиск элементов множества *A,* отвечающих запросу $f_s^{\ b}$.
5. Формирование пользовательской коллекции в соответствии с условиями визуализации.

В качестве примера реализации описанного алгоритма рассмотрим формирование коллекции материалов, относящихся к поэтам «Серебряного века», в среде некоторой условной электронной библиотеки, организованной по принципам ЭБ «Научное Наследие России» (ЭБ ННР) [12].

Пусть имеется некоторая ЭБ *X*, отражающая культурное наследие России в части, относящейся к литературным произведениям. В качестве объектов ЭБ X выступают персоны (авторы) и публикации. Метаданные персон включают следующие элементы.

Фасет **КТО:**

- фамилия;
- имя;
- отчество;

- варианты имени (псевдонимы).

Фасет **ГДЕ:**

- место рождения;
- место смерти.

Фасет **КОГДА:**

- дата рождения;
- дата смерти;

Фасет **ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ:**

биография;

библиография.

Метаданные публикаций включают следующие элементы:

Фасет **ЧТО:**

- название публикации

- вид представления публикации (ссылка на элемент множества B_1);
- тип публикации (проза, поэзия, литературоведческая работа);
- вид публикации (монография, сборник, том многотомника, выпуск сериального издания, статья из сборника или из сериального издания, прочие виды);

Фасет **КТО:**

- ссылки на персон (авторов);
- ссылки на персон (редакторов, составителей, художников).

Фасет **ГДЕ:**

- место издания (страна, город, издательство);
- ссылка на публикацию для статей из журналов, сборников и т. п.;
- информация о конкретном выпуске издания (журнала, сборника), в котором опубликован данный материал;

Фасет **КОГДА:**

- год издания публикации.

- Фасет **ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ:**

полное библиографическое описание материала;

аннотация;

примечания.

Фасет **ТЕМАТИКА:**

индексы УДК;
индексы ББК;
ключевые термины.

Публикации, представленные в виде аудиозаписей, в фасете **КОГДА** должны содержать информацию о дате создания звукозаписи, а в фасете **ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ** (или в фасете **КТО**, если участники создания звукозаписи входят в круг персон, отраженных в ЭБ Х) – информацию о ее создателях.

Задача состоит в том, чтобы сформировать пользовательскую коллекцию материалов, входящих в ЭБ Х, относящихся к поэтам «Серебряного века». Коллекция должна включать сведения об авторах и обо всех документах, связанных с ними (в том числе, выходные данные и полные тексты их произведений).

Для решения этой задачи определим сначала, к какому временному интервалу относится понятие «Серебряный век».

Согласно «Литературной энциклопедии» [13], «Серебряный век» заключен в интервале между 1890-м и 1921-м годами.

Формирование искомой пользовательской коллекции включает следующие этапы:

- На первом этапе определяются годы публикаций (отражение на множество B), тем самым «переводится» параметр запроса «Серебряный век» на «язык» метаданных.
- На втором этапе необходимо выбрать из множества объектов «публикация» те, в фасете **ЧТО** метаданных которых имеется элемент «тип публикации», содержащий значение «поэзия», а в фасете **КОГДА** – один из годов, входящих в заданный интервал. Необходимо отметить, что для упрощения задания интервалов годов поисковый интерфейс ЭБ (ЦПНЗ) должен предусматривать в фасете **КОГДА** возможность обработки логических выражений, содержащих условия «больше», «меньше», «равно», «не равно». Соответственно, программная оболочка ЭБ должна уметь обрабатывать такие условия. Публикации, найденные на этом этапе, включаются в пользовательскую коллекцию.
- На третьем этапе выбираются персоны, ссылка на которых имеется в фасете **КТО** выбранных на предыдущем этапе публикаций. Данные об

этих персонах включаются в пользовательскую коллекцию.

- На четвертом этапе осуществляется визуализация сформированной коллекции в соответствии с условиями по каждому объекту. Программные средства ЭБ (ЦПНЗ) должны обеспечивать гибкие возможности визуализации элементов пользовательской коллекции сортировку по различным элементам метаданных различных объектов (это может быть не только список объектов, упорядоченный по алфавиту (или числовому значению) заданного элемента метаданных, но и список, в котором чередуются объекты разного вида). В частности, для рассматриваемого примера это может быть информация о поэте, за которой следует список его произведений со ссылками на полные тексты.

4. ЗАКЛЮЧЕНИЕ

Предлагаемые подходы к формализации процессов представления элементов ЦПНЗ и построения пользовательских коллекций позволяют упростить алгоритмизацию построения отдельных элементов ЦПНЗ, разработки его программной оболочки и пользовательского интерфейса. Работы в этом направлении проводятся в рамках государственного задания.

СПИСОК ЛИТЕРАТУРЫ

1. *Антопольский А.Б., Каленов Н.Е., Серебряков В.А., Сотников Н.А.* Точка зрения о едином цифровом пространстве научных знаний // Вестник Российской академии наук, 2019 (в печати).
2. *Gauch S., Chaffee J., Pretschner A*. Ontology-based personalized search and browsing. // Web Intell Agent Syst. 2003. V. 1. No 3, 4. P. 219–234.
3. *Sun Y., Yu Y., Han J.* Ranking-based clustering of heterogeneous information networks with star network schema // KDD '09 Proceedings of the 15th ACM SIGKDD international Conference on Knowledge discovery and data mining. 2009. P 797–806.
4. *Wong W., Liu W., Bennamoun M.* Ontology learning from text: a look back and into the future // ACM Computing Surveys (CSUR). 2012. V. 44. Issue 4. Article No 20.

5. *Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, Jiawei Han*. Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems. 2015. V. 44. Issue 3. P. 529–558.
6. *Каленов Н.Е., Соболевская И.Н., Сотников А.Н.* Иерархические уровни представления информационных объектов в среде электронных библиотек // Информация и инновации. 2018. Т. 13. № 2. С. 25–31.
7. *Антопольский А.Б., Белоозеров В.Н., Маркарова Т.С., Дмитриева Е.Ю.* Установление соответствий рубрик ГРНТИ рубрикам других систем классификации научной и технической информации // Научно-техническая информация. Серия 1: организация и методика информационной работы. 2015. № 3. С. 3–18.
8. *Астахова Т.С.* Проблемы отражения современного научного знания в классификационных системах: новое в УДК // Сборник трудов конференции «Перспективные направления научных исследований и критические технологии в классификационных системах» / ВИНИТИ РАН, Москва, 25–27 октября 2017 г. С. 32–35.
9. *Александров П.С.* Введение в теорию множеств и общую топологию. М.: «Наука», 1977. 368 с.
10. *Ивановский А.А.* Объектная модель системы избирательного распространения информации // Научные и технические библиотеки. 2019. № 4. С. 61–75. DOI 10/33186/1027-3689-2019-4-61-75
11. *Захарова С.С.* Избирательное распространение информации и информационно-коммуникационные технологии: обзор исследований // Библиотековедение. 2017. № 6. С. 651–658. DOI: 10.25281/0869-608X-2017-66-6-651-658
12. *Каленов Н.Е., Савин Г.И., Серебряков В.А., Сотников А.Н.* Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты, системы и алгоритмы. Электронный журнал. 2012. Т. 4. № 100. С. 30– 40. Url: http://www.swsys-web.ru
13. *Литературная энциклопедия* [Электронный ресурс]. (https://dic.academic.ru/dic.nsf/enc_literature/5383/%D0%A1%D0%B5%D1%80%D0%B5%D0%B1%D1%80%D1%8F%D0%BD%D1%8B%D0%B9) (07.11.2019).

FORMALIZATION OF PROCESSES FOR FORMING USER COLLECTIONS IN THE DIGITAL SPACE OF SCIENTIFIC KNOWLEDGE

N. E. Kalenov, I. N. Sobolevskaya, A. N. Sotnikov

Joint Supercomputer Center of the Russian Academy of Sciences - Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences"

nkalenov@jscc.ru, ins@jscc.ru, ansotnikov@jscc.ru

Abstract

The task of forming a digital space of scientific knowledge (DSSK) is analyzed in the paper. The difference of this concept from the general concept of the information space is considered. DSSK is presented as a set containing objects verified by the world scientific community. The form of a structured representation of the digital knowledge space is a semantic network, the basic organization principle of which is based on the classification system of objects and the subsequent construction of their hierarchy, in particular, according to the principle of inheritance. The classification of the objects that make up the content of the DSSK is introduced. A model of the central data collection system is proposed as a collection of disjoint sets containing digital images of real objects and their characteristics, which ensure the selection and visualization of objects in accordance with multi-aspect user requests. The concept of a user collection is defined, and a hierarchical classification of types of user collections is proposed. The use of the concepts of set theory in the construction of DSSK allows you to break down information into levels of detail and formalize the algorithms for processing user queries, which is illustrated by specific examples.

Keywords*:* recursive link, knowledge cyberdomain, digital library, detail levels, data entries hierarchy.

REFERENCES

1. *Antopol'skij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov N.A*. Tochka zreniya o edinom cifrovom prostranstve nauchnyh znanij // Vestnik Rossijskoj akademii nauk, 2019 (v pechati).

2. *Gauch S., Chaffee J., Pretschner A*. Ontology-based personalized search and browsing. // Web Intell Agent Syst. 2003. V. 1. No 3, 4. P. 219–234.
3. *Sun Y., Yu Y., Han J.* Ranking-based clustering of heterogeneous information networks with star network schema // KDD '09 Proceedings of the 15th ACM SIGKDD international Conference on Knowledge discovery and data mining. 2009. P 797–806.
4. *Wong W., Liu W., Bennamoun M.* Ontology learning from text: a look back and into the future // ACM Computing Surveys (CSUR). 2012. V. 44. Issue 4. Article No 20.
5. *Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, Jiawei Han*. Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems. 2015. V. 44. Issue 3. P. 529–558.
6. *Kalenov N.E., Sobolevskaya I.N., Sotnikov A.N.* Ierarhicheskie urovni predstavleniya informacionnyh ob"ektov v srede elektronnyh bibliotek // Informaciya i innovacii. 2018. T. 13. No 2. S. 25–31.
7. *Antopol'skij A.B., Beloozerov V.N., Markarova T.S., Dmitrieva E.YU*. Ustanovlenie sootvetstvij rubrik GRNTI rubrikam drugih sistem klassifikacii nauchnoj i tekhnicheskoj informacii // Nauchno-tekhnicheskaya informaciya. Seriya 1: organizaciya i metodika informacionnoj raboty. 2015. No 3.S. 3–18.
8. *Astahova T.S.* Problemy otrazheniya sovremennogo nauchnogo znaniya v klassifikacionnyh sistemah: novoe v UDK // Sbornik trudov konferencii «Perspektivnye napravleniya nauchnyh issledovanij i kriticheskie tekhnologii v klassifikacionnyh sistemah». VINITI RAN, Moskva, 25–27 oktyabrya 2017 g. S. 32–35.
9. *Aleksandrov P.S.* Vvedenie v teoriyu mnozhestv i obshchuyu topologiyu. M.: «Nauka», 1977. 368 s.
10. *Ivanovskij A.A.* Ob"ektnaya model' sistemy izbiratel'nogo rasprostraneniya informacii // Nauchnye i tekhnicheskie biblioteki, 2019. № 4. S. 61–75.
11. *Zaharova S.S.* Izbiratel'noe rasprostranenie informacii i informacionno-kommunikacionnye tekhnologii: obzor issledovanij // Bibliotekovedenie. 2017. No 6. S. 651–658.
12. *Kalenov N.E., Savin G.I., Serebryakov V.A., Sotnikov A.N.* Principy postroeniya i formirovaniya elektronnoj biblioteki "Nauchnoe nasledie Rossii" //

Programmnye produkty, sistemy i algoritmy. Elektronnyj zhurnal. 2012. T. 4. No 100. S. 30–40. Url: http://www.swsys-web.ru

13. Literary encyclopedia [digital resource]. Url: https://dic.academic.ru/dic.nsf/enc_literature/5383/%D0%A1%D0%B5%D1%80%D0%B5%D0%B1%D1%80%D1%8F%D0%BD%D1%8B%D0%B9 (07.11.2019)

СВЕДЕНИЯ ОБ АВТОРАХ

КАЛЕНОВ Николай Евгеньевич – главный научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. т. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема.

Nikolay Evgenvich KALENOV – Chief Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing.

email: nkalenov@jscc.ru

СОБОЛЕВСКАЯ Ирина Николаевна – старший научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», к. ф.-м. н. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; 3D-моделирование.

Irina Nikolaevna SOBOLEVSKAYA – senior scientist researcher of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences". Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; 3D modeling.

email: ins@jscc.ru

СОТНИКОВ Александр Николаевич – зам. директора по научной работе Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. ф.-м. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; нейронные и семантические сети.

Aleksandr Nikolaevich SOTNIKOV – Deputy director for science of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences". Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; semantic and nerve nets.

email: ansotnikov@jscc.ru

Материал поступил в редакцию 16 ноября 2019 года

УДК 001.812 + 004.032.6 + 004.051 + 004.378.5

АУДИОВИЗУАЛЬНАЯ ЗАПИСЬ СИНХРОННЫХ ЗАНЯТИЙ ПРИ ОЧНОМ И ДИСТАНЦИОННОМ ОБУЧЕНИЯХ

Ф. О. Каспаринский

ООО «МАСТЕР-МУЛЬТИМЕДИА», г. Москва

felix@kasparinsky.pro

Аннотация

Современная информационная среда предоставляет беспрецедентные возможности по сочетанию high-tech и high-touch подходов в обучении. Можно ожидать, что в ближайшее время всеобщим трендом станет использование аудиовизуальных записей синхронных занятий, которые целесообразно применять для последующего закрепления, повторения, контроля, обобщения и систематизации знаний. В статье резюмированы результаты 10-летнего опыта создания и использования аудиовизуальных записей очных и дистанционных занятий в университетских и школьных аудиториях.

***Ключевые слова**: аудиовизуальная запись, дистанционное обучение, очное обучение, интернет, Skype, Video, high-touch, high-tech, синхронные занятия*

ВВЕДЕНИЕ

Современное дистанционное обучение сочетает два методических подхода [1]: high-tech (минимизация общения преподавателя с учащимися) и high-touch (обеспечение персонализированного внимания преподавателя к восприятию информации учащимися). High-tech подход к первичному преподнесению учебных материалов минимизирует общение учащихся с преподавателями посредством организации асинхронной передачи информации многочисленным учащимся и автоматизации самостоятельного продвижения по учебной траектории согласно результатам мониторинга активности и контрольных тестов. High-tech форма обучения эффективна для склонных к самостоятельной работе представителей поколений «Беби-бумеров» и «X» (1943–1966 и 1967–1990 годы

рождения соответственно) [2]. High-touch методы, обеспечивающие постоянное персонализированное внимание преподавателя к учащимся, целесообразно применять для детей-инвалидов [3] и представителей современного поколения «Y» (1991–2014 г. р.), которое сформировалось на фоне перехода к глобальному инфоцентрическому миру, становления веб-сервисов и социальных сетей [2]. Адаптированные к переизбытку доступной информации учащиеся поколения «Y» с «клиповым» мышлением предпочтительно усваивают авторитетно анонсированные сведения, передаваемые через экран в процессе синхронного общения, реализуемого в динамичной мультимедийной форме.

В результате 15-летней эволюции аппаратно-программные средства и онлайн-сервисы для записи сетевой трансляции медиаресурсов достигли профессионального уровня [4], позволяющего перейти к дидактически целенаправленному использованию аудиовизуальной формы передачи знаний при очно-дистанционной форме обучения.

1. ОРГАНИЗАЦИЯ ВЕБИНАРОВ И ИХ АУДИОВИЗУАЛЬНАЯ ЗАПИСЬ В ПРОЦЕССЕ ДИСТАНЦИОННОГО ОБУЧЕНИЯ ПО МЕТОДИКЕ HIGH-TECH

В современных системах с преобладающей high-tech формой дистанционного обучения общение учащихся между собой, а также с преподавателями и фасилитаторами сочетает асинхронный обмен сообщениями электронной почты, коммуникационных программ (мессенджеров), форумов, блогов и комментариев с использованием сервисов синхронных видеоконференций (вебинаров) профессионального уровня [5, 6]. Информационная среда таких вебинаров адаптирована к трансляции комментируемых преподавателем презентаций PowerPoint сотням и тысячам учащихся, циркулярному файлообмену и получению обратной реакции посредством опросов и голосований с визуализацией результатов в реальном времени.

Инфопространство high-tech вебинаров определяется интерфейсом приложения для проведения видеоконференций и, как правило, включает главное окно с презентацией, среднее окно с изображением преподавателя или видеорядом с его веб-камеры, и малое окно мессенджера для всех участников вебинара с возможностью ввода своих сообщений. Большинство сервисов предоставляет пользователям возможность видоизменять размеры и положение

окон. В последнее время всё чаще встречаются high-tech вебинары, в ходе которых используются только мессенджер и функционал демонстрации слайдов с голосовым сопровождением.

Потенциальные проблемы при проведении high-tech вебинаров – конфликты Adobe Flash Player и аналогичных надстроек браузеров с системой безопасности операционной системы и информационной средой компьютера пользователя. Для заблаговременного выявления и исправления проблем сервисы профессиональных видеоконференций обязывают участников тестировать используемое оборудование перед началом вебинара.

Видеозапись high-tech вебинаров посредством ресурсов сервиса видеоконференций осуществляется опционально и может быть связана с дополнительной оплатой. Видеозаписи high-tech вебинаров, как правило, распространяются организаторами через общедоступные файловые хостинги и видеосервисы. Альтернативный вариант формирования аудиовизуальной записи high-tech вебинаров – самостоятельный захват видеоданных экрана и системного аудиоряда посредством специализированного программного обеспечения на стороне учащегося [7].

2. ОРГАНИЗАЦИЯ АУДИОВИЗУАЛЬНОЙ ЗАПИСИ HIGH-TOUCH ЗАНЯТИЙ В СИСТЕМЕ ОЧНОГО ОБУЧЕНИЯ

Современная информационная среда большинства занятий в системе очного обучения формируется посредством экранной демонстрации наглядных материалов, комментируемых преподавателем посредством устных и графических пояснений (при наличии электронной доски или дополнительного инструментария, такого, как флипчарты и пр.). Автоматизированные средства создания видеолекций (Sonic Foundry Mediasite, Echo360 и др.), в реальном времени комбинировавшие изображение преподавателя с наглядными материалами и формировавшие медиапродукты для публикации в интернете, не нашли массового применения вследствие дороговизны приобретения и владения, а также по причине утраты совместимости с быстро эволюционирующей информационной средой (элиминация Flash-видео, несоответствие требованиям новых протоколов безопасности обмена данными и др.).

Современное изобилие и технический уровень аудиовизуальной аппаратуры делают возможным организацию аудиовизуальной записи очных занятий каждым преподавателем, который при этом становится продюсером и приобретает через три года имущественные права на медиапродукты, создаваемые им в процессе служебной деятельности. Для создания дидактически привлекательных аудиовизуальных материалов требуются аудиовизуальный захват экрана электронной доски и/или её видеозапись стендовой камерой, а также независимая запись изображения преподавателя крупным планом в сочетании со звукорядом с петличного Lavalier-микрофона. Аудиовизуальный захват контента электронной доски удобно осуществлять при помощи функционала записи слайдов Microsoft Power Point. Практический опыт показал, что звукоряд в экспортируемом аудиовизуальном ряде презентации Power Point, как правило, содержит высокочастотные помехи, индуцируемые в звуковой карте материнской платы компьютера близко расположенными электромагнитно неэкранированными микросхемами оперативной памяти. Таким образом, для качественной звукозаписи компьютер должен быть снабжен дискретной внутренней или внешней звуковой картой с электропитанием, отличным от источника питания компьютера. Альтернативный способ получения качественной звукозаписи голоса преподавателя – вывод звукоряда с петличного микрофона на диктофон или видеокамеру (при наличии соответствующего функционала). Еще один вариант – запись голоса лектора независимой видеокамерой с направленным микрофоном.

На основании практического опыта мы рекомендуем располагать камеру для видеозаписи крупного плана преподавателя (N1) слева от доски под углом 45 градусов и записывать на неё звукоряд посредством направленного микрофона. Стендовую камеру для записи содержимого экрана и дальнего плана с преподавателем (N2) целесообразно устанавливать в заднем левом углу аудитории и записывать на неё звукоряд с петличного микрофона посредством радиосистемы, работающей в диапазоне, не интерферирующем с Bluetooth. Для видеозаписи теперь можно использовать не только профессиональную аппаратуру, но и спортивные камеры с качеством записи 4К и коррекцией бочкообразных искажений широкоугольного объектива в реальном времени (чип Ambarella).

Видеоряд с презентацией (аспект кадра 4x3) при монтаже располагается в левой части окна видеоредактора (аспект кадра 16x9), а в правой части размещается изображение преподавателя, которое при монтаже можно варьировать в соответствии с необходимостью (крупный план анфас с камеры N1 и дальний план с камеры N2). Основной звукоряд при монтаже берется с камеры N2, а остальные используются для синхронизации видеорядов. Для облегчения синхронизации перед началом занятия целесообразно подавать сигнал «Мотор» в виде пары громких хлопков (специальной хлопушкой или ладонями). Опыт показал, что систематической использование сигнала начальной синхронизации положительно воспринимается слушателями и дополнительно дисциплинирует их.

3. ОРГАНИЗАЦИЯ ВИДЕОТРАНСЛЯЦИИ HIGH-TOUCH ЗАНЯТИЙ

Образовательная среда для дистанционных high-touch занятий может формироваться посредством организации видеоконференций в мини-группах при условии, что преподаватель и фасилитатор имеют возможность отслеживать реакцию учащихся, эмоционально комментировать преподаваемую информацию и оперативно корректировать траекторию обучения и модифицировать демонстрируемые медиаресурсы в соответствии с интересами аудитории. Создание благоприятной информационной среды для проведения учебных high-touch видеоконференций возможно благодаря особым методическим подходам к использованию high-tech инструментария [8].

В 2013 году специалистами ЮНЕСКО был разработан дидактический стандарт BYOD (Bring Your Own Device), учитывающий специфику поколения «Y», которое использует учебные материалы в кроссплатформенной среде личных мобильных устройств, таких, как ноутбуки, ультрабуки, нетбуки, планшеты, смартфоны, медиаплееры, микрокомпьютеры и пр. [9]. Благодаря практическому внедрению стандарта BYOD все современные устройства обладают базовой совокупностью аппаратных средств для участия в видеоконференциях (видеокамеры высокого разрешения, чувствительные микрофоны и акустические выходы для подключения наушников). Для организации high-touch формы обучения важно обеспечивать высокое качество аудиовизуальной информации со стороны преподавателя и фасилитатора. Используемые ими устройства должны

предоставлять возможность подключения внешней веб-камеры. Опыт показал, что внешняя аппаратура формирует аудиовизуальные ряды с 4K-качеством изображения и внятным звуком даже в 1–3 метрах от источника, что важно при демонстрации практических опытов, основанных на материальных технологиях.

При выборе веб-камеры для вещания целесообразно отдавать предпочтение устройствам с аппаратным кодированием потокового аудиовизуального ряда, что позволяет освободить ресурсы инфосреды преподавателя, интенсивно расходуемые в процессе high-touch занятия. К примеру, отзывчивость динамических медиаресурсов заметно улучшается при замене веб-камеры Logitech C920 HD Pro с программной обработкой медиаданных на устройство Logitech BRIO с аппаратной адаптацией аудиовизуальных данных к потоковому вещанию [10].

Для поддержания психологической атмосферы постоянного внимания к учащимся внешнюю веб-камеру преподавателя целесообразно позиционировать на мини-штативе перед частью экрана, где располагается окно приложения видеоконференции, демонстрирующее преподавателя. Веб-камеру фасилитатора имеет смысл устанавливать подобно веб-камере преподавателя или непосредственно над стеком окон участников занятия.

Сетевое аудиовизуальное вещание можно организовать посредством мессенджеров с функциями видеотелефонии [11], функционала социальных сетей [12] или специализированных видеосервисов [13]. Аппаратные свойства всех современных персональных компьютеров, начиная с нижнего ценового диапазона (процессор с частотой от 1 ГГц, оперативная память от 2 Гб, постоянная память от 32 Гб), позволяют организовать сетевую видеотрансляцию базового уровня с небольшим количеством одновременно запущенных приложений. Для демонстрации динамических мультимедийных ресурсов (ассоциативные карты и т. п.), рисования в многослойном режиме, открытия множества окон интернет-браузеров, переключения между множеством-веб-камер и пр. рекомендуется использовать компьютер с 8 Гб оперативной памяти, не менее 64 Гб постоянной памяти и аппаратным графическим ускорителем. Минимальные требования к интернет-подключению: 4 Мбит/с на приём и передачу.

4. ИНФОПРОСТРАНСТВО СИНХРОННОГО ДИСТАНЦИОННОГО HIGH-TOUCH ЗАНЯТИЯ

Характерные представители поколения «Y» игнорируют форму и декларируемое содержание, обращая внимание на своевременно представляемое содержимое [8]. Во время синхронных high-touch занятий нецелесообразно использование high-tech способов иллюстрирования обсуждаемых сведений при помощи заранее сформированных презентаций с линейной структурой (Microsoft Power Point, Adobe Presenter), которая заметно ограничивает свободу преподавания. Сравнительный анализ востребованности видеозаписей лекций показывает, что классические лекции с иллюстрациями, создаваемыми преподавателем в реальном времени мелом на доске, интересуют сетевую аудиторию на порядок больше, чем комментируемые слайды PowerPoint. Замечено, что существенные для high-touch занятий личностные особенности преподавателя проявляются сильнее при нелинейной структуре изложения учебных материалов.

В непредсказуемых условиях high-touch формы обучения следует формировать инфопространство, позволяющее в реальном времени дополнять графическими комментариями первичные медиаресурсы (pdf-публикации, статичные и анимированные изображения, звук и видео), управление которыми удобно осуществлять при помощи динамических ассоциативных карт [14]. Демонстрируемые медиаресурсы удобно располагать в окнах левой половины экранного поля, правая часть которого занята приложением видеоконференционной связи с совокупностью окон участников занятия (см. рис. 1).

При необходимости последовательной демонстрации медиаресурсов в разных окнах удобнее использовать режим трансляции всего рабочего стола, нежели переключать трансляцию между окнами. Наши эксперименты показали, что оконный режим видеотрансляции не утилитарен, поскольку демонстрируемое участникам занятия изображение ограничивается площадью первично транслируемого окна и для остальных окон преподаватель лишается возможности использовать курсор в качестве аттрактора внимания, а в процессе переключения трансляции между окнами учащиеся успевают отвлечься от занятия [8]. Быстрое распределение окон между частями экрана удобно осуществлять при

помощи устройств, поддерживающих жесты, таких, как мультисенсорные панели ноутбуков и клавиатур (Logitech Wireless Touch Keyboard K400), а также некоторые «мыши» (Logitech MX Master и т. п.).

Следует отметить, что использование устаревших программных приложений (молекулярное моделирование HyperChem, редактор химических формул ChemWin и др.) может быть доступно только в 32-битной версии операционной системы Windows 10 или запрещено вследствие несовместимости с базовыми компонентами системы (векторный графический редактор CorelDraw 13).

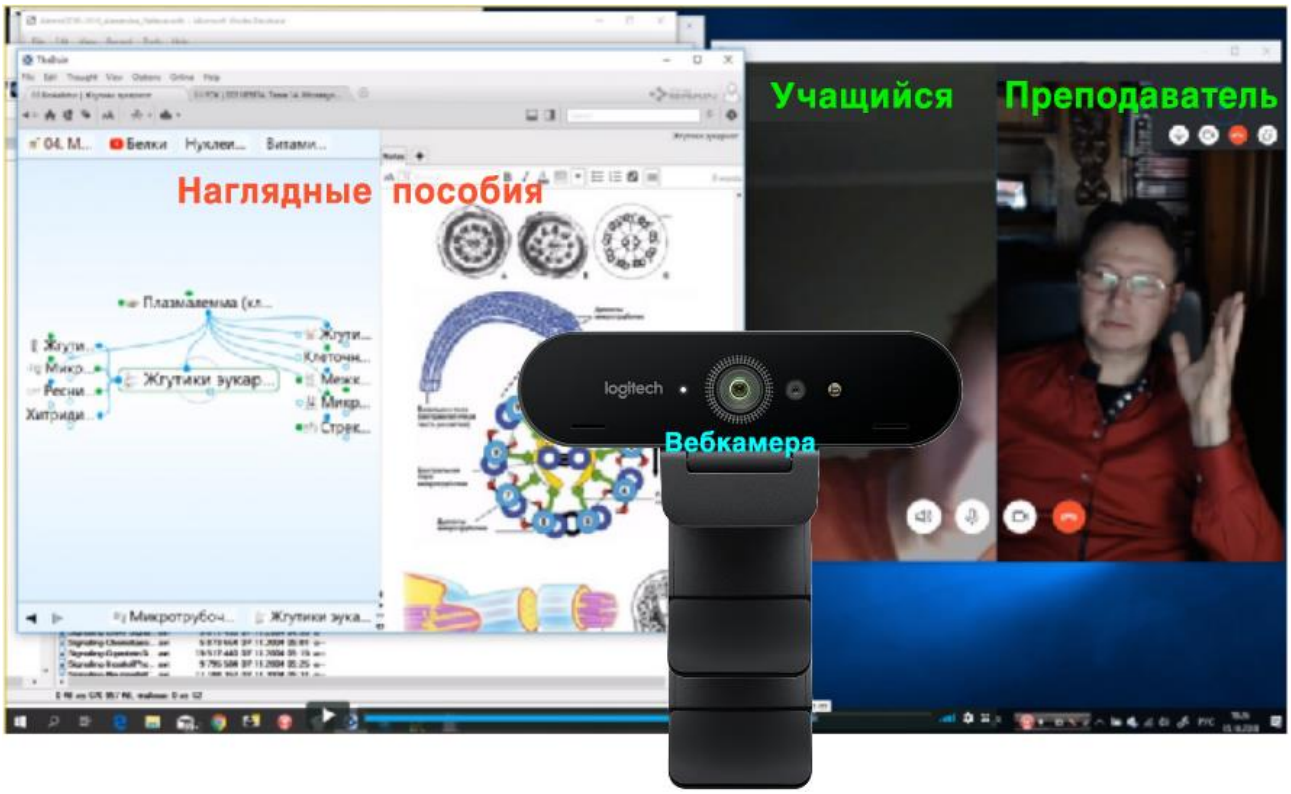


Рис. 1. Распределение экранного пространства во время high-touch занятия

При подготовке объектов, которые предполагается демонстрировать в режиме трансляции рабочего стола, необходимо учитывать, что изображение ведущего формируется в виде зеркального отражения.

5. ИНФОРМАЦИОННЫЙ ИНСТРУМЕНТАРИЙ СИНХРОННОГО ДИСТАНЦИОННОГО HIGH-TOUCH ЗАНЯТИЯ

Демонстрация первичных медиаресурсов (pdf-компиляции, графика, звук, видео) в среде Microsoft Windows 10 обеспечивается базовым набором приложений (браузер «Edge», «Фотографии», «Музыка Groove», «Кино и ТВ», соответственно).

Эффективный способ удержания внимания аудитории поколения «Y» на иллюстрациях преподаваемых сведений – рисование на доске, флипчарте, в поле кадра документ-камеры или посредством стилуса графического планшета («мыши») в совокупности с совместимыми векторными или растровыми редакторами изображений. До становления Windows 10 для рисования в процессе занятия использовались коммерческие программы Corel Painter, Corel DRAW, Adobe Photoshop, Adode Illustrator, ArtRage, Micrografx Picture Publisher и бесплатные MyPaint, GIMP, Artweaver, Inkscape). Современная альтернатива вышеперечисленным программам – бесплатное приложение Microsoft Whiteboard из Microsoft Store [15]. Векторные цифровые холсты свободного формата позволяют создавать профессионально выглядящие диаграммы и фигуры на неограниченной площади с интерфейсом, оптимизированным для работы с помощью жестов, пера и клавиатуры. Содержимое холстов в реальном времени сохраняется в облаке и синхронизируется на всех компьютерах одного аккаунта Microsoft. Особенно ценная возможность холстов Microsoft Whiteboard – присоединение всех участников занятия к одновременной работе. При необходимости содержимое холстов экспортируется в различные графические форматы.

Для дополнения существующих медиаресурсов поясняющими и надписями и изображениями ранее использовалось специфическое программное обеспечение графических планшетов (Wacom JustWrite Office и т. п.). Современный базовый набор программ операционной системы Windows 10 содержит приложение «Фрагмент и набросок» [16], позволяющее быстро добавлять векторные примечания к снимкам экрана, фотографиям и другим изображениям с помощью пера, сенсорной панели или мыши и сохранять, вставлять или отправлять их в другие приложения.

6. ПЕРВИЧНОЕ ПРЕПОДНЕСЕНИЕ ИНФОРМАЦИИ И КОНТРОЛЬ ЗНАНИЙ В ПРОЦЕССЕ СИНХРОННОГО ДИСТАНЦИОННОГО HIGH-TOUCH ЗАНЯТИЯ

Продуктивность high-touch занятий возрастает, если они проводятся по методике «flip lesson», когда базовые представления о предмете обсуждения учащиеся получают в ходе самоподготовки по материалам, предоставляемым в форме high-tech ресурсов [1]. Опыт показал, что набор из 20 вопросов, охватывающих все опорные темы 1,5-часового занятия, необходим и достаточен для контроля результатов самоподготовки и мотивации углубленного совместного рассмотрения дополнительных материалов. Контрольные вопросы целесообразно последовательно распределять между учащимися, присутствующими на занятии. Реакция учащихся на вопросы видна в реальном времени в окнах приложения для видеоконференций, любое из которых может быть максимизировано в нужный момент. При необходимости увеличение окна с видеорядом камеры преподавателя можно использовать для наглядной демонстрации моделей, привлечения внимания к мимике, жестикуляции и т. п.

7. ОРГАНИЗАЦИЯ АУДИОВИЗУАЛЬНОЙ ЗАПИСИ ИНФОПРОСТРАНСТВА СИНХРОННОГО ДИСТАНЦИОННОГО HIGH-TOUCH ЗАНЯТИЯ

В начале занятия к вещательному пространству подключается совокупность используемых преподавателем камер и/или включается демонстрация рабочего стола. При использовании специализированных видеосервисов или социальных сетей демонстрация рабочего стола преподавателя участникам занятия может оказаться невозможной, и в этом случае используется многокамерная конфигурация. Разработанные для потокового вещания веб-камеры HD и 4K-разрешения при использовании соответствующего программного обеспечения (ChromaCam от Personify) позволяют заменять фон в реальном времени на слайды презентаций или изображения с различными оптическими эффектами без использования физического хромакея. Существуют приложения (ManyCam и др.), позволяющие использовать во время прямого эфира множество веб-камер (замена плана, синхронная демонстрация множества видеорядов и т. п.).

Камеру, передающую фронтальное изображение лица преподавателя, рекомендуется размещать на середине высоты демонстрируемого экрана посред-

ством мини-штатива, установленного напротив вертикальной линии разделения левой и правой сторон экранного поля с наглядными (или контрольными) материалами и приложением видеоконференционной связи соответственно (см. рис. 1).

При организации вещания посредством функционала социальных сетей (Facebook и др.) предоставляется возможность во время подготовки к эфиру снабжать его необходимыми и дополнительными реквизитами первичного документа (автор и участники, название, аннотация, место и время создания, ссылки на дополнительные материалы). Опция сохранения и последующей публикации эфирных записей имеется в сервисах Facebook и YouTube. С 2017 года социальная сеть Facebook предоставляет возможность публикации записей прямых эфиров своих пользователей в форме мультискрипта, сохраняющего синхронную демонстрацию видеоряда, ленты комментариев зрителей с текстами и изображениями, а также реакций пользовательской аудитории в форме отображения движущихся эмодзи поверх видеоряда [12].

Аудиовизуальная запись занятий посредством специализированного программного обеспечения может осуществляться на преподавательском компьютере с достаточным запасом аппаратных ресурсов (частота процессора 2–3 МГц, не менее 8 Гб оперативной памяти, наличие графического сопроцессора с собственной оперативной памятью и система активного охлаждения). Из десятка протестированных приложений для видеозахвата экрана (Sketchman Studio Rylstim Screen Recorder, SourceForge CamStudio, Webinaria, Softronic Apowersoft Screen Recorder, Icecream Screen Recorder, FlashBack Blueberry screen recorder, Screencast-O-Matic Screen Recorder, Movavi Screen Capture Studio, Corel VideoStudio Pro Screen Capture, SolveigMultimedia HyperCam) наименее ресурсоёмким и наиболее надёжным оказался HyperCam [7], позволяющий дополнять изображение оверлейными надписями, озвучиванием различно анимируемых щелчков кнопок «мыши» и пр.

Утилитарным способом ведения аудиовизуальных записей синхронных занятий является бесплатный функционал мессенджера Skype, доступный с 2018 года [11]. После подключения всех участников к занятию можно активировать циркулярную демонстрацию экрана преподавательского компьютера, а затем включить запись, хронометраж которой отображается на фризе окна мессен-

джера инициатора записи. Остальные участники получают уведомления о начале записи. После окончания сеанса связи записи автоматически сохраняются и публикуются в лентах мессенджеров всех участников занятия в форме, доступной для онлайн-просмотра и скачивания в течение 30 дней (формат mp4). Главные преимущества этого вида видеозаписи – высокая надёжность и быстрота формирования файла с видеозаписью (несколько минут по окончании записи).

Удобная возможность видеозаписи презентации со звуковыми и графическими комментариями (указатель в режиме «лазерная указка», рисование поверх слайда пером или маркером) в сочетании с опциональным аудиовизуальным рядом вебкамеры (640x480) предоставляется современным приложением для презентаций Microsoft Power Point. При выборе опции «Запись слайдшоу» окно презентации переходит в режим отображения слайдов совместно с базовой палитрой и инструментами рисования. Отображение видеоряда вебкамеры во время презентации может быть отключено. После окончания записи она сохраняется в компактной форме в файле презентации. Записанные голосовые и графические комментарии в сочетании с синхронизированным аудиовизуальным рядом вебкамеры могут воспроизводиться непосредственно из презентации или экспортироваться в видеофайл формата mp4 с разными опциями качества (Ultra HD, Full HD, HD и SD). Перед экспортом видеоряд вебкамеры можно переместить в оптимальное место экрана или отключить. С учётом последующего монтажа целесообразно использовать презентацию с соотношением сторон кадра 4x3, которая может быть позиционирована с левого края монтажного стола видеоредактора, использующего соотношение сторон кадра 16x9, при этом остающееся справа пространство используется для размещения видеозаписи преподавателя. Недостаток записи занятий посредством Power Point – медленный экспорт аудиовизуальных рядов.

8. ПУБЛИКАЦИЯ ВИДЕОЗАПИСЕЙ И НАГЛЯДНЫХ МАТЕРИАЛОВ ЗАНЯТИЯ

Видеозаписи занятий, создаваемые посредством специализированных сервисов [13], функционала социальных сетей [12] и приложений для презентаций, после окончания занятия конвертируются в пригодный для сетевой публикации формат (mp4), адаптирующийся к устройствам пользователей в соответствии со стандартом BYOD [4, 9]. Время конвертирования аудиовизуальных за-

писей определяется планом подписки на сервис и может превышать длительность записи в несколько раз. Конвертирование 1.5-часовых записей вебинаров в мессенджере Skype происходит за 3 минуты, после чего аудиовизуальные ряды в формате mp4 становятся доступны для просмотра и скачивания. Экспорт 1.5-часовых презентаций Power Point происходит несколько часов.

При перерывах в занятии формируется несколько аудиовизуальных рядов, которые можно быстро объединить без перекодирования в приложении HyperCam Media Editor, входящем в комплект программы видеозахвата HyperCam 5 [7].

Аудиовизуальные записи, созданные на локальном компьютере или скачанные из мессенджера Skype, удобно публиковать в альбомах (ShowCase) видеохостинга Vimeo (см. рис. 2), предоставляющего расширенные возможности для дидактически целенаправленного использования в открытом и закрытом доступах [17].

Изображения экспортированных по окончании занятия холстов Miscosoft Whiteboard и сохраненных набросков на фрагментах экрана целесообразно сосредоточивать в сетевых альбомах сайтов, адаптированных для учебной работы по стандартам high-tech методик дистанционного обучения [18]. Альбомы видеозаписей Vimeo могут использоваться непосредственно или публиковаться на учебных сайтах [19].

При формировании набора медиаресурсов для самостоятельной работы учащихся следует учитывать, что информационная среда большинства устройств Apple не поддерживает воспроизведение Flash-видео и swf-анимаций, которые целесообразно заблаговременно замещать на аналоги, использующие технологии HTML-5.

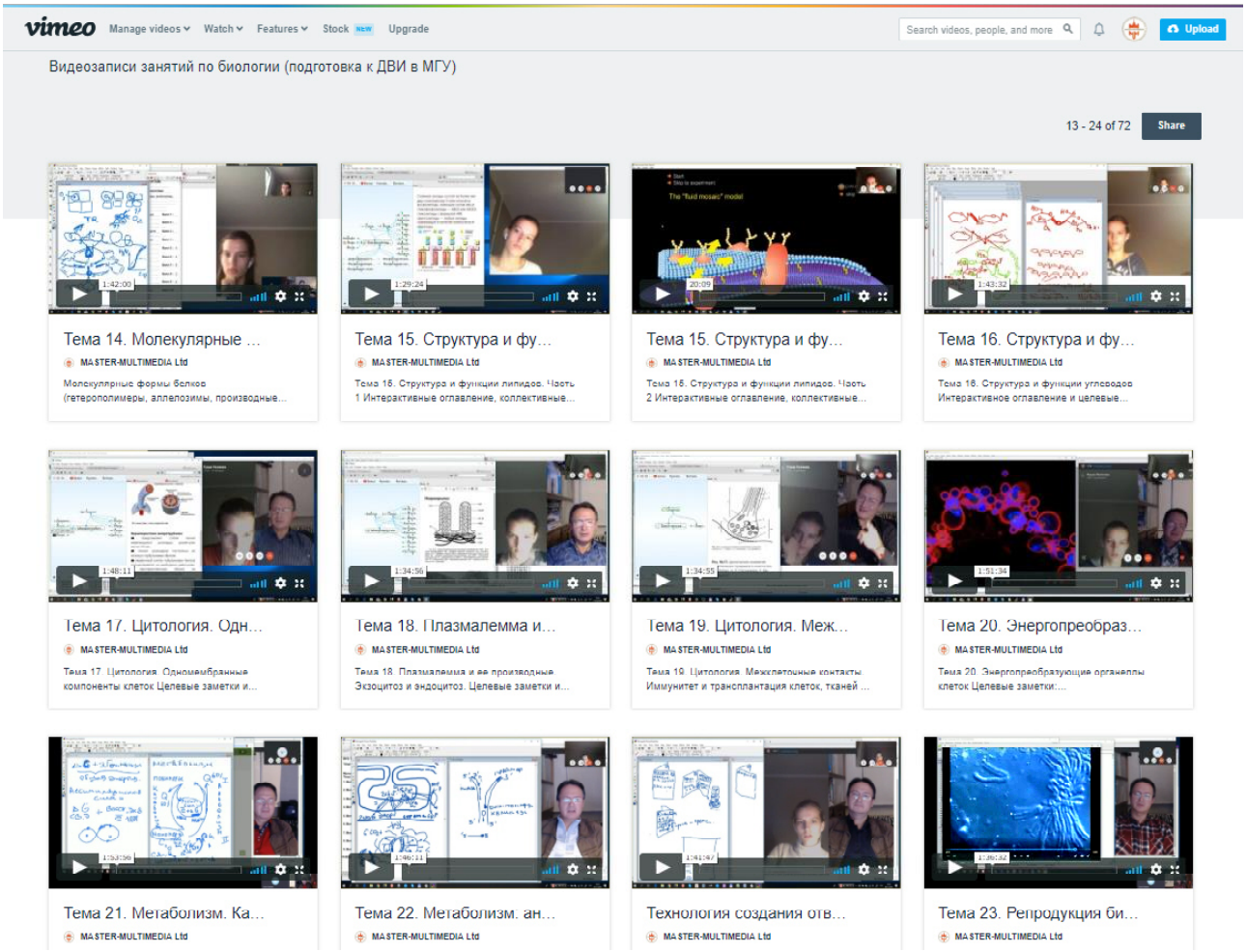


Рис. 2. Онлайн-альбом видеозаписей high-touch запнятий на хостинге Vimeo

9. ОРГАНИЗАЦИЯ ПРОЦЕССОВ ПОВТОРЕНИЯ И ЗАКРЕПЛЕНИЯ МАТЕРИАЛОВ ЗАПИСАННОГО ЗАНЯТИЯ

Аудиовизуальные записи, опубликованные на видеосервисе Vimeo, могут быть использованы в специальном режиме, позволяющем создавать целевые заметки, акцентирующие внимание на требуемом месте любого кадра видеоряда посредством размещения интерактивной метки [17]. На хронометражной полосе кадры с метками обозначаются светлыми вертикальными полосками, что облегчает навигацию (см. рис. 3). К целевым заметкам могут быть прикреплены гиперссылки на соответствующие изображения, созданные в ходе занятия и опубликованные в альбомах учебных сайтов. Во время создания метки формируется соответствующий пункт интерактивного оглавления аудиовизуального ряда, нажатие на который впоследствии обеспечивает переход к нужному кадру и визуализацию маркера. Длина имени и количество пунктов такого оглавления

не ограничены. Пользователи имеют возможность комментировать текст, сопровождающий метку, что создаёт замечательную возможность для коллективного обсуждения помеченного содержимого видеоряда в рамках научно-образовательной, деловой, общественной и любой другой деятельности. Всем участникам обсуждения по электронной почте рассылаются уведомления о появлении новых заметок и пользовательских реакциях на них.

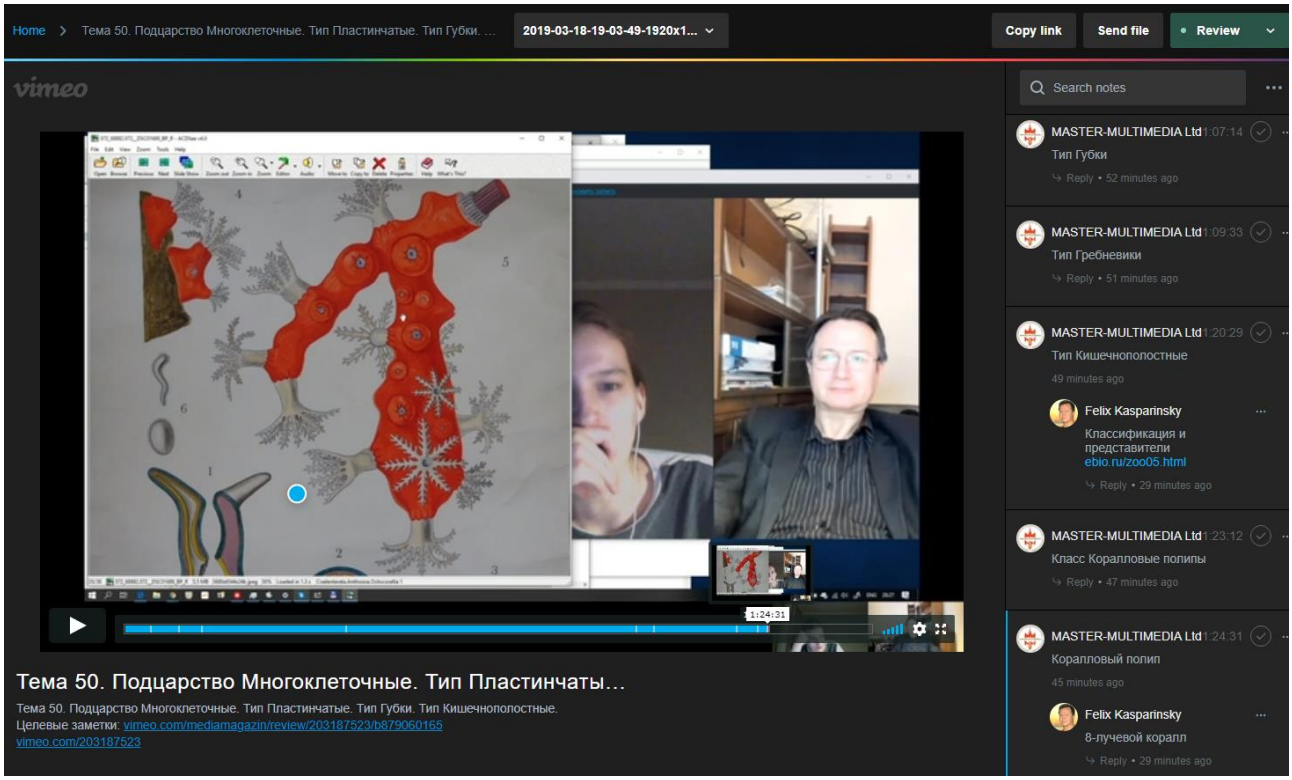


Рис. 3. Целевые заметки видеозаписей дистанционных занятий на хостинге Vimeo

В живой ленте учебной группы high-tech ресурса дистанционного обучения могут публиковаться опросы и комментарии, организовываться обсуждения дидактически существенных кадров аудиовизуальных записей high-touch занятий, прямые ссылки на которые формируются функционалом целевых заметок.

10. СИСТЕМАТИЗАЦИЯ И ОБОБЩЕНИЕ УЧЕБНЫХ МАТЕРИАЛОВ С ИСПОЛЬЗОВАНИЕМ АУДИОВИЗУАЛЬНОЙ ЗАПИСИ ДИСТАНЦИОННОГО ЗАНЯТИЯ

Применяемый видеохостингом Vimeo сервис целевых коллективных заметок [17] позволяет посредством гиперссылок перейти к любому кадру видеозаписи и обратить внимание на его определённое место. Целевые заметки обес-

печивают возможность систематизации содержимого каждой видеозаписи курса занятий и объединения фрагментов разных видеозаписей с общей тематикой.

ЗАКЛЮЧЕНИЕ

Высокое качество аудиовизуальных материалов, получаемых при помощи современных общедоступных аппаратных и программных средств, в сочетании с интерактивным функционалом сетевых сервисов создает предпосылки для возникновения нового тренда: повышения качества образования посредством всеобщего дидактически целенаправленного создания и использования видеозаписей синхронных занятий в процессе очного и дистанционного обучения.

СПИСОК ЛИТЕРАТУРЫ

1. *Крашенинникова Л.В.* Сочетание high-tech и high-touch подходов как способ достижения конкурентного преимущества в дистанционном образовании // Качество дистанционного образования: концепции, проблемы, решения (DEQ-2013). Материалы XV Международной научно-практической конференции (6 декабря 2013 г., Москва). М.: МГИУ, 2013. С. 98–100.

2. *Каспаринский Ф.О., Полянская Е.И.* Инфоцентризм как дидактическая стратегия // Вестник Международного института менеджмента ЛИНК (5). М.: МИМ ЛИНК, 2014. С. 65–73.

3. Методические рекомендации по организации обучения на дому детей-инвалидов с использованием дистанционных образовательных технологий // Министерство образования и науки Российской Федерации, Департамент государственной политики в сфере защиты прав детей. Письмо от 10 декабря 2012 г. № 07-832. URL: https://usperm.ru/content/pismo-minobrnauki-rossii-ot-10122012-no-07-832

4. *Каспаринский Ф.О., Полянская Е.И.* Аудиовизуальные ресурсы для мобильного дистанционного обучения // Формирование системы независимой оценки квалификации и качество дистанционного образования: концепции, проблемы, решения (DEQ-2014). Материалы Всероссийской конференции. Жуковский: МИМ ЛИНК, 2014. С. 46–49.

5. Adobe web conferencing software | Adobe Connect. URL: https://www.adobe.com/products/adobeconnect.html

6. Платформа Webinar нового поколения для онлайн-мероприятий. Запускается на всех браузерах и без дополнительного ПО. https://webinar.ru

7. HyperCam 5.0 – Удобная запись экрана, игр, фильмов. URL: http://www.solveigmm.com/ru/products/hypercam/

8. *Каспаринский Ф.О., Полянская Е.И.* Организация high-touch формы дистанционного обучения посредством Skype-видеоконференций // Качество дистанционного образования: концепции, проблемы, решения (DEQ-2015). Материалы Международной конференции 11 декабря 2015 г. Жуковский: АНО ВО «Международный институт менеджмента ЛИНК», 2016. С. 42–45.

9. *Каспаринский Ф.О.* Публикация интернет-ресурсов дистанционного обучения в соответствии со стандартом BYOD // Качество открытого дистанционного образования: концепции, проблемы, решения (DEQ-2017). Молодежь и наука. Материалы XIX международной научно-практической конференции. Жуковский: Международный институт менеджмента ЛИНК, 2018. С. 89–94.

10. Веб-камеры для видеоконференций и видеосвязи // Logitech.com. URL: https://www.logitech.com/ru-ru/video/webcams

11. Skype — общение без ограничений. Звоните, переписывайтесь, делитесь любыми файлами — и все это бесплатно // Microsoft. URL: https://www.skype. com/ru/

12. *Каспаринский Ф.О., Полянская Е.И.* Вариативность инструментов публикации медиаресурсов в социальных сетях // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2017. С. 218–226. doi:10.20948/abrau-2017-28

13. Live streaming, without limits. The home for high-quality live streaming and video hosting. // Vimeo. URL: https://vimeo.com/features/livestreaming

14. *Каспаринский Ф.О.* Представление наглядных материалов учащимся поколения Сети посредством динамических ассоциативных карт // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2017. С. 207–217. doi:10.20948/abrau-2017-27

15. Microsoft Whiteboard // Microsoft Store. URL: https://www.microsoft.com/ store/ productId/9MSPC6MP8FM4

16. Фрагмент и набросок // Microsoft Store. URL: https://www.microsoft.com/store/productId/9MZ95KL8MR0L

17. *Каспаринский Ф.О., Полянская Е.И.* Информационно-навигационный сервис сетевых аудиовизуальных ресурсов // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 284–294. URL: http://keldysh.ru/abrau/2018/theses/42.pdf doi:10.20948/abrau-2018-42

18. *Каспаринский Ф.О., Полянская Е.И.* Адаптация ресурсов дистанционного обучения к компетентностному формату // Открытое образование. Научно-практический журнал. М: Московский государственный университет экономики, статистики и информатики. 2014. №4. С. 11–19.

19. *Каспаринский Ф.О., Полянская Е.И.* Организация структурированных образовательных видеотек под управлением CMS 1C-Bitrix // Качество дистанционного образования: концепции, проблемы, решения (DEQ-2012). Материалы XIV Международной научно-практической конференции 7 декабря 2012 г. М.: МГИУ, 2012. С. 71–74.

AUDIOVISUAL RECORDING OF SYNCHRONOUS LESSONS DURING FULL-TIME AND DISTANCE LEARNING

F. O. Kasparinsky

MASTER-MULTIMEDIA LLC, Moscow

felix@kasparinsky.pro

Abstract

The modern information environment provides unprecedented opportunities for combining high-tech and high-touch learning approaches. It can be expected that in the near future, the general trend will be the use of audio-visual recordings of synchronized classes, which should be used for subsequent consolidation, repetition, control, generalization and systematization of knowledge. The article summarizes the results of 10 years of experience in creating and using audio-visual recordings of full-time and distance learning in university and school classrooms.

Keywords*: audiovisual recording, distance learning, full-time learning, Internet, Skype, Vimeo, high-touch, high-tech, synchronous lessons*

REFERENCES

1. *Krasheninnikova L.V.* Sochetanie high-tech i high-touch podkhodov kak sposob dostizheniia konkurentnogo preimushchestva v distantsionnom obrazovanii // Kachestvo distantsionnogo obrazovaniia: kontseptsii, problemy, resheniia (DEQ-2013). Materialy XV Mezhdunarodnoi nauchno-prakticheskoi konferentsii (6 dekabria 2013 g., Moskva). M.: MGIU, 2013. S. 98–100.

2. *Kasparinskii F.O., Polianskaia E.I.* Infotsentrizm kak didakticheskaia strategiia // Vestnik Mezhdunarodnogo instituta menedzhmenta LINK (5). M.: MIM LINK, 2014. S. 65–73.

3. Metodicheskie rekomendatsii po organizatsii obucheniia na domu detei-invalidov s ispolzovaniem distantsionnykh obrazovatelnykh tekhnologii // Ministerstvo obrazovaniia i nauki Rossiiskoi Federatsii, Departament gosudarstvennoi politiki v sfere zashchity prav detei. Pismo ot 10 dekabria 2012 g. No 07-832. URL: https://usperm.ru/content/pismo-minobrnauki-rossii-ot-10122012-no-07-832

4. *Kasparinskii F.O., Polianskaia E.I.* Audiovizualnye resursy dlia mobilnogo distantsionnogo obucheniia // Formirovanie sistemy nezavisimoi otsenki kvalifikatsii i kachestvo distantsionnogo obrazovaniia: kontseptsii, problemy, resheniia (DEQ-2014). Materialy Vserossiiskoi konferentsii. Zhukovskii: MIM LINK, 2014. S. 46–49.

5. Adobe web conferencing software | Adobe Connect. URL: https://www.adobe.com/products/adobeconnect.html

6. Platforma Webinar novogo pokoleniia dlia onlain-meropriiatii. Zapuskaetsia na vsekh brauzerakh i bez dopolnitelnogo PO. https://webinar.ru

7. HyperCam 5.0 – Udobnaia zapis ekrana, igr, filmov. URL: http://www.solveigmm. com/ru/products/hypercam/

8. *Kasparinskii F.O., Polianskaia E.I.* Organizatsiia high-touch formy distantsionnogo obucheniia posredstvom Skype-videokonferentsii // Kachestvo distantsionnogo obrazovaniia: kontseptsii, problemy, resheniia (DEQ-2015). Materialy Mezhdunarodnoi konferentsii 11 dekabria 2015 g. Zhukovskii: ANO VO «Mezhdunarodnyi institut menedzhmenta LINK», 2016. S. 42–45.

9. *Kasparinskii F.O.* Publikatsiia internet-resursov distantsionnogo obucheniia v sootvetstvii so standartom BYOD // Kachestvo otkrytogo distantsionnogo obrazovaniia: kontseptsii, problemy, resheniia (DEQ-2017). Molodezh i nauka. Materialy XIX mezhdunarodnoi nauchno-prakticheskoi konferentsii. Zhukovskii: Mezhdunarodnyi institut menedzhmenta LINK, 2018. S. 89–94.

10. Veb-kamery dlia videokonferentsii i videosviazi // Logitech.com. URL: https://www.logitech.com/ru-ru/video/webcams

11. Skype — obshchenie bez ogranichenii. Zvonite, perepisyvaites, delites liubymi failami — i vse eto besplatno. // Microsoft. URL: https://www.skype.com/ru/

12. *Kasparinskii F.O., Polianskaia E.I.* Variativnost instrumentov publikatsii mediaresursov v sotsialnykh setiakh // Nauchnyi servis v seti Internet: trudy XIX Vserossiiskoi nauchnoi konferentsii (18–23 sentiabria 2017 g., g. Novorossiisk). M.: IPM im. M.V. Keldysha, 2017. S. 218–226. doi:10.20948/abrau-2017-28

13. Live streaming, without limits. The home for high-quality live streaming and video hosting // Vimeo. URL: https://vimeo.com/features/livestreaming

14. *Kasparinskii F.O.* Predstavlenie nagliadnykh materialov uchashchimsia pokoleniia Seti posredstvom dinamicheskikh assotsiativnykh kart // Nauchnyi servis v

seti Internet: trudy XIX Vserossiiskoi nauchnoi konferentsii (18–23 sentiabria 2017 g., g. Novorossiisk). M.: IPM im. M.V. Keldysha, 2017. S. 207–217. doi:10.20948/abrau-2017-27 .

15. Microsoft Whiteboard // Microsoft Store. URL: https://www.microsoft.com/store/ productId/9MSPC6MP8FM4

16. Fragment i nabrosok // Microsoft Store. URL: https://www.microsoft. com/ store/ productId/9MZ95KL8MR0L

17. *Kasparinskii F.O., Polianskaia E.I.* Informatsionno-navigatsionnyi servis setevykh audiovizualnykh resursov // Nauchnyi servis v seti Internet: trudy XX Vserossiiskoi nauchnoi konferentsii (17–22 sentiabria 2018 g., g. Novorossiisk). M.: IPM im. M.V. Keldysha, 2018. S. 284–294. URL: http://keldysh.ru/abrau/2018/theses/ 42.pdf doi:10.20948/abrau-2018-42

18. *Kasparinskii F.O., Polianskaia E.I.* Adaptatsiia resursov distantsionnogo obucheniia k kompetentnostnomu formatu // Otkrytoe obrazovanie. Nauchno-prakticheskii zhurnal. M: Moskovskii gosudarstvennyi universitet ekonomiki, statistiki i informatiki. 2014. No 4. S. 11–19.

19. *Kasparinskii F.O., Polianskaia E.I.* Organizatsiia strukturirovannykh obrazovatelnykh videotek pod upravleniem CMS 1C-Bitrix // Kachestvo distantsionnogo obrazovaniia: kontseptsii, problemy, resheniia (DEQ-2012). Materialy XIV Mezhdunarodnoi nauchno-prakticheskoi konferentsii 7 dekabria 2012 g. M.: MGIU, 2012. S. 71–74.

СВЕДЕНИЯ ОБ АВТОРЕ

КАСПАРИНСКИЙ Феликс Освальдович – кандидат биологических наук, основатель и научный руководитель Лаборатории мультимедийных технологий Биологического факультета МГУ имени М.В. Ломоносова, учредитель и Генеральный директор ООО «МАСТЕР-МУЛЬТИМЕДИА» Сфера научных интересов – формирование информационной среды, дидактически целенаправленное использование мультимедийных технологий.

Felix Osvaldovich KASPARINSKY – Founder and Scientific Director of Multimedia Technologies Laboratory (Biological Faculty, M.V. Lomonosov Moscow State University), Founder and General Director of MASTER-MULTIMEDIA LLC. Research interests include creating an information environment and didactically targeted use of multimedia technologies.

email: felix@kasparinsky.pro

Материал поступил в редакцию 15 ноября 2019 года

УДК 004.432

ОПРЕДЕЛЕНИЕ ЗАВИСИМОСТЕЙ ПО ДАННЫМ СРЕДСТВАМИ ДИНАМИЧЕСКОГО АНАЛИЗА СИСТЕМЫ SAPFOR

Н. А. Катаев [1], А. А. Смирнов [2], А. Д. Жуков [3]

[1, 2]*Институт прикладной математики им. М.В. Келдыша РАН, г. Москва;*
[3]*Московский государственный университет им. М.В. Ломоносова, г. Москва*

[1]kataev_nik@mail.ru, [2] smiraland@gmail.com, [3]andreyzkk@yandex.ru

Аннотация

Использование указателей и косвенной адресации в программе, а также сложная структура графа потока управления являются одними из основных препятствий при выполнении статического анализа программ. Обнаруженные в результате такого анализа свойства программы слишком консервативно описывают ее поведение и часто оказываются недостаточными для принятия решений о возможности ее параллельного выполнения. Использование динамического анализа программ позволяет расширить возможности средств автоматизации распараллеливания. В системе SAPFOR (System FOR Automated Parallelization) реализован инструмент динамического анализа, опирающийся на инструментацию программ в представлении LLVM, что позволяет исследовать программы на языках C и Fortran. Чтобы снизить накладные расходы на время выполнения инструментированной программы, сохранив при этом полноту проводимого анализа, используются возможности статического анализа, реализованного в SAPFOR. В процессе динамического анализа часть обращений к памяти, информация о которых была получена в процессе статического анализа, может быть проигнорирована. Разработанный инструмент был протестирован на тестах производительности из пакета NAS Parallel Benchmarks для языков C и Fortran. В процессе динамического анализа кроме традиционных видов зависимостей (flow, anit, output) также определяются переменные, зависимость по которым может быть устранена за счет приватизации или конвейерного выполнения циклов. Совместно с возможностями DVM и OpenMP это существенно облегчает, в том числе, и ручное распараллеливание, облегчая задание соответствующих директив компилятора.

***Ключевые слова**: анализ программ, динамический анализ, автоматизация распараллеливания, SAPFOR, DVM, LLVM*

ВВЕДЕНИЕ

Целью разработки системы автоматизированного распараллеливания SAPFOR [1, 2] (System FOR Automated Parallelization) является снижение сложности разработки параллельных программ. Разностороннее развитие архитектурных и программных составляющих современных вычислительных систем приводит к необходимости совместного использования различных технологий параллельного программирования (MPI, SHMEM, OpenMP, CUDA, OpenACC, OpenCL). Система SAPFOR использует в качестве целевого языка программирования DVM-языки (Fortran-DVMH и C-DVMH), входящие в состав DVM-системы [3, 4]. Данные языки инкапсулируют специфичные для различных программных систем особенности, облегчая как ручную, так и автоматизированную разработку параллельных программ. Разработка системы SAPFOR связана с проведением исследований в трех основных направлениях: анализ программ, автоматическое распараллеливание «хорошо» написанных потенциально параллельных программ и приведение последовательных программ к потенциально параллельному виду. Каждое из этих направлений, как по отдельности, так и в целом, призвано оказать помощь в разработке параллельных программ. Все три упомянутых направления исследований опираются на исследования свойств распараллеливаемой программы. Таким образом, анализ программ является неотъемлемой составляющей любого процесса распараллеливания, не обязательно автоматического.

SAPFOR обеспечивает как статический, так и динамический анализ программ. По отдельности каждого вида анализа оказывается недостаточно. С одной стороны, статический анализ во многих случаях оказывается слишком консервативным, указывая на наличие реально отсутствующих зависимостей с целью сохранения корректности программы. Авторы работы [5], исследуя возможности автоматического распараллеливания в компиляторах Intel и PGI на примере набора модельных приложений для научных расчетов OmpSRC [6], выделяют как минимум две основные проблемы, препятствующие статическому анализу: использование указателей и сложный граф потока управления. В обоих случаях компиляторы вынуждены принимать консервативные решения о наличии зави-

симостей в программе. С другой стороны, динамический анализ, во-первых, анализирует программу для определенного набора входных данных, во-вторых, является ресурсоемким с точки зрения потребляемых времени и памяти.

Кроме факта наличия или отсутствия зависимости в программе, важно понимать возможные способы ее устранения. Например, обратные зависимости и зависимости по выходу могут быть устранены за счет приватизации данных (т. е. создания локальной копии данных для каждого процесса/нити). Соответствующий статический анализ скалярных переменных, основанный на анализе потока данных, реализован в SAPFOR [7]. Но для приватизации массивов необходимо определять множества элементов, используемых на каждой итерации цикла. Адресная арифметика, косвенная индексация, зависимость индексных выражений от параметров функций и переменных, вычисляемых в процессе выполнения программы, приводят к невозможности статического определения таких массивов. При этом устранение такого рода зависимостей является одним из ключевых преобразований, требуемых для распараллеливания тестов производительности из набора NAS Parallel Benchmarks [8, 9]. Другим источником параллелизма являются прямые зависимости по данным, расстояние которых ограничено константой. Циклы с такими зависимостями допускают конвейерное выполнение. Соответствующие директивы ACCROSS предусмотрены в DVM-системе, и их задание требует указания предварительно вычисленного расстояния зависимости. Чтобы справиться с указанными трудностями, в системе SAPFOR был реализован динамический анализ зависимостей по данным.

1. СУЩЕСТВУЮЩИЕ ПОДХОДЫ К ДИНАМИЧЕСКОМУ АНАЛИЗУ

Основными недостатками динамического анализа являются зависимость от полноты входных данных и большие накладные расходы. Степень представительности входных данных существенно влияет на достоверность динамического анализа, использование плохо подобранных наборов данных может приводить к пропуску существующих зависимостей. В этом смысле динамический анализ возможен только под пристальным контролем пользователя, занимающегося распараллеливанием программы, и удобен для использования в системах автоматизации распараллеливания, таких, как SAPFOR.

Для получения информации о программе в процессе динамического анализа широко применяются два следующих подхода: семплирование, то есть получение профиля выполнения программы через заранее заданные промежутки времени, и использование инструментации (вставки в код программы вызовов некоторой внешней библиотеки) для исследования поведения программы в заранее заданных точках. Семплирование значительно менее требовательно к потребляемым ресурсам, но страдает от вероятности потери информации, поэтому для анализа зависимостей по данным оно малоприменимо.

Для определения зависимостей по данным наиболее подходящим подходом является использование инструментации. В свою очередь, вставлять вызовы функций динамического анализа можно на уровне исходного кода, на уровне некоторого внутреннего представления или, выполняя бинарную инструментацию. Инструментация программ на уровне исходного кода требует отдельной реализации инструментов для каждого поддерживаемого языка программирования (в случае SAPFOR это языки Fortran и C). Приходится отдельно обрабатывать все возможные синтаксические конструкции поддерживаемых языков. Кроме того, с целью сокращения накладных расходов в определенных случаях программа может быть предварительно модифицирована и инструментирована не полностью. Такого рода преобразования также удобнее выполнять над некоторым единым для разных языков представлением программы (например, LLVM IR). Бинарная инструментация наиболее эффективна с точки зрения полноты покрытия исполняемой программы, так как допускает инструментацию даже в случае отсутствия исходных кодов, например, позволяя анализировать уже скомпилированные модули, для которых применение других видов инструментации невозможно. Недостатком является то, что возможность соотнесения полученной информации с исходным кодом анализируемой программы сильно зависит от полноты доступной отладочной информации.

В системе SAPFOR в качестве внутреннего представления для анализа программ (как статического, так и динамического) используется LLVM [10]. LLVM IR является универсальным для различных языков программирования, поддерживает возможность выполнения дополнительных анализов и преобразований с целью выборочной инструментации программы, многие из которых уже реализованы в LLVM. Отладочная информация, доступная в LLVM IR, может быть дополнительно

расширена с целью более полного описания исходной программы. С этой целью в SAPFOR обеспечивается соответствие между определенными конструкциями исходной программы, представленными в виде синтаксического дерева (AST) и конструкциями LLVM IR (циклы, переменные, функции и их вызовы). LLVM IR одновременно существует в трех видах: структура классов языка C++ 11, бинарное и текстовое представления. Доступность понятного текстового представления существенно повышает удобство отладки проводимых преобразований и является еще одним преимуществом перед использованием бинарной инструментации. Недостатком является то, что заранее скомпилированные участки программы не могут быть инструментированы и должны подвергаться консервативному анализу.

Одним из наиболее широко применяемых инструментов динамического анализа зависимостей по данным является Intel Advisor, входящий в состав Intel Parallel Studio [11]. Данный инструмент позволяет определять три типа зависимостей по данным: прямые (flow, RAW), обратные (anti, WAR), по выходу (output, WAW). Более детальной классификации, например, с целью выделения переменных которые могут быть объявлены приватными, а также определения расстояний зависимостей, не выполняется. Для анализа используется бинарная инструментация, поэтому с целью корректного соотнесения полученных результатов с исходным кодом рекомендуется отключение оптимизаций. Чтобы запустить анализ зависимостей по данным, нужно предварительно получить профиль выполнения программы: в нем можно будет выбрать циклы, которые должны быть проанализированы. Для получения профиля выполняется семплирование, поэтому количество обнаруженных циклов зависит от размера шага, который используется для сбора данных. Но даже для минимального шага не все циклы будут идентифицированы, например, для программы LU (класс S) при минимальном шаге будет обнаружена только треть циклов (порядка 60 из 187 циклов программы). Это связано в первую очередь с тем, что данные класса S – это тестовые данные очень маленького размера, на которых программа выполняется доли секунды. При этом указанный набор данных достаточно полно описывает поведение программы в целом.

В Таблице 1 приведены замедление Intel Advisor для программы LU (класс S), а также количество циклов, доступных для анализа при минимальном шаге семплирования. Замедление указывается в виде отношения времен выполнения для

программ собранных с разными оптимизационными опциями (-O0 и -O3) с включенным и отключенным режимом анализа зависимостей по данным. Стоит отметить, что рекомендованные режимы анализа предполагают использование опции -O0, чтобы обеспечить анализ исходной программы без оптимизаций и обеспечить точное соответствие выдаваемой информации объектам исходного кода. В этом случае замедление составляет порядка 4160 раз при том, что будет проанализирована только треть всех циклов программы.

Тестирование выполнялось на Intel Xeon CPU E5-1660 v2, 3.70 GHz, с отключенным Turbo Boost. Для компиляции и анализа программ использовалась Intel Parallel Studio 2019 с опорой на системные библиотеки GCC 7.4.

Таблица 1. Замедление Intel Advisor 2019 при анализе теста LU (класс S)

Опции	**-O0 (дин. анализ)**	**-O3 (дин. анализ)**
-O0 (без анализа)	850 раз / 62 цикла	433 раз / 27 циклов
-O3 (без анализа)	**4160 раз / 62 цикла**	2384 раз / 27 цикл

Идея реализованного в SAPFOR алгоритма динамического анализа похожа на алгоритм попарного сравнения (pairwise method) обращений к памяти, описанный в работе [12]. Авторами были предложены улучшения данного алгоритма, позволяющие сократить накладные расходы, но нам не удалось получить доступ к разработанному ими инструменту. При его реализации авторы опираются на бинарную инструментацию, хотя и говорят о возможности использования LLVM. Кроме того, динамический анализ выполнялся не над всеми циклами программы, а только над наиболее ресурсоемкими. Поэтому оценить реальные издержки при анализе всей программы довольно сложно (общее количество циклов в программах не приводится). Авторы статьи выполняли анализ с целью дальнейшего распараллеливания для систем с общей памятью, и в этом случае выборочный анализ циклов оказывается оправданным. В SAPFOR необходимо принимать глобальные решения о распределении данных и вычислений, следовательно, нельзя исключать циклы из анализа, так как они могут обращаться к распределяемым данным (в том числе, косвенно, то есть средствами статического анализа отследить данные ситуации может быть невозможно).

2. АЛГОРИТМ ДИНАМИЧЕСКОГО АНАЛИЗА

В рамках системы SAPFOR требуется, чтобы динамический анализ обнаруживал для каждого цикла программы тип зависимости по данным и возможность ее устранения. Для определения возможности конвейерного выполнения цикла требуется знать расстояние зависимости (максимальное и минимальное). Также необходимо определять ситуации, когда зависимость может быть устранена за счет приватизации данных. В этом случае кроме локального анализа тела цикла необходимо убедиться, что значение переменной, вычисленное в цикле, не используется после выхода из него.

Так как не требуется определять конкретные операторы, порождающие зависимости по данным, можно не хранить список всех обращений к памяти, а только сам факт обращения к данной ячейке памяти и некоторую дополнительную информацию, необходимую для выявления желаемых свойств. При этом в момент обращения к ячейке памяти приходится выполнять элементарную обработку накопленных данных об этой ячейке. Такой подход позволяет снизить требования к памяти в случае, когда на каждой итерации цикла происходят множественные обращения по одному и тому же адресу. Время динамического анализа изменяется неочевидным образом, поскольку с одной стороны в момент обращения происходит небольшая обработка текущих данных, что немного замедляет выполнение, с другой стороны, после выхода из цикла требуется обработать меньший объем накопленных данных, что ускоряет работу. Поэтому время работы сильно зависит от структуры анализируемой программы.

Для каждого вложенного цикла имеется своя собственная информация о памяти, с которой была зафиксирована работа в рамках этого цикла. В момент обращения к памяти информация о ней обновляется только в самом вложенном цикле. После выхода из цикла накопленная информация используется для обновления данных в объемлющем цикле, а также запоминаются найденные зависимости для данного цикла в глобальном хранилище.

Динамический анализатор использует две основные структуры данных: глобальное хранилище результатов анализа и стек контекстов. В глобальном хранилище результатов анализа для каждого цикла программы содержится информация обо всех найденных зависимостях: тип зависимости (прямая, обратная, по выходу),

расстояние зависимости (максимальное и минимальное), возможность приватизации переменной, вызвавшей зависимость. Контекстом называется множество адресов памяти в связке с дополнительной информацией, необходимой для выявления интересующих свойств. При входе в очередной цикл или функцию анализируемой программы создается новый пустой контекст и добавляется в стек. В контексте хранится текущая итерация соответствующего цикла. В случае, когда контекст соответствует функции, итерация не имеет значения и может быть любой. При переходе на следующую итерацию цикла вызывается функция библиотеки анализа, которая изменяет итерацию в соответствующем контексте. В момент обращения к памяти в верхнем контексте стека находится или создается новый объект с информацией об обращениях по данному адресу, при этом используется текущая итерация, хранящаяся в контексте. Например, если по адресу происходит чтение, то в контексте находится объект, соответствующий указанному адресу, и в нем отмечается, что последнее чтение было произведено на такой-то итерации. Если при этом записано, что на другой итерации имелась запись, то фиксируется факт зависимости и вычисляется расстояние между итерациями. Для вычисления расстояния необходимо хранить не только номер итерации последнего чтения/записи, но и номер итерации первого чтения/записи.

При выходе из цикла или функции происходит удаление контекста из стека. Для каждого адреса из удаленного контекста, если было зафиксировано чтение/запись по этому адресу, фиксируются соответствующие операции в вершине стека. В глобальное хранилище добавляется информация об обнаруженных зависимостях для цикла, соответствующего удаленному контексту.

Определение факта использования переменной после цикла происходит с использованием других структур данных, за исключением глобального хранилища.

Во время выполнения программы при выходе из цикла выполняются следующие действия. Для каждой переменной, к которой производилось обращение в цикле (в том числе, и неявно, внутри вызываемых функций), запоминается адрес структуры данных в глобальном хранилище, содержащей информацию о зависимостях в данном цикле. Запомненные адреса хранятся в списке, который назовем «списком циклов». Для каждой переменной будет создан свой список циклов.

При обращении к переменной происходит поиск соответствующего ей списка циклов. В случае отсутствия списка не требуется предпринимать каких-либо действий. Если же список найден и к переменной обратились на чтение, то для каждого цикла в списке отмечается, что данная переменная используется после этого цикла. После указанных действий список удаляется.

Для корректной обработки автоматических переменных, располагающихся на стеке, необходимо при выходе из функции удалять списки циклов, соответствующие этим локальным переменным.

В момент завершения программы информация из глобального хранилища выдается в заданном формате.

3. ДЕТАЛИ РЕАЛИЗАЦИИ

Библиотека динамического анализа реализована на C++11, инструментация выполняется на уровне LLVM IR. Инструментация заключается во вставке вызовов функций библиотеки динамического анализа. Для каждого инструментируемого объекта динамическому анализатору передается описывающая его метаинформация. Данная информация может быть использована в динамическом анализаторе для определения объекта исходного кода, соответствующего инструментированному объекту. Поддерживается инструментация для программ на языках Fortran и C/C++. LLVM IR не зависит от исходного языка, но отладочная информация, используемая для описания инструментируемых объектов, может несколько отличаться. Например, для описания COMMON-блоков языка Fortran используются структуры специального вида. Поэтому для других языков может отсутствовать высокоуровневое описание инструментируемых данных.

Динамический анализ состоит из следующей последовательности шагов:

1. Получение LLVM IR для каждого файла, который должен быть проанализирован;

2. Компоновка полученных файлов, содержащих LLVM IR, с помощью инструмента llvm-link, входящего в состав LLVM;

3. Инструментация полученного единого файла (модуля LLVM IR);

4. Компиляция инструментированного файла, компоновка его с остальными частями анализируемого проекта, а также с библиотекой динамического анализа.

Результатом запуска полученного исполняемого файла будет либо текстовое описание всех найденных зависимостей по данным, либо описание зависимостей в формате JSON. Желаемый способ представления результатов анализа может быть задан с помощью переменных окружения непосредственно перед исполнением анализируемой программы.

Структура генерируемого JSON-файла приведена в Таблице 2 и включает два основных блока информации: список переменных, для которых выполнялся анализ, и информация о результатах анализа для каждого инструментированного цикла программы. Полученный JSON-файл может быть передан статическому анализатору системы SAPFOR для уточнения результатов анализа. Для соотнесения результатов динамического анализа с внутренним представлением программы в системе SAPFOR используется информация о расположении объектов в исходном коде программы (имя файла, номер строки и столбца), для переменных дополнительно используются их имена.

В разделе приватизируемых переменных указываются, во-первых, переменные, использование которых локализовано в рамках каждой итерации цикла (т. е. для каждой итерации цикла может быть создана своя копия переменной), во-вторых, факт использования значений переменных, полученных на какой-либо итерации цикла после выхода из цикла (UseAfterLoop). Спецификации типа *private*, доступные как в DVMH-языках, так и в OpenMP, приводят к созданию локальных копий исходных переменных, таким образом, при параллельном выполнении после выхода из цикла исходные переменные будут иметь значения, отличные о тех, которые они имели бы при последовательном выполнении. Для указания таких ситуаций используется свойство UseAfterLoop, в этом случае распараллеливание может потребовать дополнительных действий, и применение только спецификации *private* не допустимо. Например, если известно, что используемое после выхода из цикла значение переменной было вычислено на последней итерации цикла, то можно использовать спецификацию *lastprivate* OpenMP-языков.

Таблица 2. Структура JSON-файла отражающего результаты динамического анализа программы

{ "Vars": [{ "File": <path-to-file>, "Line": <number>, "Column": <number>, "Name": <name>, }, "Loops": [{ "File": <path-to-file>, "Line": <number>, "Column": <number>,	Список зарегистрированных переменных. Список зарегистрированных циклов.
"Write": [<var-list>], "Read": [<var-list>],	Режим доступа к зарегистрированным переменным
"Private": [<var-list>], "UseAfterLoop": [<var-list>],	Приватизируемые переменные.
"Flow": [{<var-id> : {"Min": <distance>, "Max": <distance>},...], "Anti": [{<var-id> : {"Min": <distance>, "Max": <distance>},...], "Output": [{<var-id> : {"Min": <distance>, "Max": <distance>},...] } }	Зависимости по данным с указанием расстояние зависимости.

Дополнительным источником параллелизма в циклах является наличие прямых и обратных зависимостей по данным, расстояние которых ограничено константой. Для параллельного выполнения таких циклов DVMH-языки предусматривают спецификацию *across*, а OpenMP-языки – спецификацию *ordered*. Для использования обеих спецификаций требуется явно указать, между какими итерациями цикла существует зависимость. Необходимая для этого информация будет указана при описании зависимостей по данным в результатах динамического анализа. В приведенном в Таблице 3 примере присутствуют как прямая, так и обратная зависимости по данным расстояния один, которое в явном виде указано в спецификациях *across* и *ordered*.

Таблица 3. Пример использования спецификаций across и ordered для параллельного выполнения гнезда циклов с зависимостями

```
#pragma dvm parallel([i][j] on A[i][j]) across(A[1:1][1:1])
for (i = 1; i < N-1; i++)
  for (j = 1; j < N-1; j++)
    A[i][j]=(A[i][j-1]+A[i][j+1]+A[i-1][j]+A[i+1][j])/4.;
```

```
#pragma omp parallel for ordered(2)
for (int i = 0; i < M; i++)
  for (int j = 0; j < N; j++) {
    #pragma omp ordered depend (sink: i - 1, j) depend (sink: i, j - 1)
    A[i][j]=(A[i][j-1]+A[i][j+1]+A[i-1][j]+A[i+1][j])/4.;
    #pragma omp ordered depend (source)
}
```

Инструментация модуля LLVM IR включает в себя:

- вставку объявлений функций динамического анализатора;
- вставку вызовов данных функций в тела функций инструментируемого модуля;
- объявление глобального пула, содержащего мета информацию, описывающую инструментируемые объекты модуля;
- создание вспомогательных функций, отвечающих за инициализацию мета информации, регистрацию типов и глобальных данных;
- вставку вызовов функций инициализации в начало функции, являющейся точкой входа программы.

В процессе инструментации регистрируются вызовы функций, обращения к памяти на чтение и запись, начало, конец, а также начало каждой итерации циклов, связь между фактическими и формальными параметрами.

4. ПРИМЕНЕНИЕ ДИНАМИЧЕСКОГО АНАЛИЗА

Разработанный инструмент был протестирован на тестах производительности из пакета NAS Parallel Benchmarks в версиях 3.3.1 для языков Fortran [8] и C [9]. Оценка накладных расходов на время выполнения приведена на Рис. 2 и Рис. 3. Рост потребляемой памяти (в количестве раз) указан на Рис. 4 и 5.

Тестирование выполнялось на Intel Xeon CPU E5-1660 v2, 3.70 GHz, с отключенным Turbo Boost. Для получения LLVM IR и компиляции программ использовались компиляторы Clang и Flang версий 7.1.0 с опорой на библиотеки компилятора GCC 7.4. Компиляция выполнялась с использованием опции –O3. В отличие от применения бинарной инструментации использование данной опции допускается, так как входящие в ее состав оптимизации применяются уже после инструментации исходной программы. Также приведены накладные расходы при использовании Intel Advisor. В данном случае анализ зависимостей выполнялся с опцией –O0, чтобы гарантировать получение достоверных результатов. При этом замедление указано относительно программ, собранных с опцией –O3.

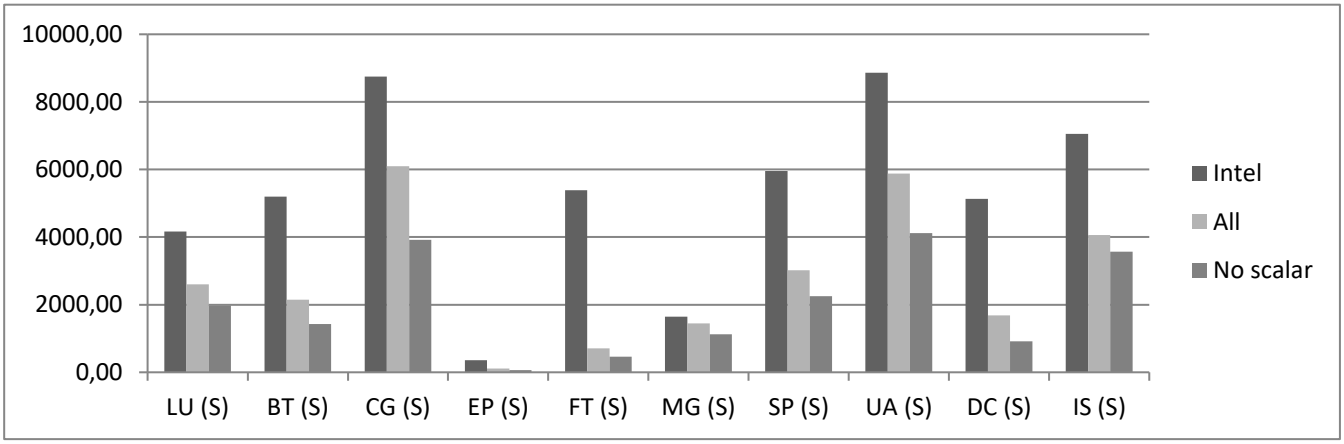


Рис. 2. Замедление C-тестов из набора NAS Parallel Benchmarks при полной (All) и выборочной (No scalar) инструментациях и использовании Intel Advisor

С целью снижения накладных расходов были использованы реализованные в SAPFOR средства статического анализа [7]. В большинстве случаев их оказывается достаточно для анализа скалярных переменных, к которым не применяются операции взятия адреса и которые не участвуют в операциях адресной арифметики.

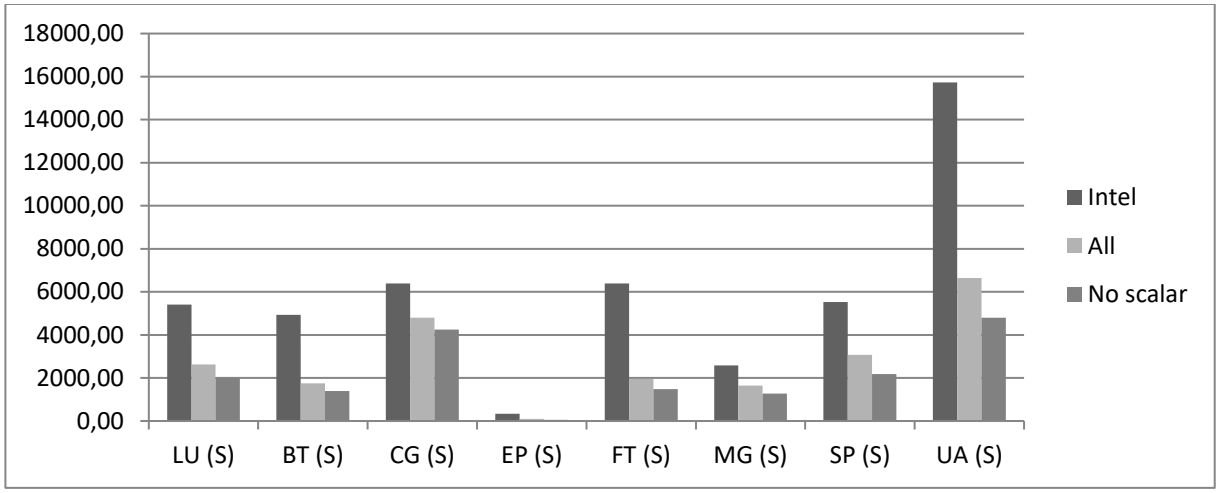


Рис. 3. Замедление Fortran-тестов из набора NAS Parallel Benchmarks при полной (All) и выборочной (No scalar) инструментациях и использовании Intel Advisor

Для зависимостей по данным переменным выполняется их классификация с целью устранения зависимостей за счет использования редукционных операций и приватизации соответствующих данных. Данные переменные могут быть проигнорированы в процессе динамического анализа, более того, обращения к данным переменным могут быть оптимизированы средствами LLVM, например, они могут быть размещены на регистрах. Применение данной оптимизации позволило сократить замедление анализируемых программ до 40% (на тесте EP).

Таким образом, с учетом данной оптимизации время выполнения увеличивается в среднем до 2000 раз. Но совместно с применением статического анализа обеспечивается полный анализ всей программы. При этом анализ трети всех циклов, выполняемый средствами Intel Advisor, приводит к замедлению программы в среднем до 5000 раз.

Также была реализована дополнительная возможность, позволяющая выбрать функцию, начиная с которой должна выполняться инструментация. В данном случае будут проанализированы указанная функция, а также все функции, которые из нее вызываются. Выборочный анализ отдельных функций, важных для распараллеливания конкретного цикла, значительно ускоряет анализ программ. Например, для теста LU (класс S) рост времени для анализа одного из основных циклов с регулярной зависимостью по данным, которую можно устранить за счет конвейерного выполнения цикла, составляет 707 раз.

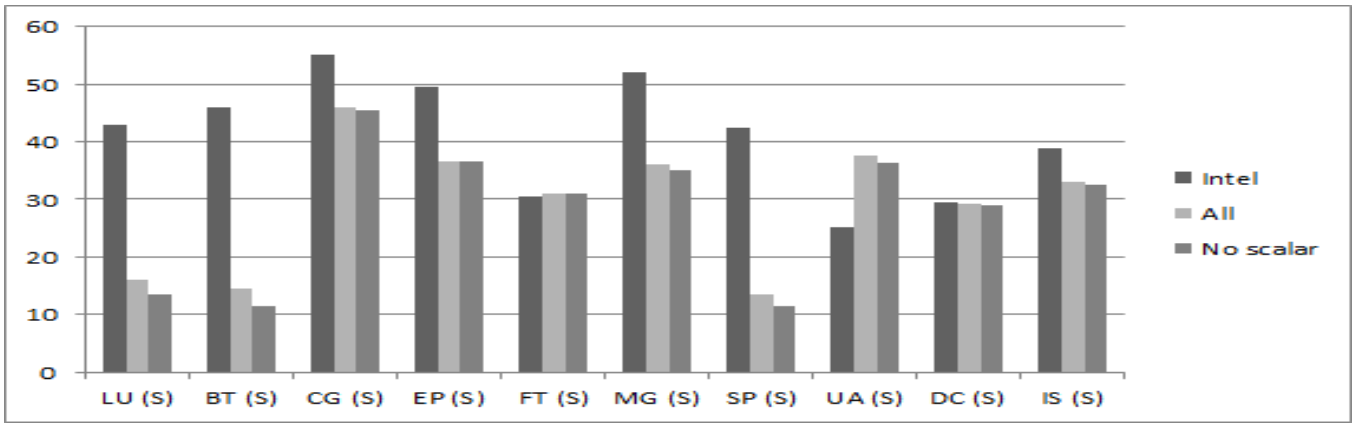


Рис. 4. Рост потребления памяти C-тестов из набора NAS Parallel Benchmarks при полной (All) и выборочной (No scalar) инструментациях и использовании Intel Advisor

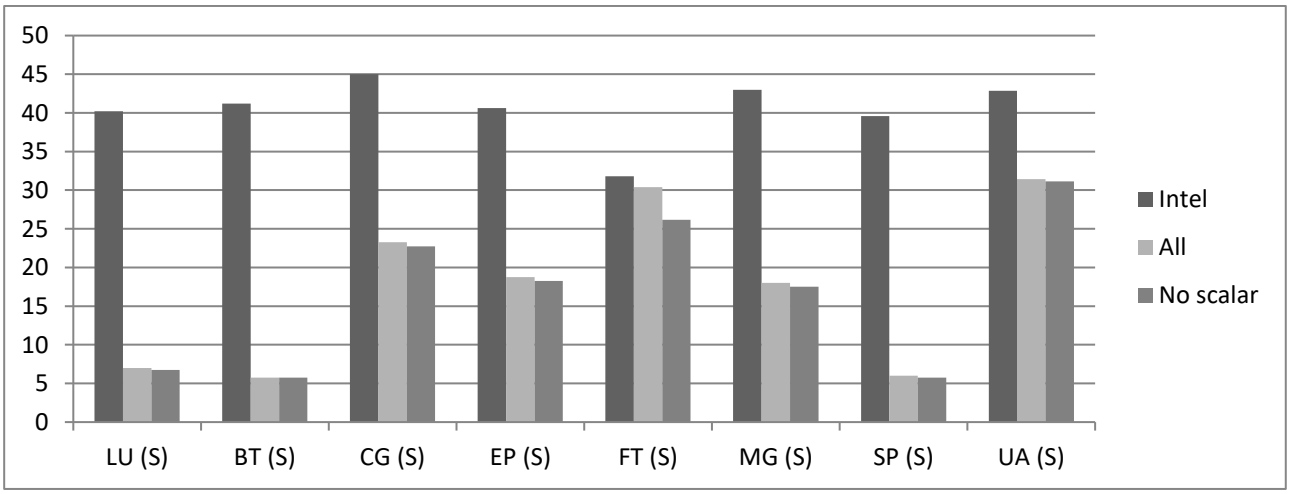


Рис. 5. Рост потребления памяти Fortran-тестов из набора NAS Parallel Benchmarks при полной (All) и выборочной (No scalar) инструментациях и использовании Intel Advisor

ЗАКЛЮЧЕНИЕ

В работе рассмотрен инструмент, предназначенный для динамического анализа зависимостей по данным, реализованный в рамках системы SAPFOR. Инструмент может быть использован как для получения результатов анализа с целью автоматизации распараллеливания программ в системе SAPFOR, так и с целью ручного распараллеливания программ. Помимо определения основных типов зависимостей (flow, anti, output) также предоставлена рекомендация о том, какие зависимости могут быть устранены за счет приватизации переменных, в том числе, массивов

(что особенно важно из-за ограниченности возможностей статического анализа), а также о том, для каких циклов может быть организовано конвейерное выполнение. При этом предоставляемой информации о расстоянии зависимостей достаточно для расстановки директив ACROSS DVM-языков, позволяющих организовать конвейер автоматически. Статический анализ скалярных переменных позволил снизить накладные расходы на проведение динамического анализа, не потеряв полноты результатов. Использование инструментации LLVM IR вместо бинарной инструментации помимо проведения предварительного статического анализа позволяет выполнять оптимизацию программы после инструментации без потери точности результатов и их соотнесения с объектами исходного кода. Дальнейшие исследования планируется направить на большее снижение накладных расходов, так как распараллеливание для вычислительных систем с распределенной памятью требует анализа всей программ, сильно ограничивая возможности выборочного анализа циклов.

Исходные коды системы SAPFOR доступны на GitHub [13].

СПИСОК ЛИТЕРАТУРЫ

1. *Клинов М.С., Крюков В.А.* Автоматическое распараллеливание Фортран-программ. Отображение на кластер // Вестник Нижегородского университета им. Н.И. Лобачевского, 2009. № 2. С. 128–134.
2. *Бахтин В.А., Жукова О.Ф., Катаев Н.А., Колганов А.С., Крюков В.А., Поддерюгина Н.В., Притула М.Н., Савицкая О.А., Смирнов А.А.* Автоматизация распараллеливания программных комплексов // Труды XVIII Всероссийской научной конференции «Научный сервис в сети Интернет», Новороссийск, Россия, 19–24 сентября 2016 г. М.: ИПМ им. М.В. Келдыша, 2016. С. 76–85. doi:10.20948/abrau-2016-31
3. *Konovalov N.A., Krukov V.A, Mikhajlov S.N., Pogrebtsov A.A.* Fortan DVM: a Language for Portable Parallel Program Development // Programming and Computer Software. 1995. V. 21. No. 1. P. 35–38.
4. *Бахтин В.А., Клинов М.С., Крюков В.А., Поддерюгина Н.В., Притула М.Н., Сазанов Ю.Л.* Расширение DVM-модели параллельного программирования для кластеров с гетерогенными узлами // Вестник Южно-Уральского государственного университета, серия «Математическое моделирование и про-

граммирование», 2012. №18 (277), выпуск 12. Челябинск: Издательский центр ЮУрГУ. С. 82–92.

5. *Kim M., Kim H., Luk C.K.* Prospector: A dynamic data-dependence profiler to help parallel programming // HotPar'10: Proceedings of the USENIX workshop on Hot Topics in parallelism, 2010.
6. *Dorta A.J., Rodríguez C., de Sande F., Gonzalez-Escribano A.* The OpenMP Source Code Repository // Parallel, Distributed, and Network-Based Processing, Euromicro Conference, 2005.
7. *Kataev N.A.* Application of the LLVM Compiler Infrastructure to the Program Analysis in SAPFOR // Voevodin V., Sobolev S. (eds) Supercomputing. RuSCDays 2018. Communications in Computer and Information Science, 2018. Vol 965. Springer, Cham. P. 487-499. doi:10.1007/978-3-030-05807-4_41
8. NAS Parallel Benchmarks. UFL: https://www.nas.nasa.gov/publications/npb.html
9. *Seo S., Jo G., Lee J.* Performance Characterization of the NAS Parallel Benchmarks in OpenCL // 2011 IEEE International Symposium on. Workload Characterization (IISWC), 2011. P. 137–148.
10. *Lattner C., Adve V.* LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation // Proc. of the 2004 International Symposium on Code Generation and Optimization (CGO'04). Palo Alto, California, 2004.
11. Intel Parallel Studio. URL: https://software.intel.com/en-us/parallel-studio-xe
12. *Kim M., Kim H., Luk C.K.* SD3: A Scalable Approach to Dynamic Data-Dependence Profiling // 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture. IEEE, 2011. doi:10.1109/MICRO.2010.49
13. SAPFOR. URL: https://github.com/dvm-system.

INVESTIGATION OF DATA DEPENDENCIES BY DYNAMIC ANALYSIS OF SAPFOR

N.A. Kataev [1], A.A. Smirnov [2], A.D. Zhukov [3]

[1,2]*Keldysh Institute of Applied Mathematic RAS;* [3] *Lomonosov Moscow State University*

[1]kataev_nik@mail.ru, [2] smiraland@gmail.com, [3]andreyzkk@yandex.ru

Abstract

The use of pointers and indirect memory accesses in the program, as well as the complex control flow are some of the main weaknesses of the static analysis of programs. The program properties investigated by this analysis are too conservative to accurately describe program behavior and hence they prevent parallel execution of the program. The application of dynamic analysis allows us to expand the capabilities of semi-automatic parallelization. In the SAPFOR system (System FOR Automated Parallelization), a dynamic analysis tool has been implemented, based on on the instrumentation of the LLVM representation of an analyzed program, which allows the system to explore programs in both C and Fortran programming languages. The capabilities of the static analysis implemented in SAPFOR are used to reduce the overhead program execution, while maintaining the completeness of the analysis. The use of static analysis allows to reduce the number of analyzed memory accesses and to ignore scalar variables, which can be explored in a static way. The developed tool was tested on performance tests from the NAS Parallel Benchmarks package for C and Fortran languages. The implementation of dynamic analysis, in addition to traditional types of data dependencies (flow, anit, output), allows us to determine privitizable variables and a possibility of pipeline execution of loops. Together with the capabilities of DVM and OpenMP these greatly facilitates program parallelization and simplify insertion of the appropriate compiler directives.

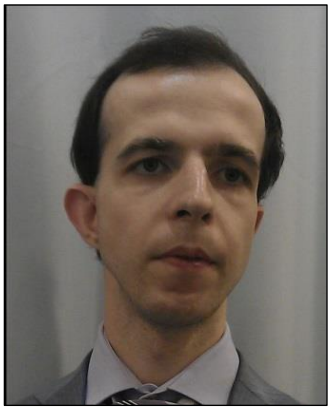
Keywords*: program analysis, dynamic analysis, semi-automatic parallelization, SAPFOR, DVM, LLVM*

REFERENCES

1. *Klinov M.S., Kriukov V.A.* Avtomaticheskoe rasparallelivanie Fortran-programm. Otobrazhenie na klaster // Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo, 2009. No 2. S. 128–134.
2. *Bakhtin V.A., Zhukova O.F., Kataev N.A., Kolganov A.S., Kriukov V.A., Podderiugina N.V., Pritula M.N., Savitskaia O.A., Smirnov A.A.* Avtomatizatsiia rasparallelivaniia programmnykh kompleksov // Trudy XVIII Vserossiiskoi nauchnoi konferentsii «Nauchnyi servis v seti Internet», Novorossiisk, Russia, 19–24 sentiabria. M.: IPM im. M.V. Keldysha, 2016. P. 76–85. doi:10.20948/abrau-2016-31
3. *Konovalov N.A., Krukov V.A, Mikhajlov S.N., Pogrebtsov A.A.* Fortan DVM: a Language for Portable Parallel Program Development // Programming and Computer Software, 1995. V. 21. No. 1. P. 35–38.
4. *Bakhtin V.A., Klinov M.S., Kriukov V.A., Podderiugina N.V., Pritula M.N., Sazanov Iu.L.* Rasshirenie DVM-modeli parallelnogo programmirovaniia dlia klasterov s geterogennymi uzlami // Vestnik Iuzhno-Uralskogo gosudarstvennogo universiteta, seriia "Matematicheskoe modelirovanie i programmirovanie", No 18 (277), vypusk 12. Cheliabinsk: Izdatelskii tsentr IuUrGU, 2012. S. 82–92.
5. *Kim M., Kim H., Luk C.K.* Prospector: A dynamic data-dependence profiler to help parallel programming // HotPar'10: Proceedings of the USENIX workshop on Hot Topics in parallelism, 2010.
6. *Dorta A.J., Rodríguez C., de Sande F., Gonzalez-Escribano A.* The OpenMP Source Code Repository // Parallel, Distributed, and Network-Based Processing, Euromicro Conference, 2005.
7. *Kataev N.A.* Application of the LLVM Compiler Infrastructure to the Program Analysis in SAPFOR // Voevodin V., Sobolev S. (eds) Supercomputing. RuSCDays 2018. Communications in Computer and Information Science, 2018. Vol 965. Springer, Cham. P. 487–499. doi:10.1007/978-3-030-05807-4_41
8. NAS Parallel Benchmarks. URL: https://www.nas.nasa.gov/publications/npb.html
9. *Seo S., Jo G., Lee J.* Performance Characterization of the NAS Parallel Benchmarks in OpenCL // 2011 IEEE International Symposium on. Workload Characterization (IISWC), 2011. P. 137–148.

10. *Lattner C., Adve V.* LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation // Proc. of the 2004 International Symposium on Code Generation and Optimization (CGO'04). Palo Alto, California, 2004.
11. Intel Parallel Studio. URL: https://software.intel.com/en-us/parallel-studio-xe
12. *Kim M., Kim H., Luk C.K.* SD3: A Scalable Approach to Dynamic Data-Dependence Profiling // 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture. IEEE, 2011. doi:10.1109/MICRO.2010.49
13. SAPFOR. URL: https://github.com/dvm-system

СВЕДЕНИЯ ОБ АВТОРАХ



КАТАЕВ Никита Андреевич – научный сотрудник ИПМ им. М.В. Келдыша, специалист в области системного программирования. Сфера научных интересов – компиляторные технологии, автоматизация распараллеливания программ.

Nikita Andreevich KATAEV – Researcher of KIAM RAS, a specialist in system programming. Research interests include compiler technologies, semi-automatic program parallelization.

email: kataev_nik@mail.ru

СМИРНОВ Александр Андреевич – научный сотрудник ИПМ им. М.В. Келдыша, специалист в области системного программирования. Сфера научных интересов – компиляторные технологии, автоматизация распараллеливания программ.

Alexander Andreevich SMIRNOV – Researcher of KIAM RAS, a specialist in system programming. Research interests include compiler technologies, semi-automatic program parallelization.

email: smiraland@gmail.com

ЖУКОВ Андрей Дмитриевич – студент 2 курса магистратуры факультета ВМиК МГУ им М.В. Ломоносова.

Andrey Dmitrievich ZHUKOV – student of CMC faculty of Lomonosov Moscow State University

email: andreyzkk@yandex.ru

Материал поступил в редакцию 15 ноября 2019 года

УДК 004.774.2 + 004.774.6

ИСПОЛЬЗОВАНИЕ МИКРОРАЗМЕТОК ДЛЯ ДОБАВЛЕНИЯ В КОНТЕНТ ВЕБ-СТРАНИЦЫ ДАННЫХ ВНЕШНИХ РЕСУРСОВ

Е. Л. Китаев[1], Р. Ю. Скорнякова [2]

Институт прикладной математики им. М.В. Келдыша Российской академии наук, г. Москва;

[1]kitaev@keldysh.ru, [2]rimmaskorn@gmail.com

Аннотация

В семантических разметках Всемирной паутины накоплено большое число данных, и их количество продолжает расти. Однако потенциал этих данных реализуется, на наш взгляд, не в полной мере. Данные, заключенные в семантических разметках, или микроразметках, широко используются поисковыми системами, отчасти социальными сетями, использование же этих данных разработчиками приложений, как правило, основано на приведении данных к стандарту RDF и выполнении SPARQL-запросов, что требует хорошего знания этого языка и умения программировать. В настоящей работе предложено использовать имеющиеся в Сети семантические разметки для автоматического включения их содержимого в контент других веб-страниц и описан инструмент для реализации такого включения, не требующий от разработчика веб-страницы владения какими-либо языками программирования помимо широко известных HTML и CSS. Инструмент не требует установки, работу выполняют подключаемые стартовые скрипты. В настоящий момент инструмент поддерживает семантические данные, заключенные в популярных типах разметок «микроданные» и JSON-LD, в тегах <meta> HTML-документов и свойствах документов Word и PDF.

***Ключевые слова**: семантическая паутина, семантические технологии, семантическая разметка, микроразметка, микроданные, JSON-LD, веб-разработка, веб-технологии*

ВВЕДЕНИЕ

Бурно начавшись, развитие Семантической паутины [1] к началу 2010-х годов затормозилось. Реализация идеи превращения Всемирной паутины в Се-

мантическую столкнулась с определенными трудностями, одной из которых стала трудность освоения веб-разработчиками языка RDFa[1] (Resourse Description Framework in attributes) – основного языка семантической разметки веб-страниц. Из-за сложности RDFa (оборотной стороны его универсальности) использование этого языка оказалось ограниченным, а имеющиеся разметки содержали большое число ошибок. Новый импульс семантическому наполнению Всемирной паутины придало появление на рубеже 2010-х годов альтернативных синтаксисов семантической разметки «микроданные»[2], RDFa Lite[3] и JSON-LD[4], обладающих лучшим соотношением гибкости и сложности. По данным аналитической компании W3Techs[5], собирающей статистику об использовании различных веб-технологий, частота использования разметок «микроданные»[6] (≈15%) и JSON-LD[7] (≈27%) на 10 миллионах наиболее популярных сайтах в настоящий момент уже выше, чем частота использования Generic RDFa[8] (≈13%). При этом, если рост использования «микроданных» приостановился, то использование JSON-LD продолжает расти. На данный момент этот тип микроразметки представляется наиболее перспективным для дальнейшего семантического наполнения Всемирной паутины и использования в различных приложениях.

Популярность разметок «микроданные» и JSON-LD помимо более простого синтаксиса объясняется также тем, что их вместе со словарем Schema.org[9] активно используют поисковые системы. Содержащиеся в разметках данные служат для представления информации о веб-станицах в виде так называемых «расширенных сниппетов» (rich snippets)[10]. В отличие от обычных сниппетов, представляющих собой выдержку из неструктурированного текста, расширенные сниппеты содержат структурированную информацию, лучше отражающую содержание веб-страницы. И хотя напрямую семантические разметки при ран-

[1] http://rdfa.info/
[2] https://html.spec.whatwg.org/multipage/microdata.html
[3] https://www.w3.org/TR/rdfa-lite/
[4] https://json-ld.org/
[5] https://w3techs.com/
[6] https://w3techs.com/technologies/details/da-microdata/all/all
[7] https://w3techs.com/technologies/details/da-jsonld/all/all
[8] https://w3techs.com/technologies/details/da-genericrdfa/all/all
[9] https://schema.org/
[10] https://developers.google.com/search/docs/guides/mark-up-content/

жировании сайтов в настоящее время поисковыми системами не используются, представление в виде расширенного сниппета увеличивает «кликабельность» (click-through rate) и тем самым косвенно сказывается на ранжировании.

Другой вариант использования данных из семантических разметок – создание массивов структурированных веб-данных для проприетарного или свободного использования. Например, компания Google использует данные семантических разметок наряду с другими источниками для наполнения своей базы знаний Google Knowledge Graph[11], информация из которой частично доступна через Google Knowledge Graph Search API[12]. Интересным представляется проект Web Data Commons[13] [2], извлекающий и сохраняющий в формате RDF N-Quads[14] структурированные данные из самого большого из общедоступных массивов веб-данных Common Crawl[15]. Упакованные с помощью GZIP файлы свободно доступны для скачивания и могут использоваться для анализа веб-данных с помощью SPARQL[16] -запросов. Примеры использования данных Web Data Commons можно найти в работах [3, 4].

Для работы с данными, заключенными в семантических разметках, разработано различное программное обеспечение. Существуют инструменты для извлечения и валидации структурированных данных, заключенных в семантических разметках веб-страниц, для преобразования одних форматов структурированных данных в другие, инструменты, реализующие запросы к структурированным данным.

Для валидации структурированных можно использовать инструмент проверки структурированных данных[17], разработанный компанией Google или валидатор микроразметки[18] компании Яндекс. Для извлечения структурированных данных при просмотре веб-страниц разработаны различные плагины к браузе-

[11] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
[12] https://developers.google.com/knowledge-graph
[13]http://webdatacommons.org/
[14] https://www.w3.org/TR/n-quads/
[15] https://commoncrawl.org/
[16] https://www.w3.org/TR/rdf-sparql-query/
[17] https://search.google.com/structured-data/testing-tool/u/0/
[18] https://webmaster.yandex.ru/tools/microtest/

рам, например, плагины к браузеру Google Chrome Microdata.reveal[19], Semantic inspector[20], JSON-LD Tester[21], Structured Data Testing Tool[22]. Они позволяют нажатием на иконку выделить и представить в отдельном окне структурированные данные текущей страницы.

Среди инструментов, предназначенных для программистов и позволяющих извлекать и преобразовывать структурированные данные из одного формата в другой, одним из наиболее популярных является свободно распространяемый фреймворк Apache Any23[23] (Anything To Triples), включающий библиотеку на языке Java, инструмент командной строки и веб-службу. Поддерживаются формат RDF[24] в различных сериализациях, RDFa, микроформаты[25], JSON-LD, микроданные, CSV, YAML[26]. Для преобразования «микроданных» в RDF используют также библиотеку на языке Python pyMicrodata[27]. Для извлечения структурированных данных из веб-страниц может служить API Валидатора микроразметки[28] компании Яндекс.

Для работы со структурированными данными формата RDF широко используется фреймворк с открытым кодом Apache Jena[29], написанный на языке Java. Он включает набор различных API и инструментов командной строки. Пример использования Apache Jena для обработки данных о товарах одного из интернет-магазинов приведен в блоге [5] сотрудника компании Commonwealth Computer Research, Inc, специалиста в области Semantic Web Боба ДюШарма. Данные о товарах, заключенные в разметках формата JSON-LD, преобразуются в

[19] https://chrome.google.com/webstore/detail/microdatareveal/olapakiakkblfdaajcifgldandnikpdh?hl=ru
[20] https://chrome.google.com/webstore/detail/semantic-inspector/jobakbebljifplmcapcooffdbdmfdbjh
[21] https://chrome.google.com/webstore/detail/json-ld-tester/aohmciehgjboidolkmoaofcbnejmokan?hl=de
[22] https://chrome.google.com/webstore/detail/structured-data-testing-t/kfdjeigpgagildmolfanniafmplnplpl
[23] https://any23.apache.org/
[24] https://www.w3.org/RDF/
[25] http://microformats.org/
[26] https://learn.getgrav.org/16/advanced/yaml
[27]https://github.com/RDFLib/pymicrodata
[28] https://yandex.ru/dev/validator/
[29] https://jena.apache.org/

формат RDF, а затем с помощью процессора SPARQL-запросов ARQ, входящего в состав Apache Jena, делаются выборки товаров с определенными характеристиками.

Более подробное изложение стандартов, концепций, приложений и инструментов для разработчиков, относящихся к области Semantic Web, имеется в книге [6]. В этой книге, как и во множестве других работ (см., например, [7, 8]), основной подход к разработке приложений, использующих структурированные данные, связан с использованием стандарта RDF и языка запросов SPARQL. Такой подход безусловно является мощным и универсальным, позволяющим создавать разнообразные приложения, однако для его реализации разработчик должен был специалистом в этой области, что ограничивает, на наш взгляд, возможности использования размещенных в Сети семантических данных. Эта область нуждается в инструментах, которые могли бы использовать не только специалисты в области Semantic Web, но и другие разработчики.

В настоящей работе мы предлагаем одно из возможных использований семантических данных при разработке веб-страниц и инструмент для его реализации, не требующий каких-либо дополнительных знаний и умений, помимо владения минимумом, необходимым веб-разработчику: языком разметки HTML и языком стилей CSS, а также знания основ синтаксиса микроразметок. Инструмент позволяет «на лету», в момент загрузки веб-страницы, включать в контент веб-страницы данные, заключенные в семантических разметках внешних ресурсов, и при этом не требует программирования.

1. МИКРОРАЗМЕТКИ И ДИНАМИЧЕСКАЯ АГРЕГАЦИЯ ВЕБ-ДАННЫХ БЕЗ ПРОГРАММИРОВАНИЯ

Как правило, использование семантических данных, размещенных в Сети, предполагает их предварительное извлечение и промежуточное хранение для дальнейшей обработки. Однако, если не стоит задача отбора веб-ресурсов по определенным условиям, гиперссылки на ресурсы известны, и данные необходимы только для показа пользователю, организация промежуточного хранения не является необходимой. Примером может служить составление для размещения в Сети разного рода списков: списков организаций с контактными данными, библиографических списков, подборок кулинарных рецептов, таблиц с ценами

на один и тот же товар в разных интернет-магазинах и т. п. В этих случаях данные можно динамически извлекать непосредственно из веб-ресурсов, где они размещены, и это особенно актуально, если данные не постоянны, как, например, контактные данные, цены, даты последней редакции в ссылках на живые публикации [9, 10] и т. п.

В общем случае задача извлечения информации из Сети весьма сложна, поскольку данные слабо структурированы и предназначены в первую очередь для прочтения человеком. В этом случае необходимо кодирование алгоритма извлечения данных, и для разных веб-ресурсов это алгоритм может быть разным. И хотя извлечение данных из Сети (веб-скрейпинг) является весьма популярной задачей и существует множество инструментов, помогающих в ее решении[30], создание на их основе инструмента, который позволил бы веб-разработчику без программирования динамически включать в контент данные из разных веб-ресурсов, не представляется возможным.

При наличии семантической разметки задача существенно упрощается, поскольку алгоритм извлечения данных зависит только от типа разметки. Ее успешно решают, например, упоминавшиеся выше фреймворк Apache Any23[31] и API Валидатора микроразметки[32] компании Яндекс. Однако мы не стали их использовать в качестве основы для нашего инструмента из-за ряда недостатков при одновременном извлечении данных из нескольких источников: для получения данных надо делать отдельный запрос для каждого из источников, из-за чего увеличивается общее время получения данных; в запросе нельзя указать тип данных – результат включает все, имеющиеся в разметке – и т. п.

Созданный нами инструмент StructScraper [11] позволяет динамически извлекать и включать в контент веб-страницы семантические данные, заключенные в разметках «микроданные» и JSON-LD, в тегах <meta> HTML-документов и свойствах документов Word и PDF.

Стандарт разметки «микроданные» был предложен в 2008 году сотрудником Google и на тот момент участником консорциума W3C Яном Хиксоном как

[30] http://scraping.pro/software-for-web-scraping/, https://www.garethjames.net/a-guide-to-Web-scrapping-tools/
[31] https://any23.apache.org/
[32] https://yandex.ru/dev/validator/

часть стандарта HTML5 с целью добавления семантики к имеющимся html-элементам. Для этого в HTML были введены специальные глобальные атрибуты itemscope, itemtype, itemprop, itemid, itemref. Атрибут itemscope помечает html-элемент как узел микроданных, атрибут itemtype задает тип данных (как правило, тип выбирается из какого-нибудь словаря), атрибут itemprop задает имя свойства, атрибут itemid предназначен для глобальных идентификаторов, атрибут itemref предназначен для ссылки на свойство, не содержащееся внутри узла микроданных. В листинге 1 приведен пример разметки микроданными контактных данных организации с использованием словаря Schema.org.

```
<div itemscope itemtype="http://schema.org/Organization">
    <span itemprop="name">ИПМ им. М.В.Келдыша РАН</span>
    <div>
        Контакты
        <div itemprop="address"
             itemscope
             itemtype="http://schema.org/PostalAddress">
            Адрес:
            <span itemprop="postalCode">125047</span>,
            <span itemprop="addressLocality">Москва</span>,
            <span itemprop="streetAddress">
                Миусская пл.,д.4
            </span>
        </div>
        Телефон:
        <span itemprop="telephone">+7 499 978-13-14</span>,
        Факс:
        <span itemprop="faxNumber">+7 499 972-07-37</span>,
    </div>
</div>
```

Листинг 1. Микроразметка с использованием микроданных

Стандарт разметки JSON-LD (JSON for Linking Data) был предложен в 2010 году как альтернатива формату RDFa программистами, которые в своем проекте использовали для внутреннего хранения данных широко распространенный в веб-приложениях формат JSON. При этом данные, которые они извлекали и об-

рабатывали, хранились в Сети в формате RDFa. Идея состояла в том, чтобы данные в веб-документах, которые предназначаются для программной обработки, тоже хранились в формате JSON. В этом случае отпадает необходимость преобразования одного формата в другой. В 2014 году JSON-LD стал официальным стандартом, поддерживаемым консорциумом W3C. JSON-LD совместим с RDF, он является еще одним способом сериализации RDF.

В html-документе данные формата JSON-LD помещаются в тег <script> с атрибутом type="application/ld+json". Ключевые слова, входящие в синтаксис JSON-LD, начинаются с символа "@". В листинге 2 приведен пример разметки JSON-LD для научной публикации.

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Живая публикация",
  "author": {
       "@type": "Person",
       "name": "Горбунов-Посадов, Михаил Михайлович"
            },
  "datePublished": "2011",
  "dateModified": "2019-05-01",
  "url": "https://keldysh.ru/gorbunov/live.htm"
}
</script>
```

Листинг 2. Микроразметка с использованием JSON-LD и словаря Schema.org

Для использования предлагаемого нами инструмента автору веб-страницы достаточно соответствующим образом подготовить разметку HTML-страницы (вставив нужные атрибуты) и подключить стартовые скрипты (вставив в страницу фрагмент заранее подготовленного кода) – вся остальная работа по включению данных из семантических разметок внешних ресурсов выполняется автоматиче-

ски в процессе загрузки страницы. Инструкция по оформлению веб-страницы размещена на посвященном инструменту сайте[33].

StructScraper могут использовать как профессиональные разработчики, так и непрофессиональные авторы, создающие собственные страницы, поскольку для его использования необходимо знать только основы HTML и CSS. Он может быть полезен блогерам, авторам страниц с кулинарными рецептами, научным работникам для создания персональных страниц и списков публикаций, его можно использовать для сравнения цен на товары, рейтингов сайтов и т.п.

2. ПРИМЕРЫ ИПОЛЬЗОВАНИЯ

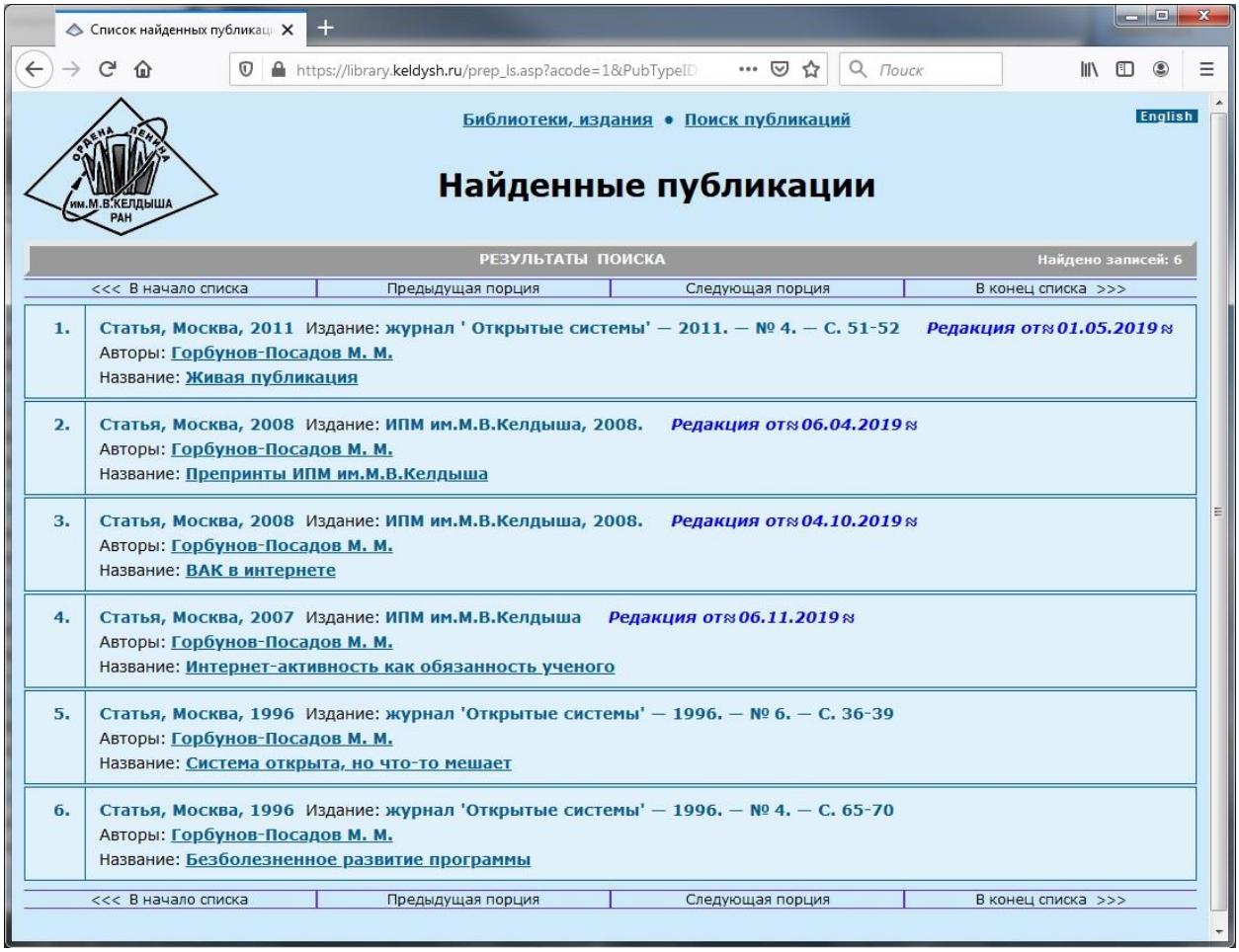


Рис. 1. Пример использования инструмента: дата последней редакции в ссылке на «живую» публикацию

С необходимостью автоматического включения в контент веб-страницы данных внешних веб-ресурсов мы впервые столкнулись в работе над инструмен-

[33] http://struct-scraper.keldysh.ru/doc-page.html

тами поддержки «живых» публикаций [9]. «Живая» публикация – это размещенная в свободном доступе в интернете научная работа, которая постоянно совершенствуется и развивается ее автором. Дата последней редакции является важным атрибутом такой публикации, показывающим, насколько актуальна работа. Эта дата должна храниться в самой публикации и автоматически обновляться в онлайн-ссылках на нее. Ручное обновление смысла не имеет, поскольку может происходить с запаздыванием. Ссылки на работу могут размещаться на веб-страницах разных сайтов, и не все авторы этих страниц могут быть профессиональными фронтенд-разработчиками, поэтому необходим был простой в использовании инструмент [10]. Инструмент был реализован и внедрен в ИПМ им. М.В. Келдыша в 2017 году (рис. 1).

Эта задача не единственная, где автоматическое включение в контент веб-страницы данных внешних ресурсов может быть полезным. Поэтому возникла идея разработать более общий инструмент, пригодный для разных случаев использования.

Примером такого использования может служить мониторинг цен. Сейчас многие интернет-магазины добавляют микроразметку к описаниям товаров, представленных в каталогах, включающую цену на товар. Веб-страница, доступная по адресу https://struct-scraper.keldysh.ru/test_pages/product.html, содержит таблицу с ценами на одну и ту же модель смартфона (рис. 2), подгружаемыми «на лету», что гарантирует их актуальность на момент загрузки. Изначальная разметка для каждой строки таблицы имеет вид, как в листинге 3. Она содержит только гиперссылку на модель смартфона. В html-код добавлен также вызов готового плагина jQuery, которому в качестве параметров передаются адрес REST API и тип Product из словаря Schema.org. При такой разметке подгружаются все свойства Product и записываются в теги . Какие из них должны быть видны пользователю, определяется в CSS.

Смартфон Samsung Galaxy Note 10

Интернет-магазин	Модель	Цена
1Click	Samsung Galaxy Note 10	47690
Болтун	Samsung Galaxy Note 10 8/256GB N970 Черный РСТ	56489
Комус	Смартфон Samsung Galaxy Note 10 256 ГБ черный (SM-N970FZKDSER)	76990.00
М-Видео	Смартфон Samsung Galaxy Note10 Black (SM-N970F)	76990
Мегафон	Смартфон Samsung Galaxy Note10 Чёрный	76 990
МТС	Смартфон Samsung N970 Galaxy Note 10 8/256Gb Черный	76990.00
Плеер.ру	Сотовый телефон Samsung SM-N970F Galaxy Note 10 8Gb/256Gb Black	61599
Самсунг	Samsung Galaxy Note 10, 256 Гб, Чёрный	76990
Технопарк	Смартфон Samsung Galaxy Note10 черный	76990.00
ТехноСити	Смартфон Samsung SM-N970 Galaxy Note 10, 256 Gb, чёрный (SM-N970FZKDSER)	76990.00
Allo.market	Samsung Galaxy Note 10 8/256GB (EXINOS)	55890
Ant-Shop	Смартфон Samsung Galaxy Note 10 8/256GB Черный	69990
Appleavenue	Samsung Galaxy Note 10 8/256Gb (SM-N970F) (Aura White)	52550
Elecity	Смартфон Samsung Galaxy Note 10 8/256GB черный	63130
FLASH	Смартфон Samsung Galaxy Note 10 (2019) SM-N970 8/256GB черный, SM-N970FZKDSER	54 380.–
Galaxystore	Galaxy Note10 256 ГБ, черный	76990
GPod	iMac	45912
HiFi Zona	Смартфон Samsung Galaxy Note 10 Black (SM-N970F) черный	58970
Likemobmarket	Samsung Galaxy Note 10 256GB Aura Glow	47679
lite-mobile	Смартфон Samsung Galaxy Note 10 SM-N970F/DS 256Gb (Цвет: Aura Black)	56750

Рис. 2. Пример использования инструмента: мониторинг цен

```
<tr class="import-struct">
  <td>Мегафон</td>
  <td>
    <a class="struct-url"
       href="https://moscow.shop.megafon.ru/mobile/117195.html">
    </a>
   </td>
</tr>
```

Листинг 3. HTML-разметка для строки таблицы с ценами

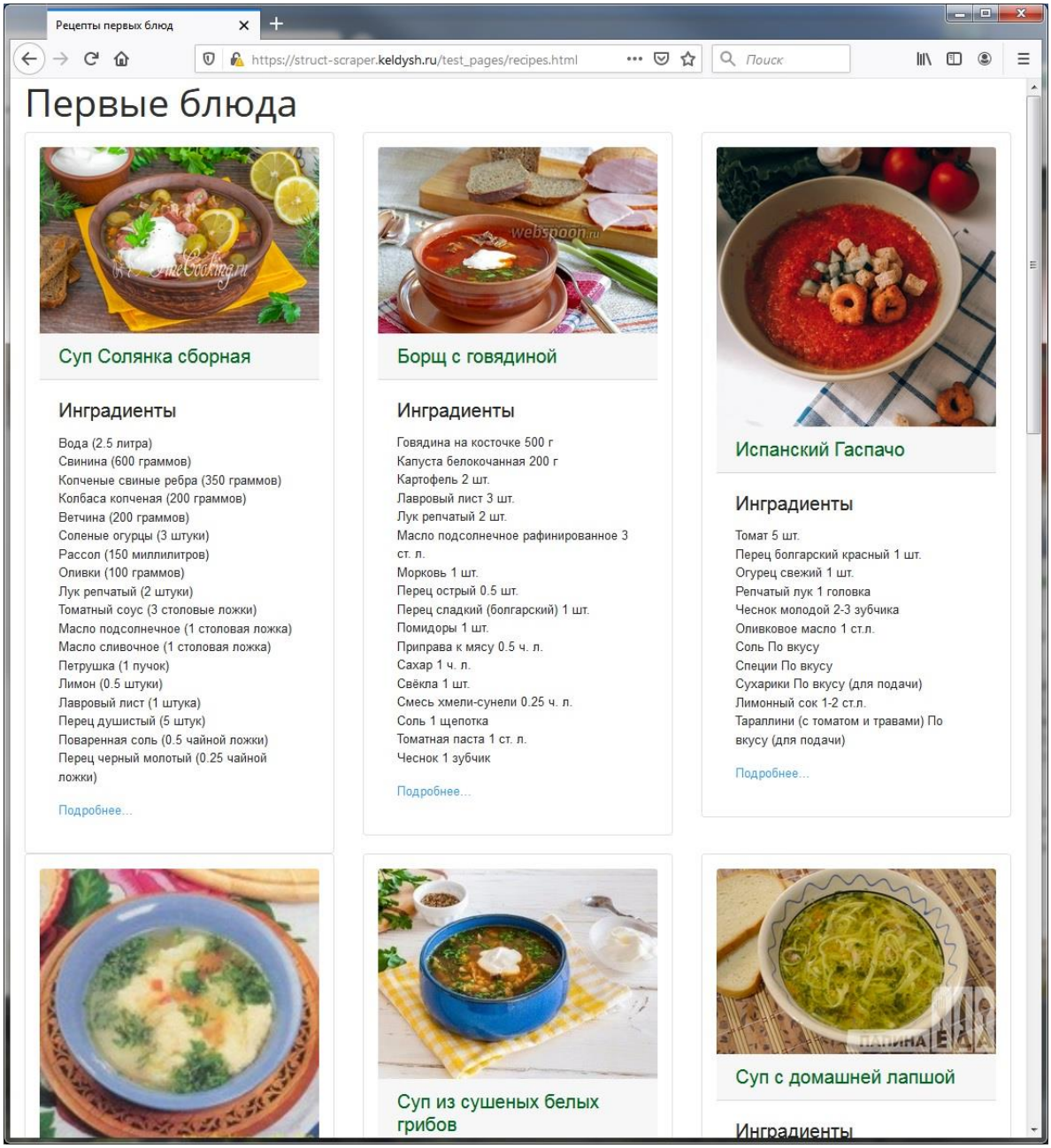


Рис. 3. Пример использования инструмента: страница с кулинарными рецептами

```
<div class="card import-struct">
    <a href="https://papinaeda.ru/244">
        <img class="card-img-top"
             itemprop="resultPhoto"
             alt="Фото"
             data-unique="true">
    </a>
    <a href="https://papinaeda.ru/244">
        <img class="card-img-top"
             itemprop="image"
             alt="Фото"
             data-unique="true">
    </a>
    <h5 class="card-header">
        <a class="struct-url"
           href="https://papinaeda.ru/244" itemprop="name">
           Суп с домашней лапшой
        </a>
    </h5>
    <div class="card-body">
        <h5 class="card-title">Ингрaдиенты</h5>
        <ul class="list-unstyled">
            <li itemprop="recipeIngredient"></li>
            <li itemprop="ingredients"></li>
        </ul>
        <a href="https://papinaeda.ru/244"
           class="card-link">
           Подробнее...
        </a>
    </div>
</div>
```

Листинг 4. HTML-разметка для кулинарного рецепта

Еще один пример – страница с подборкой кулинарных рецептов (рис. 3). Рецепты могут располагаться на разных сайтах. При использовании предлагаемого инструмента составителю подборки нет необходимости копировать их на свою страницу. Достаточно только вставить гиперссылки на рецепты в шаблон html-разметки, и рецепты автоматически загрузятся на страницу. Пример доступен по адресу https://struct-scraper.keldysh.ru/test_pages/recipes.html. Поскольку в этом случае на страницу добавляются из внешних сайтов не только тексты, но и изображения, здесь используется тип разметки, отличный от предыдущего примера (листинг 4).

На сайте инструмента [11] и его странице[34] на веб-сервисе GitHub можно найти дополнительные примеры использования.

3. РЕАЛИЗАЦИЯ ИНСТРУМЕНТА

Реализация предлагаемого нами инструмента StructScraper включает серверную и клиентскую части. Серверная часть представляет собой REST API для извлечения данных – ее можно использовать как вместе с клиентской частью, так и самостоятельно. Клиентская включает jQuery плагины, вызов которых при загрузке веб страницы выполняет работу по добавлению данных в контент.

REST API StructScraper реализован на языке C#, имеющем встроенную поддержку асинхронного программирования, с использованием технологии Microsoft ASP.NET Web API.

Обращения к REST API производятся методом POST с передачей параметров в формате JSON. Параметры включают адреса веб-ресурсов, из которых необходимо извлечь данные, и сведения о том, какие именно данные должны быть извлечены. Для метаданных, извлекаемых из тегов и свойств документов, передаются названия, для микроданных и разметки JSON-LD передается список типов из словаря Schema.org.

Обработка нескольких URL на сервере происходит асинхронно, время ответа равно максимальному времени ответа от одного ресурса. Поэтому время ответа на клиентский запрос не больше, чем если бы запросы с клиентской стороны для каждого URL производились отдельно, а за счет сокращения времени на установку отдельных соединений к REST API, а также из-за отсутствия ограничений на число одновременно выполняемых асинхронных запросов, оно становится меньше.

REST API StructScraper допускает CORS, запросы к нему могут производиться из клиентских частей веб-приложений любых доменов. CORS[35] (Cross Origin Resource Sharing) – это технология, реализованная в современных браузерах, частично снимающая ограничения правила одного источника, введенного из соображений безопасности с целью дать возможность коду из веб-страницы одного сайта свободно взаимодействовать с ресурсами этого же сайта и максимально

[34] https://github.com/RimmaSkorn/struct-scraper
[35] https://fetch.spec.whatwg.org/#http-cors-protocol

ограничить такое взаимодействие с ресурсами других сайтов. Без такого ограничения скрипт из страницы, загруженной с какого-нибудь сайта, мог бы обратиться, например, к почтовому серверу пользователя через сессию, открытую в другом окне браузера, получить его почту или отправить от его имени письмо, что непременно бы использовали злоумышленники. В соответствии с правилом одного источника браузеры запрещают производить ajax-запросы к стороннему серверу. До появления технологии CORS такие запросы были запрещены полностью. Технология CORS, реализованная как надстройка над HTTP протоколом, позволяет их осуществлять, если сторонний сервер дает на это явное разрешение. При этом сервер также контролирует детали кросс-доменных запросов: разрешенные методы, возможности передачи авторизующих заголовков и т.п.

Работу по загрузке клиентом данных на веб-страницу осуществляет JavaScript код, оформленный в виде плагинов jQuery. Для того чтобы плагин загрузил данные на веб-страницу, в нее должна быть добавлена специальная разметка, по которой код плагина определяет, из каких веб-ресурсов должны быть загружены данные и какие именно данные необходимо загрузить. Имеется несколько типов разметок с использованием только атрибутов class и с использованием атрибутов class и микроданных.

StructScraper является инструментом с открытым кодом, доступном[36] на веб-сервисе GitHub.

ЗАКЛЮЧЕНИЕ

Сегодня во Всемирной паутине еще редко встречаются страницы, предъявляющие посетителю информацию из нескольких активно обновляемых сайтов, собранную в момент обращения к этой странице. В настоящей работе

— предложено использовать для включения в контент веб-станицы данные, заключенные в получивших широкое распространение семантических разметках;

— описан созданный авторами инструмент для реализации такого включения, не требующий установки и написания программного кода;

— приведены примеры использования созданного инструмента.

[36] https://github.com/RimmaSkorn/struct-scraper

В дальнейшем предполагается расширить возможности предлагаемого инструмента за счет добавления поддержки разметки RDFa Lite, реализации выбора пользователем словаря семантической разметки, одновременного включения в контент нескольких объектов из одного внешнего ресурса и др. Предполагается также добавить шаблоны разметок для других случаев использования, в частности, для автоматического формирования библиографических ссылок по гиперссылкам на размещенные в Сети научные публикации.

СПИСОК ЛИТЕРАТУРЫ

1. *Bizer C., Heath T., Berners-Lee T.* Linked Data – The Story so far // International Journal on Semantic Web and Information Systems. 2009. V. 5. No. 3. P. 1–22.

2. *Meusel R., Petrovski P., Bizer C.* The WebDataCommons Microdata, RDFa and Microformat Dataset Series // Proceedings of the 13th International Semantic Web Conference: «The Semantic Web – ISWC 2014», Part I, Riva del Garda, Italy, October 19–23, 2014. Lecture Notes in Computer Science, vol. 8796. Springer: 2014. P 277–292.

3. *Lehmberg O., Ritze D., Ristoski P., Meusel R., Paulheim H., Bizer C.* The Mannheim Search Join Engine // Journal of Web Semantics. 2015. V. 35. Part. 3. P. 159–166.

4. *Lohvynenko C., Nedbal D.* Usage of Semantic Web in Austrian Regional Tourism Organizations // Proceedings of the 15th International Conference on Semantic Systems: «SEMANTiCS 2019», Karlsruhe, Germany, September 9–12, 2019. Lecture Notes in Computer Science, vol. 11702. Springer: 2019. P. 3–18.

5. *DuCharme B.* Exploring JSON-LD. URL: http://www.bobdc.com/blog/json-ld/

6. *Yu Liyang.* A Developer's Guide to the Semantic Web. Second Edition. Heidelberg: Springer, 2014. 829 p. DOI:10.1007/978-3-662-43796-4.

7. *Апанович З.В.* Ресурсы и инструменты для преподавания методов и средств Semantic Web // Системная информатика. 2017. № 11. С. 1–20.

8. *Апанович З.В.* Преподавание методов Semantic Web разработчикам программного обеспечения // Труды XIX Всероссийской научной конференции «Научный сервис в сети Интернет», г. Новороссийск, 18–23 сентября 2017 г. М.: ИПМ им. М.В. Келдыша: 2017. С. 9–20. URL: http://keldysh.ru/abrau/2017/37.pdf. DOI:10.20948/abrau-2017-37

9. *Горбунов-Посадов М.М.* Живая публикация // Открытые системы. 2011. № 4. С. 51–52. URL: http://keldysh.ru/gorbunov/live.htm

10. *Горбунов-Посадов М.М., Скорнякова Р.Ю.* Обновляемая дата последней редакции в ссылке на живую публикацию // Препринты ИПМ им. М.В. Келдыша. 2017. № 82. 14 с. DOI:10.20948/prepr-2017-82, URL: http://library.keldysh.ru/ preprint.asp?id=2017-82

11. *StructScraper.* URL: https://struct-scraper.keldysh.ru/

LEVERAGING SEMANTIC MARKUPS FOR INCORPORATING EXTERNAL RESOURCES DATA TO THE CONTENT OF A WEB PAGE

E. L. Kitaev[1], R. Y. Skornyakova [2]

Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow;

[1]kitaev@keldysh.ru, [2]rimmaskorn@gmail.com

Abstract

The semantic markups of the World Wide Web have accumulated a large amount of data and their number continues to grow. However, the potential of these data is, in our opinion, not fully utilized. The semantic markups contents are widely used by search systems, partly by social networks, but the usual approach to using that data by application developers is based on converting data to RDF standard and executing SPARQL queries, which requires good knowledge of this language and programming skills. In this paper, we propose to leverage the semantic markups available on the Web to automatically incorporate their contents to the content of other web pages. We also present a software tool for implementing such incorporation that does not require a web page developer to have knowledge of any programming languages other than HTML and CSS. The developed tool does not require installation, the work is performed by JavaScript plugins. Currently, the tool supports semantic data contained in the popular types of semantic markups “microdata” and JSON-LD, in the <meta> tags of HTML documents and the properties of Word and PDF documents.

Keywords*: semantic web, semantic technologies, semantic markup, microdata, JSON-LD, web development, web technologies*

REFERENCES

1. *Bizer C., Heath T., Berners-Lee T.* Linked Data – The Story so far // International Journal on Semantic Web and Information Systems. 2009. V. 5. No. 3. P. 1–22.

2. *Meusel R., Petrovski P., Bizer C.* The WebDataCommons Microdata, RDFa and Microformat Dataset Series // Proceedings of the 13th International Semantic Web Conference: «The Semantic Web – ISWC 2014», Part I, Riva del Garda, Italy, Oc-

tober 19–23, 2014. Lecture Notes in Computer Science, vol 8796. Springer: 2014, P 277–292.

3. *Lehmberg O., Ritze D., Ristoski P., Meusel R., Paulheim H., Bizer C.* The Mannheim Search Join Engine // J. of Web Semantics. 2015. V. 35. Part. 3. P. 159–166.

4. *Lohvynenko C., Nedbal D.* Usage of Semantic Web in Austrian Regional Tourism Organizations // Proceedings of the 15th International Conference on Semantic Systems: «SEMANTiCS 2019», Karlsruhe, Germany, September 9–12, 2019. Lecture Notes in Computer Science, vol 11702. Springer: 2019. P. 3–18.

5. *DuCharme B.* Exploring JSON-LD. URL: http://www.bobdc.com/blog/json-ld/

6. *Yu Liyang.* A Developer's Guide to the Semantic Web. Second Edition. Heidelberg: Springer, 2014. 829 p. DOI:10.1007/978-3-662-43796-4.

7. *Apanovich Z.V.* Resursy i instrumenty dlia prepodavaniia metodov i sredstv Semantic Web // Sistemnaia informatika. 2017. No 11. S. 1–20.

8. *Apanovich Z.V.* Prepodavanie metodov Semantic Web razrabotchikam programmnogo obespecheniia // Trudy XIX Vserossiiskoi nauchnoi konferentsii «Nauchnyi servis v seti Internet», g. Novorossiisk, 18–23 sentiabria 2017 g. M.: IPM im. M.V. Keldysha: 2017. S. 9–20. URL: http://keldysh.ru/abrau/2017/37.pdf. DOI:10.20948/abrau-2017-37

9. *Gorbunov Posadov M.M.* Zhivaia publikatsiia // Otkrytye sistemy. 2011. No 4. S. 51–52. URL: http://keldysh.ru/gorbunov/live.htm

10. *Gorbunov Posadov M.M., Skorniakova R.Iu.* Obnovliaemaia data poslednei redaktsii v ssylke na zhivuiu publikatsiiu // Preprinty IPM im. M.V. Keldysha. 2017. № 82. 14 s. DOI:10.20948/prepr-2017-82 URL: http://library.keldysh.ru/preprint.asp?id=2017-82.

11. *StructScraper.* URL: https://struct-scraper.keldysh.ru/

СВЕДЕНИЯ ОБ АВТОРАХ

КИТАЕВ Евгений Львович – инженер-исследователь Института прикладной математики им. М.В. Келдыша РАН, специалист в области веб-технологий и информационных систем.

Evgeny L'vovich KITAEV – Research Engineer at the Keldysh Institute of Applied Mathematics RAS, specialist in web technologies and information systems.

email: kitaev@keldysh.ru

СКОРНЯКОВА Римма Юрьевна – научный сотрудник Института прикладной математики им. М.В. Келдыша РАН, специалист в области разработки информационных систем.

Rimma Yuryevna SKORNYAKOVA – Researcher at the Keldysh Institute of Applied Mathematics RAS, specialist in the development of information systems.

email: rimmaskorn@gmail.com

Материал поступил в редакцию 15 ноября 2019 года

УДК 004.912

ОПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКОЙ БЛИЗОСТИ НАУЧНЫХ ЖУРНАЛОВ И КОНФЕРЕНЦИЙ С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА ГРАФА СОАВТОРСТВА

А. С. Козицын, С. А. Афонин, Д. А. Шачнев

НИИ механики МГУ им. М.В. Ломоносова, г. Москва

alexanderkz@mail.ru, serg@msu.ru, mitya57@gmail.com

Аннотация

Количество публикуемых в мире журналов очень велико. В этой связи, необходим программный инструментарий, который позволит анализировать тематические связи журналов. Разработанный авторами и представленный в этой работе алгоритм использует для анализа тематической близости журналов граф соавторства. Алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации. Апробация алгоритма проводилась в наукометрической системе ИАС ИСТИНА. В разработанном для этих целей интерфейсе пользователь может выбрать один близкий ему по тематике журнал, и система автоматически сформирует подборку журналов, которые могут представлять интерес для пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей. В перспективе разработанный алгоритм можно адаптировать для поиска похожих по тематике конференций, сборников публикаций и научных проектов. Наличие такого инструмента увеличит публикационную активность молодых сотрудников, повысит цитируемость статей и цитируемость между журналами. Результаты работы алгоритма определения тематической близости между журналами, сборниками, конференциями и научными проектами также могут использоваться для построения правил в моделях разграничения доступа к данным на основе онтологий предметной области.

Ключевые слова*: тематическая классификация, библиографические данные, граф соавторства, информационные системы*.

1. МОДЕЛЬ

Количество публикуемых в настоящий момент научных журналов очень велико. Например, в информационно-аналитической системе (ИАС) «ИСТИНА» [1] зарегистрировано более 70 тысяч научных журналов и еще более 200 тысяч различных сборников научных публикаций и материалов конференций. В этой связи молодым ученым, аспирантам и студентам необходимы сервисы, которые позволят автоматически подбирать журналы, которые наиболее соответствуют по тематике их научным интересам. Для решения этой задачи может использоваться аккумулированный опыт всего научного сообщества.

Существует несколько возможных способов решения задачи определения тематической близости журналов, позволяющих в автоматическом режиме определить меру сходства между каждой парой журналов. Первый способ основан на использовании тематического анализа полнотекстовых данных, таких, как полнотекстовые описания журналов, тексты опубликованных в журналах статей, их полных или кратких аннотаций, а также указанных авторами ключевых слов. Для проведения тематического анализ полнотекстовых данных в настоящее время разработан широкий спектр различных методов: использование деревьев решений [2]; преобразование текста документа в вектор в многомерном пространстве [3] с использованием частотных характеристик слов [4] и последующим применением геометрических методов классификации (SVN, K-means и аналогичные); нейронные сети. Следует отметить, что в большинстве методов полнотекстовой классификации требуется проведение предобработки текстов, в том числе с использованием методов морфологического анализа [5], которые существенно зависят от языка текста, и требуют предварительной настройки на каждый из используемых языков.

На основе результатов проведенного полнотекстового тематического анализа с использованием указанных методов классификации или кластеризации полных текстов публикаций или их аннотаций возможно построение оценки смысловой близости публикаций в журнале с описанной информационной потребностью конкретного пользователя. Область своих научных интересов пользователь может описать при помощи задания достаточного количества ключевых

слов или загрузить в систему полные тексты своих статей для автоматического построения тематического портрета пользователя. В рамках такого подхода для проведения анализа необходимо иметь достаточно точно описанные тематические профили всех журналов или полные тексты статей, публикуемых в этих журналах.

Получение достаточно полных полнотекстовых данных является сложной задачей, поскольку во многих журналах открытая публикация полных текстов статей не разрешена. Вместе с тем, использование только ключевых слов для проведения тематического анализа может давать слишком общие результаты. В первую очередь это объясняется тем фактом, что во многих случаях ключевые слова статьи характеризуют в большей степени не ее тематику, а связь статьи с одним из приоритетных направлений развития науки, технологий и техники в Российской Федерации. Например, ключевое слово «Нанотехнология», которое часто упоминается в приоритетных направлениях развития науки, технологий и техники в РФ, встречается в статьях совершенно различной тематики: «Разработка новой медицинской нанотехнологии для поражения раковых клеток при детских острых лимфобластных лейкозах»; «Разработка и производство новых наноструктурированных алмазоподобных углеродных покрытий трибологического назначения»; «Разработка и создание сверхчувствительных полевых и зарядовых наноструктур для считывающих и сенсорных устройств наноэлектроники»; «Использование радионуклидов и источников ионизирующего излучения в нанохимии, ядерной медицине и для исследования процессов, происходящих в окружающей среде». Таким образом, тематическая классификация статей по данному ключевому слову определит не столько тематику статьи, сколько участие авторов статьи в проектах по определенному приоритетному направлению. В этой связи использование полнотекстового тематического анализа для решения поставленной выше задачи в наукометрических системах может сталкиваться с определенными трудностями.

Альтернативным методом оценки тематической близости журналов является анализ графа соавторства статей, публикуемых в этих журналах. Такой подход может использоваться как автономно, так и совместно с методами анализа по ключевым словам [6] или текстам. Граф соавторства – это двудольный граф, в котором множество вершин-авторов связано ребрами с множеством вершин-статей. При реализации такого подхода предполагается, что в большинстве случаев

авторы публикуют свои результаты проводимых научных исследований в тематически близких журналах. Вследствие этого в близких по тематике журналах большое количество статей связано в графе соавторов путем длины 2. Основанный на использовании графов соавторства подход, в отличие от методов полнотекстового тематического анализа, не требует наличия полнотекстовой информации о статьях и использует только библиографические данные статей, публикуемых в журналах. Такие данные могут быть получены из наукометрических систем (например, ИАС «ИСТИНА») или систем цитирования (например, WoS).

2. АЛГОРИТМ ОЦЕНКИ ТЕМАТИЧЕСКОЙ БЛИЗОСТИ ЖУРНАЛОВ

Формально задачу оценки близости журналов можно сформулировать следующим образом. Необходимо построить граф, вершинами которого являются журналы, а веса ребер соответствуют их тематической близости.

Разработанный алгоритм на первом шаге для каждой пары журналов вычисляет все пары статей, опубликованных в этих журналах одним автором. Если паре журналов соответствует только одна пара статей, то такие пары считаются не связанными. Если паре журналов соответствует несколько пар статей, то журналы считаются связанными ребром с определенным весом.

В рамках настоящей работы рассматривалось несколько методов определения веса ребра. Наиболее простым методом является определение веса ребра равным количеству уникальных авторов среди соответствующих пар статей. Основным недостатком такого метода является невозможность учитывать значимость авторов для каждой статьи. Во многих случаях статьи пишутся только одним автором, фамилия которого ставится на первом месте в ее библиографическом описании. Остальные соавторы могут участвовать в работе над статьей незначительно, и их основное направление научной деятельности может не совпадать с ее тематикой.

Для проверки гипотезы о значимости порядка авторов при проведении тематического анализа была проведена оценка доли статей, в которых порядок авторов определяется лексикографическим порядком, а не значимостью в работе над статьей. Из наукометрической системы МГУ были отобраны для анализа все статьи в журналах за 2014–2017 гг. с количеством авторов от 2 до 7.

Для каждого из указанного количества авторов были посчитаны проценты

статей L, для которых правильный набор авторов определяется лексикографическим порядком. Результаты расчета для различного количества авторов приведены в таблице 1. Из данных, приведенных в таблице, можно сделать вывод, что в большинстве случаев основным автором является тот, который указан в библиографическом описании первым. Для учета этого факта была разработана формула расчета веса ребер с учетом позиции автора в библиографическом описании статьи. Вес автора для каждой статьи определяется как 1/2+1/(2K) для первого автора и 1/(2K) для остальных соаторов, где K – количество соавторов в статье. Степень связи по заданному автору для двух журналов определяется как минимум из максимумов его весов по подмножествам статей в каждом из журналов. Окончательный вес ребра связи между двумя журналами может быть рассчитан как сумма степеней их связи по всем авторам.

Количество авторов	L
2	24%
3	16%
4	9%
5	6%
6	6%
7	3%

Таблица 1. Процент статей с лексикографическим порядком.

3. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

При выборе языка для программной реализации алгоритма учитывались такие особенности алгоритма, как большой объем обрабатываемых данных, необходимость быстрого доступа к хранящимся в СУБД данным, небольшие требования к объемам памяти для создания временных структур данных и отсутствие необходимости вести диалог с пользователем. Учитывая эти требования, для реализации был выбран язык PL/SQL. Расчет тематической близости между журналами производится с заданными интервалами времени и сохраняется в таблицы СУБД.

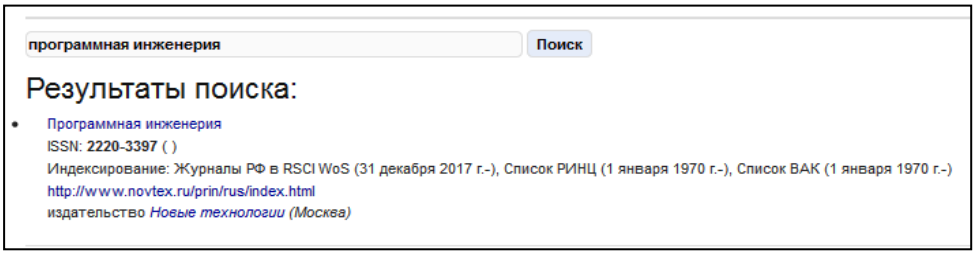


Рис. 1. Интерфейс контекстного поиска журналов

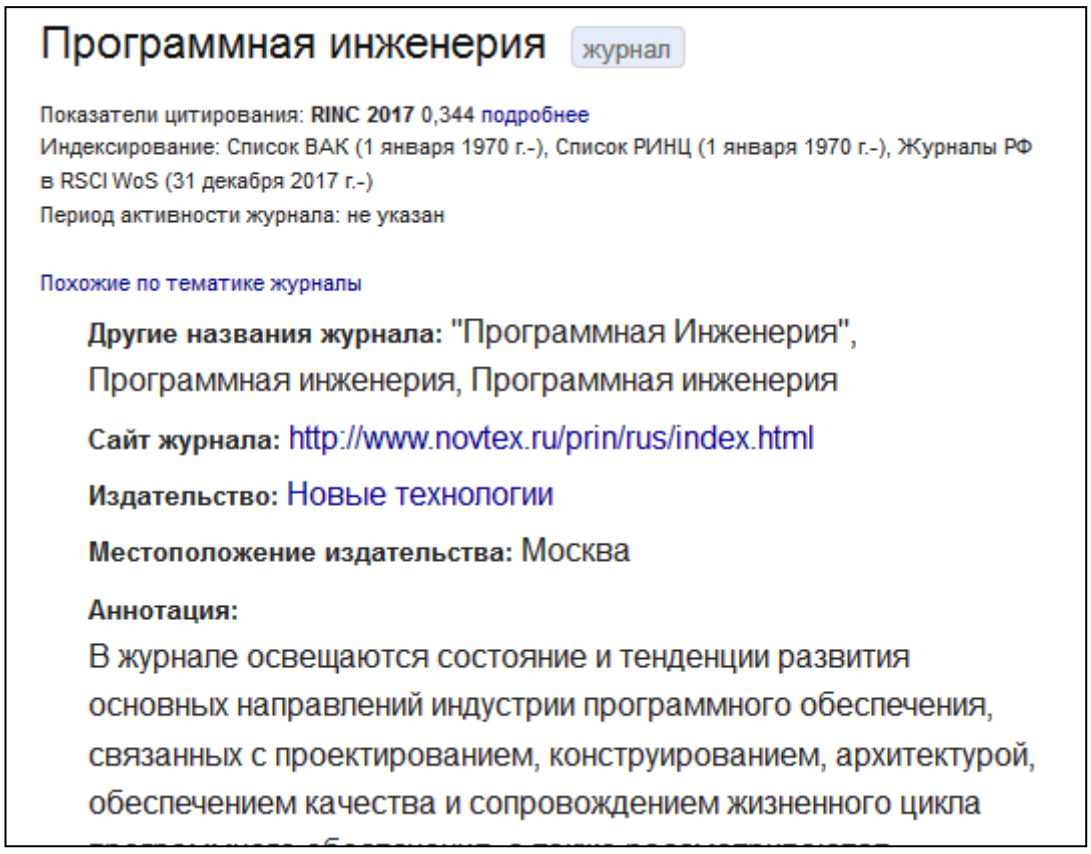


Рис. 2. Карточка журнала

В разработанном для этих целей интерфейсе [7] пользователь может выбрать один журнал, близкий ему по тематике (рис. 1, 2), и система автоматически сформирует подборку журналов, которые могут представлять интерес для пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей (рис. 3). Веб-интерфейс реализован с использованием открытой библиотеки DataTables [8]. В информационную карточку каждого журнала добавлена ссылка для перехода к таблице со списком тематически похожих журналов. В этой таблице указываются названия близких по тематике журналов и меры сходства. Кроме того, для возможности быстрой оценки

авторитетности каждого журнала из списка в таблице приводятся данные о количестве публикаций в этом журнале за 5 лет (зарегистрированных в системе «ИСТИНА»), а также данные Web of Science и РИНЦ. С целью удобной навигации по графу близости журналов в разработанном интерфейсе также реализована возможность перехода по ссылкам на список похожих журналов непосредственно из каждого элемента списка. Средствами библиотеки DataTables для быстрого поиска по названиям журналов реализован механизм быстрой фильтрации по части названия журнала.

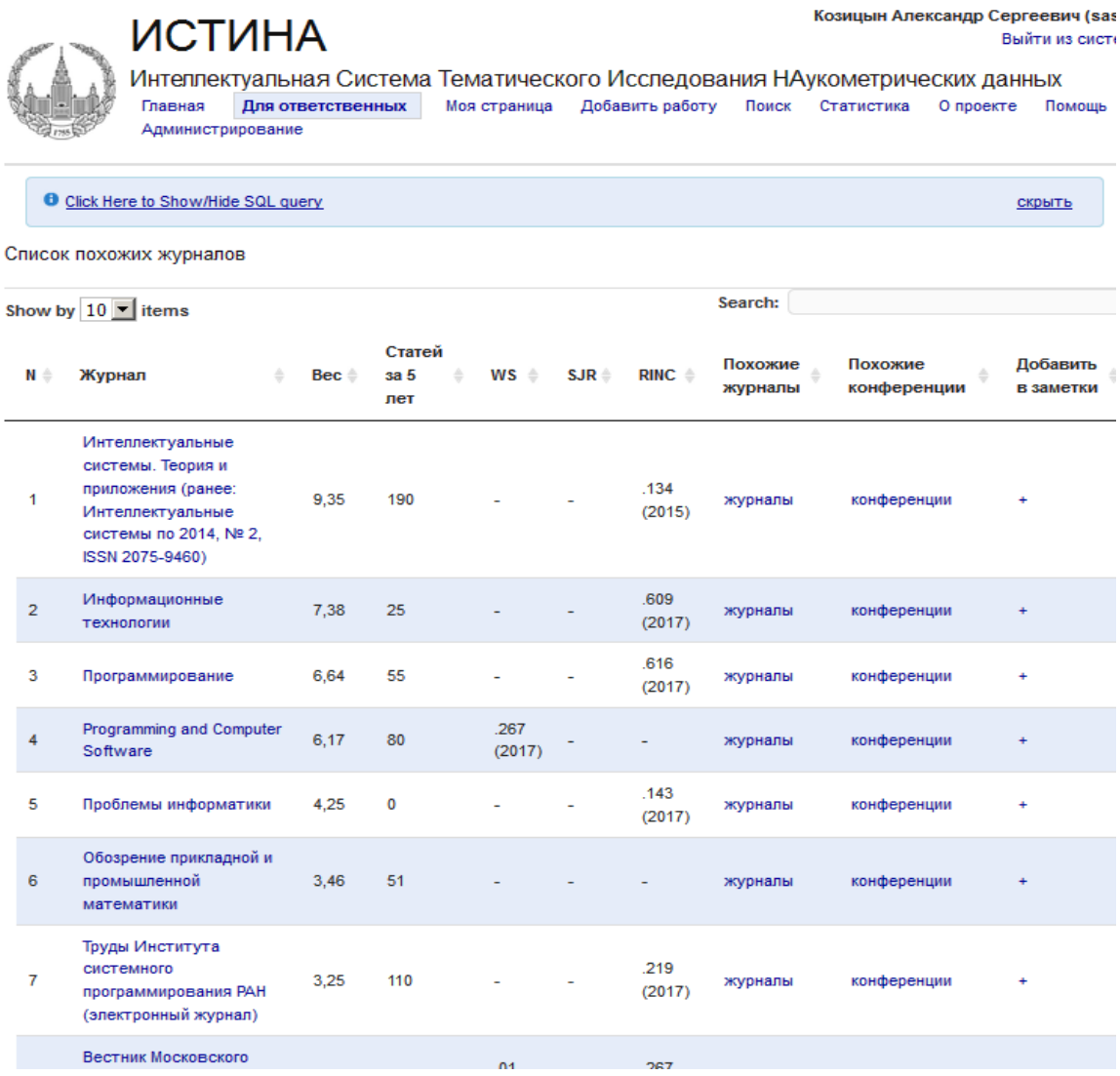


Рис. 3. Поиск похожих по тематике журналов

Для удобства работы пользователя предусмотрена возможность добавлять выбранный журнал в заметки, которые впоследствии можно просматривать, редактировать, а также использовать при последующем поиске. Дополнительно предоставляется возможность подбора похожих по тематике конференций (рис.

4).

Список похожих на журнал конференций

Show by 10 items Search:

N	Конференция	Вес	Количество докладов	Похожие конференции	Похожие журналы
1	Ломоносовские чтения - 2018. Секция "Механика"(2018)	2,54	129	конференции	журналы
2	Ломоносовские чтения-2018, секция "Вычислительная математика и кибернетика"(2018)	1,42	134	конференции	журналы
3	Знания-Онтологии-Теории (ЗОНТ-2017)(2017)	1,17	5	конференции	журналы
4	Знания-Онтологии-Теории (ЗОНТ-2019)(2019)	1,13	3	конференции	журналы
5	«Ломоносовские чтения - 2019». Секция «ВМК»(2019)	1,04	111	конференции	журналы
6	Научный сервис в сети Интернет 2019(2019)	0,95	5	конференции	журналы
7	«Modern Network Technologies, MoNeTec-2018»:(2018)	0,92	7	конференции	журналы
8	2018 Annual International Conference on Biologically Inspired Cognitive Architectures Ninth Annual Meeting of the BICA Society, Ninth Annual Meeting of the BICA Society(2018)	0,84	6	конференции	журналы

Рис. 4. Поиск похожих по тематике конференций

Тестирование разработанной программной реализации алгоритма проводилось по следующей методике. Из полученных результатов случайным образом было выбрано 200 пар связей журналов. Экспертами была проведена ручная оценка совпадения тематик журналов с простановкой баллов (2 – точная; 1 – не совсем точная; 0 – ошибочная). Общая сумма баллов делилась на удвоенное количество анализируемых связей. Оценка точности по этой методике составила 78%.

В качестве примера ошибок алгоритма можно привести, например, список журналов, которые определены как близкие по тематике к изданию «Труды Высшей школы Министерства внутренних дел СССР»: «Философские науки»; «Логические исследования»; «Известия МГТУ МАМИ»; «Логико-философские исследования»; «Вестник Московского университета. Серия 7: Философия». Такие ошибки могут возникать как следствие слишком широкой тематической области принимаемых в журнал статей.

ЗАКЛЮЧЕНИЕ

Алгоритм, описанный в настоящей работе, позволяет автоматически оценивать степень тематической близости научных журналов на основе библиографического описания статей и без использования полнотекстовых версий статей. Следует отметить, что алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации.

В перспективе разработанный алгоритм можно адаптировать для поиска похожих по тематике конференций, сборников публикаций и научных проектов. Наличие такого инструмента увеличит публикационную активность молодых сотрудников, повысит цитируемость статей и цитируемость между журналами.

Результаты работы алгоритма определения тематической близости между журналами, сборниками, конференциями и научными проектами также могут использоваться для построения правил в моделях разграничения доступа к данным на основе онтологий предметной области [9].

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-07-01055.

СПИСОК ЛИТЕРАТУРЫ

1. *Садовничий В.А., Васенин В.А.* Интеллектуальная система тематического исследования наукометрических данных: предпосылки создания и методология разработки. Часть 1 // Программная инженерия. 2018. Т. 9. № 2. С. 51–58.

2. Воронцов К.В. Лекции по логическим алгоритмам классификации URL:http://www.ccas.ru/voron/download/LogicAlgs.pdf

3. Шундеев А. С. Об изменении размерности векторного представления текстовых данных. Программная инженерия, , 2019 Т. 10. № 6. С. 265-273.

4. *Бурлаева Е.И. Павлыш В.Н.* Анализ методов преобразования текстов в форму объектов векторного пространства//Программная инженерия. 2019, Т. 10. № 1. с.30-37.

5. *Трофимов И.В.* Морфологический анализ русского языка: обзор прикладного характера//Программная инженерия. 2019, Т. 10. № 9. с. 391–399

6. Vasenin V., Lunev K., Afonin S., Shachnev D. Methods for intelligent data analysis based on keywords and implicit relations: The case of "istina" data analysis

system//In Actual Problems of Systems and Software Engineering – APSSE 2019, IEEE Conference Proceedings, pages 151-155, United States, 2019

7. ИАС ИСТИНА. URL: https://istina.msu.ru.
8. Библиотека datatables. URL: https://datatables.net/
9. *Afonin S.* Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6. doi: 10.1109/RPC.2018.8482178.

DETERMINING THE THEMATIC PROXIMITY OF SCIENTIFIC JOURNALS AND CONFERENCES USING BIG DATA TECHNOLOGIES

A. S. Kozitsin, S. A. Afonin, D. A. Shachnev

Institute of Mechcanics Lomonosov Moscow State University, Moscow

alexanderkz@mail.ru, serg@msu.ru, mitya57@gmail.com

Abstract

The number of scientific journals published in the world is very large. In this regard, it is necessary to create software tools that will allow analyzing thematic links of journals. The algorithm presented in this paper uses graphs of co-authorship for analyzing the thematic proximity of journals. It is insensitive to the language of the journal and can find similar journals in different languages. This task is difficult for algorithms based on the analysis of full-text information. Approbation of the algorithm was carried out in the scientometric system IAS ISTINA. Using a special interface, a user can select one interesting journal. Then the system will automatically generate a selection of journals that may be of interest to the user. In the future, the developed algorithm can be adapted to search for similar conferences, collections of publications and research projects. The use of such tools will increase the publication activity of young employees, increase the citation of articles and quoting between journals. In addition, the results of the algorithm for determining thematic proximity between journals, collections, conferences and research projects can be used to build rules in the ontology models for access control systems.

Keywords*: thematic classification, bibliographic data, graph of co-authorship, Information Systems*

REFERENCES

1. *Sadovnichii V.A., Vasenin V.A.* Intellektualnaia sistema tematicheskogo issledovaniia naukometricheskikh dannykh: predposylki sozdaniia i metodologiia razrabotki. Chast 1 // Programmnaia inzheneriia. 2018. T. 9. No 2. P. 51–58.

2. Voroncov K.V. Lekcii po logicheskim algoritmam klassifikacii. URL:http://www.ccas.ru/voron/download/LogicAlgs.pdf.

3. Shundeev A.S., Ob izmenenii razmernosti vectornogo predstavlenija tekstovyh dannyh// Programmnaia inzheneriia. 2019, T.10. No 6. p.265-273.

4. Burlaeva E.I., Pavlysh V.N., Analiz metodov preobrazovaniya tekstov v formu obyektov vektornogo prostarnstva// Programmnaia inzheneriia. 2019. T. 10. No 1. S. 30–37.

5. Trofimov I.V. Morfologicheskiy analiz russkogo yazyka: obzor prikladnogo haraktera// Programmnaia inzheneriia. 2019. T. 10. No 9. S. 391–399.

6. Vasenin V., Lunev K., Afonin S., Shachnev D. Methods for intelligent data aalysis based on keywords and implicit relations: The case of "istina" data analysis system//In Actual Problems of Systems and Software Engineering - APSSE 2019, IEEE Conference Proceedings, pages 151–155, United States, 2019.

7. IAS ISTINA. URL: https://istina.msu.ru.

8. Datatables. URL: https://datatables.net/

9. *Afonin S.* Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6. doi: 10.1109/RPC.2018.8482178

СВЕДЕНИЯ ОБ АВТОРАХ

КОЗИЦЫН Александр Сергеевич – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационного поиска и баз данных.

Alexander Sergeevich KOZITSIN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru, ORCID: 0000-0002-8065-9061

АФОНИН Сергей Александрович – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области регулярных языков и информационных систем.

Sergey Alexandrovich AFONIN – Leading Researcher, Ph. D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of regular languages and information systems.

email: serg@msu.ru, ORCID:0000-0003-3058-9269

Шачнев Дмитрий Алексеевич – программист, окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационных систем.

Dmitiy Alekseevich SHACHNEV – programmer, graduated from M.V. Lomonosov Moscow State University. Specialist in information systems.

email: mitya57@gmail.com, ORCID: 0000-0002-5940-9180

Материал поступил в редакцию 12 ноября 2019 года

УДК 519.178: 004.738.5

СИЛЬНЫЕ И СЛАБЫЕ СВЯЗИ В НАУЧНО-ОБРАЗОВАТЕЛЬНОМ ВЕБЕ

А. А. Печников

Институт прикладных математических исследований — обособленное подразделение ФИЦ «Карельский научный центр Российской академии наук», г. Петрозаводск

pechnikov@krc.karelia.ru

Аннотация

Веб-граф является наиболее популярной моделью фрагментов реального Веба, применяемой в науке о Вебе. Исследование сообществ в веб-графе способствует лучшему пониманию организации фрагмента Веба и процессов, происходящих в нём. Предложено выделить в веб-графе коммуникационный граф, содержащий только те вершины (и дуги между ними), которые имеют встречные дуги, и в нём исследовать задачу разбиения на сообщества. По аналогии с социальными исследованиями связи, реализуемые через ребра в коммуникационном графе, предложено называть «сильными», а все остальные – «слабыми». На сильных связях строятся тематические сообщества, имеющие содержательную интерпретацию. В то же время слабые связи способствуют коммуникациям между сайтами, не имеющими общих признаков по сфере деятельности, географии, подчиненности и т. д., и в основном сохраняют связность фрагментов Веба даже при отсутствии сильных связей. Эксперименты, проведенные для фрагмента научно-образовательного Веба России, показали возможность содержательной интерпретации полученных результатов и перспективность такого подхода.

***Ключевые слова**: веб-граф, коммуникационный граф, сообщество в графе, сила связей*

ВВЕДЕНИЕ

Исследование графов реального (и виртуального) мира является важной задачей во многих областях, таких, как биология, социология, социальные сети,

вебометрика и многие другие, поскольку позволяют понять структуру объектов и проанализировать их свойства.

Например, как говорится в [1], социальная сеть – это совокупность людей, каждый из которых знаком с некоторым подмножеством других. Социальные сети были предметом как эмпирического, так и теоретического изучения в социальных науках в течение, по крайней мере, пятидесяти лет, отчасти из-за присущего им интереса в паттернах человеческого взаимодействия, но и потому, что их структура имеет важное значение для распространения информации и болезней. Ясно, например, что изменение только среднего числа знакомств, которые имеют люди (также называемое средней степенью сети), может существенно повлиять на распространение слухов, моды, шутки или гриппа в этом году. Социальная сеть может быть представлена в виде набора точек (или вершин), обозначающих людей, соединенных попарно линиями (или ребрами), обозначающими знакомых. Можно, в принципе, построить социальную сеть для компании или фирмы, школы или университета, любого другого сообщества, вплоть до всего мира. Обратим внимание на то, что в такой постановке речь идет о неориентированных сетях.

Примером такой неориентированной сети может являться сеть ученых, построенная по принципу соавторства: два ученых считаются связанными, если они совместно написали статью. Как сказано в [1]: «… это кажется разумным определением научного знакомства: большинство людей, которые написали статью вместе, будут хорошо знать друг друга. Это – умеренно строгое определение, так как есть много ученых, которые знают друг друга до некоторой степени, но никогда не сотрудничали в написании статьи. Строгость, однако, не является по своей сути плохой вещью. Строгое условие знакомства вполне приемлемо при условии, как и в данном случае, что его можно применять последовательно». В таких сетях «… обнаружены “маленькие миры”, … наличие кластеризации и ряд очевидных различий в моделях сотрудничества между различными областями знаний» [1].

А что будет, если в качестве принципа установления связи между учеными принять не соавторство, а ссылку, сделанную в работе учёного *A* на работу учёного *B*? В этом случае мы получим ориентированную сеть. Конечно, при этом возникает большое количество дополнительных вопросов, первым из которых

будет временной интервал, в течение которого были сделаны ссылки. Наверняка будут учёные, жившие 100 лет назад и являющиеся классиками в некоторой области знаний, на работы которых сделано очень много ссылок, по понятным соображениям из такого исследования выпадут. К сожалению, у автора на сегодня не хватает объективных данных для проведения подобного исследования.

Более понятными и исследованными ориентированными сетями являются фрагменты Веба. Достаточно давно известно, что вузовские и академические фрагменты Веба как России, так и других стран [2–4] обладают достаточно специфическими свойствами, характеризующими их структуру. В частности, в соответствующем веб-графе имеются большая компонента сильной связности и значительное количество «висячих» сайтов (имеющих либо только ссылки, сделанные с них, либо, что реже, ссылки, сделанные только на них).

Компонента сильной связности сама по себе является интересным объектом исследований, позволяющим установить, например, наличие или отсутствие свойства «малого мира» [5], которая, однако, мало что дает в объяснении процессов возникновения гиперссылок между сайтами фрагментов Веба. В этом смысле гораздо больше пищи для размышлений дают такие структурные элементы графа, как сообщества сайтов, когда «внутри» сообществ сделано больше ссылок, чем между сообществами. Но, опять-таки, попытки разбиения вершин, составляющих максимальную компоненту связности веб-графа на сообщества, приводят к сложно интерпретируемым результатам.

Основная идея, рассматриваемая в данной работе, заключается в том, чтобы построить некий «жесткий каркас» для фрагмента Веба, оставив только те сайты, которые имеют встречные гиперссылки, и уже на этом «каркасе» проверить свойства разбиения на сообщества (такой каркас далее будем называть коммуникационным графом). И уже далее с помощью известных алгоритмов исследовать вопрос о структуре сообществ вершин коммуникационного графа и «слабого» веб-графа, из которого удален коммуникационный граф. В качестве реального объекта исследований взят научно-образовательный фрагмент Веба, для которого дана содержательная интерпретация полученных результатов.

1. ИСПОЛЬЗУЕМЫЕ ПОНЯТИЯ, МЕТОДЫ И ИНСТРУМЕНТЫ

Целевое множество сайтов задается прямым перечислением их доменных имен, и в результате сканирования находятся все связывающие их гиперссылки. В случае, когда сайты относятся к одному виду деятельности, такое множество называется тематическим. Соответственно, (тематический) фрагмент Веба – это целевое множество сайтов и множество связывающих их гиперссылок [4].

Веб-граф фрагмента Веба – это ориентированный граф без петель и кратных дуг, множество вершин которого представлено целевым множеством сайтов, а множество дуг строится следующим образом: две вершины связаны дугой, если есть хотя бы одна гиперссылка, связывающая соответствующие сайты.

Сообщество (кластер, модуль, группа) графа неформально можно определить как множество вершин с бóльшим количеством дуг между собой, чем с остальными вершинами графа [6].

Более строго разбиение графа на сообщества можно определить через понятие модулярности. Модулярность – это свойство графа и некоторого разбиения его на подграфы (модули-сообщества). Мера модулярности показывает, насколько данное разбиение качественно в том смысле, что существует много дуг, лежащих внутри подграфов-сообществ, и мало дуг, лежащих вне подграфов (соединяющих сообщества между собой).

Меру модулярности Q можно задать следующей формулой [6]:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) * \delta(c_i, c_j).$$

Здесь m – количество дуг в графе, A_{ij} – элемент матрицы смежности графа, k_i, k_j – степени соответствующих вершин, $\delta(c_i, c_j)$ – символ Кронекера, вычисляемый по формуле:

$$\delta(c_i, c_j) = \begin{cases} 1 : if\ c_i = c_j \\ 0 : if\ c_i \neq c_j \end{cases}$$

, где c_i, c_j – метки класса соответствующих вершин.

Соответственно, наилучшим разбиением на сообщества считается разбиение, на котором достигается максимум Q.

Для сбора и анализа данных использовались программа для поиска и сбора внешних гиперссылок *BeeCrawler* [7] и базы данных внешних гиперссылок [8]. Для исследования веб-графов применялась открытая платформа для визуализа-

ции графов *Gephi* [9], в которой, среди многих, реализованы программы вычисления коэффициента модулярности и поиска сообществ в графе с использованием алгоритмов, предложенных в [10].

2. ФРАГМЕНТ НАУЧНО-ОБРАЗОВАТЕЛЬНОГО ВЕБА РОССИИ

Исходное целевое множество сайтов фрагмента научно-образовательного Веба России сформировано на конец 2017 года, то есть после присоединения Российской академии медицинских наук и Российской академии сельскохозяйственных наук (РАСХН) к РАН, но в процессе создания укрупненных структур типа федеральных исследовательских центров, а также укрупнения вузов. Такой «footprint» позволяет рассчитывать на то, что фрагмент Веба содержит достаточно устойчивые ссылки между сайтами, которые сформировались в течение нескольких предыдущих лет. Исходное целевое множество содержит 867 сайтов (279 университетов и 588 научных учреждений). Под научными учреждениями понимаются организации РАН от уровня институтов до региональных научных центров и отделений РАН (сайт собственно РАН www.ras.ru в целевое множество не включён, поскольку является мощным коммуникатором, существенно искажающим картину в свою пользу). Далее для краткости научные учреждения будем называть «институтами», понимая несколько расширительное толкование этого термина.

Сканирование всех 867 сайтов позволяет получить около 20000 гиперссылок, сделанных между этими сайтами, считая кратные ссылки. Заменим кратные ссылки на одинарные дуги и получим веб-граф, содержащий 867 вершин и 5030 дуг. Проверка на связность и сильную связность показала наличие 73 изолированных вершин и 236 висячих (36 вершин не имеют входящих дуг и 200 (!) – исходящих). В содержательном плане заметим, что 73 изолированные вершины в подавляющем большинстве относятся к сайтам институтов, ранее входившим в состав РАСХН.

Единственная компонента сильной связности (КСС) содержит 534 вершины и 4026 дуг. Веб-граф, равный максимальной КСС, имеет диаметр, равный 10, и коэффициент модулярности 0.398 [11], и при этом разбивается на 7 сообществ, содержащих от 40 до 139 вершин. Коэффициент модулярности говорит о невысоком стремлении сайтов организовываться в сообщества: значения Q принадлежат отрезку [–1, 1], а «хорошей» считается кластеризация при Q больше 0.6.

Тем не менее, одно из построенных сообществ имеет хорошее содержательное объяснение. В него входят 40 вершин, соответствующих 4 сайтам университетов и 36 сайтам институтов, и всего 97 дуг (что в среднем существенно меньше, чем в веб-графе КСС). Из 40 сайтов 35 принадлежат институтам нынешнего Отделения сельскохозяйственных наук РАН, а также Красноярскому научному центру СО РАН, Кемеровскому технологическому институту пищевой промышленности, Кубанскому государственному технологическому университету, Воронежскому государственному лесотехническому университету и Санкт-Петербургскому горному университету.

Возможные объяснения:

- Красноярский научный центр СО РАН на период исследования успел стать федеральным исследовательским центром, в состав которого входят 2 сельскохозяйственных института;
- Кемеровский технологический институт пищевой промышленности близок по роду деятельности к сельскому хозяйству;
- Кубанский государственный технологический университет имеет тесные связи с сельскохозяйственными институтами Кубани;
- Воронежский государственный лесотехнический университет на своем сайте имеет ссылку на сайт Центральной научной сельскохозяйственной библиотеки (которая попадает под термин «институты»);
- Санкт-Петербургский горный университет имеет на своем сайте ссылку на сайт Всероссийского института генетических ресурсов растений имени Н.И. Вавилова, где расположен раздел журнала «АПК», выписываемый библиотекой университета, и ещё одну ссылку на сайт Центральной научной сельскохозяйственной библиотеки.

Не считая последних двух случаев, можно сказать, что данное сообщество имеет ярко выраженную сельскохозяйственно-агропромышленную тематику.

3. КОММУНИКАЦИОННЫЙ ГРАФ ВЕБ-ГРАФА

Ньюман и соавторы при решении задачи о разбиении графа на сообщества избегают ориентированности графа, когда речь идет о веб-графах. Цитируя [11, с. 026113-5]: «… Некоторые сети являются ориентированными, т. е. их ребра работают только одном только направлении. Веб является таким примером; ссыл-

ки в Вебе указывают в одном направлении, только от одной веб-страницы до другой. … Однако мы обнаружили, что во многих случаях лучше игнорировать направленный характер сети при нахождении структура сообществ. Часто ребро действует просто как указание на связь между двумя узлами, и направление не имеет значения». А если направление (ориентация) имеет значение? Посмотрим рисунок 1, где изображено сельскохозяйственно-агропромышленное сообщество.

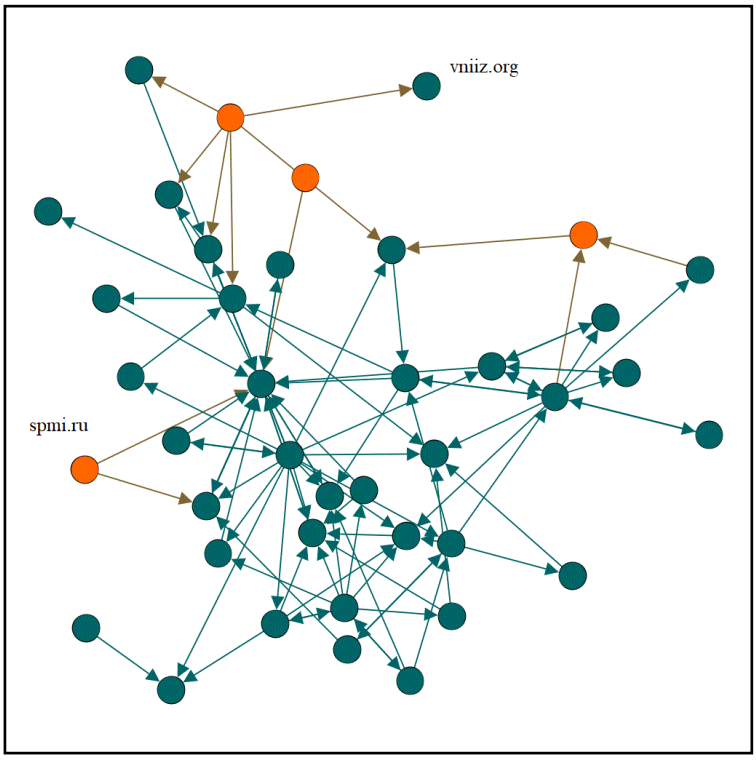


Рис. 1. Граф «сельскохозяйственно-агропромышленного сообщества»

Возникают вопросы: можно ли сайт Всероссийского НИИ зерна и продуктов его переработки (вершина *vniiz.org*) считать полноправным участником сообщества, хотя он не имеет ни одной ссылки на другие сайты? И можно ли считать участником сообщества сайт Санкт-Петербургского горного университета (*spmi.ru*), когда на него нет ни одной ссылки с других участников? Понятно, что обе этих вершины попали в данное сообщество в том числе по причине того, что ориентация дуг не учитывалась.

Термин «сообщество» (*Gemeinschaft*) возник в германской социологии в конце XIX века, и подразумевает совместную деятельность его участников, имеющих общие цели [12], и в этом контексте ответ на сформулированные вопросы должен быть отрицательным.

В работе [13] приведен аналитический обзор различных подходов кластеризации и нахождения сообществ в ориентированных сетях, основанных «… на методологических принципах и алгоритмических подходах. Насколько нам известно, это первое предложение по классификационной схеме методов кластеризации графов для направленных сетей».

Далее приведем предлагаемую классификацию в соответствии с [13]:

* *Наивный подход к преобразованию графа*: алгоритмы, которые игнорируют направленность ребер и рассматривают граф как неориентированный. Из-за наивного преобразования графа базовая семантика графа не сохраняется, и полезная информация не учитывается при выполнении задачи кластеризации. Это мы наблюдаем на рисунке 1.

* *Преобразования, поддерживающие направленность:* ориентированный граф преобразуется в неориентированный граф специального вида, например, двудольный, и направление дуги осмысленно поддерживается в создаваемой сети, а затем к двудольному графу применяются соответствующие алгоритмы кластеризации.

* *Распространение целевых функций и методологий кластеризации на ориентированные сети:* эта категория включает подходы, которые представляют собой расширение методологий из неориентированных случаев. Таким образом, объективные критерии (типа приведенного ранее коэффициента кластеризации *Q*) расширяются в соответствии с требованиями задачи. Задача кластеризации графов обычно выражается как задача оптимизации, где объективный критерий, фиксирующий требуемые свойства кластера, оптимизируется путем перераспределения узлов в кластеры. Естественным способом решения ориентированной версии задачи является расширение этих мер для ориентированных сетей, где направленность ребер рассматривается как неотъемлемая сетевая характеристика.

* *Альтернативные подходы*: эта категория включает подходы, которые следуют различным методологиям, главным образом отличным от тех, которые описаны в предыдущих трех категориях. Выделяются три основных типа методов, а именно:

(i) теоретико-информационные,

(ii) методы, основанные на вероятностных моделях и статистическом выводе, и

(iii) стохастические методы блочного моделирования.

Метод, который будет изложен далее, можно отнести к теоретико-информационным. Рассмотрим неориентированный граф, который по построению подразумевает «совместную деятельность», а именно, встречные гиперссылки. Коммуникационный граф веб-графа – это неориентированный граф, имеющий то же самое множество вершин, что и веб-граф, а множество его рёбер строится по следующему правилу: ребро (*i,j*) принадлежит множеству ребер коммуникационного графа тогда и только тогда, когда в веб-графе существуют дуги (*i,j*) и (*j, i*).

Коммуникационный граф, построенный таким образом, может иметь несколько компонент связности и/или изолированные вершины. В этом случае мы исключаем изолированные вершины (поскольку они не влияют на связность) и изучаем компоненты связности каждую по отдельности, начиная с максимальной. Коммуникационный граф веб-графа научно-образовательного Веба содержит 313 вершин и 468 ребер (рис. 2). Из 313 вершин 67 относятся к университетам, а 246 – к институтам, то есть доля университетов в коммуникационном графе сократилась на одну десятую. Раскраска вершин дана в соответствии с полученным разбиением на сообщества.

Коэффициент модулярности данного разбиения равен 0.695, то есть достаточно высок. Построенные 13 сообществ содержат от 7 до 51 вершины, из них 4 можно достаточно точно идентифицировать по одному из двух признаков: география или научное направление.

Некоторым из построенных сообществ можно дать содержательную (тематическую) интерпретацию.

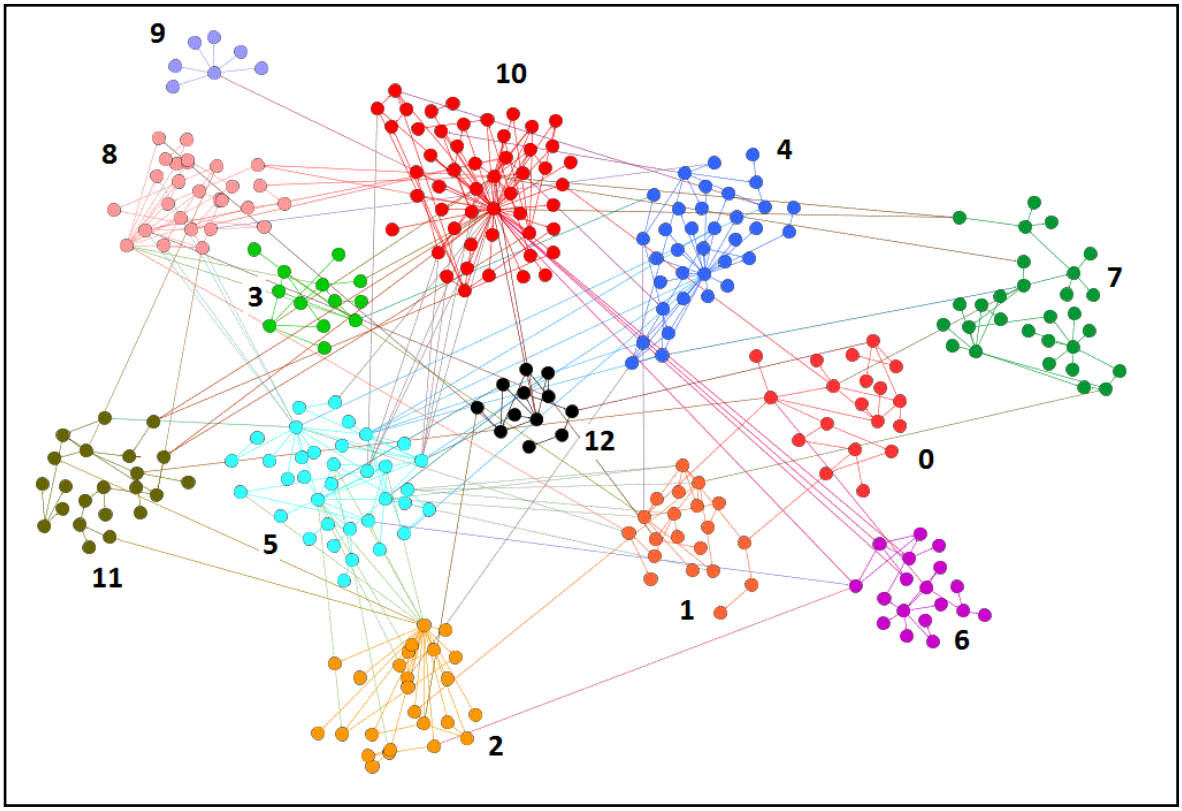


Рис. 2. Коммуникационный граф научно-образовательного Веба

Четко выделяются 2 «географических» сообщества:

№2 – «Урал»: 1 университет (Пермский политехнический), 23 института УрО РАН и примкнувший к ним Научный центр волоконной оптики РАН из Москвы;

№10 – «Сибирь»: 4 сибирских университета, 47 сибирских институтов и 1 институт из Севастополя.

«Гуманитарное» сообщество №12 содержит 10 институтов (археология, этнография, лингвистические исследования, языкознание, история, мировая литература, славяноведение, философия) и два университета (Алтайский и Горно-Алтайский).

«Медицинское» сообщество №9 представляет собой «пучок» сайтов медицинских институтов Сибири и Дальнего востока, связанных с сайтом Сибирского отделения медицинских наук, но не связанных между собой.

Участники любого сообщества коммуникационного графа присутствуют в веб-графе в силу построения, хотя не обязательно, чтобы все участники из одного сообщества коммуникационного графа входили в одно и то же сообщество

веб-графа (в нашем случае это верно для значительной части сообществ коммуникационного графа). Обратное неверно: сайты-участники агропромышленного сообщества веб-графа полностью отсутствуют в коммуникационном графе.

Теперь удалим из веб-графа КСС все пары встречных дуг, соответствующие ребрам коммуникационного графа. Полученный «слабый» веб-граф достаточно велик – 528 вершин и 3090 дуг, но имеет низкий коэффициент модулярности 0,356 и разбивается на 8 сообществ, не имеющих убедительной содержательной интерпретации.

Найдем максимальную КСС «слабого» веб-графа, удалим все вершины, не вошедшие в КСС, и получим КСС «слабый» веб-граф с 461 вершиной и 2744 дугами. Опять-таки, сохраняется низкий коэффициент модулярности.

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ. СИЛЬНЫЕ И СЛАБЫЕ СВЯЗИ

Сравним динамику изменений доли институтов и университетов в рассмотренных графах, для чего сведем данные в таблицу 1.

Таблица 1. Доля сайтов институтов и университетов в графах

Граф	*Кол-во вершин*	*Кол-во дуг (ребер)*	*Доля институтов*	*Доля университетов*	*Диаметр*	*Средняя длина пути*
Начальный веб-граф	867	5030	0,68	0,32	-	-
КСС веб-граф	534	4026	0,67	0,33	10	3,6
Коммуникационный граф	313	936 (468)	0,78	0,22	9	3,45
«Слабый» веб-граф	528	3090	0,66	0,34	-	-
КСС «слабый» веб-граф	461	2744	0,63	0,37	11	4,2

Более 45 лет назад Грановеттером [14] была предложена концепция силы социальных связей. Все связи разделяются на две категории – сильные и слабые – с целью формализации межличностных отношений на основе длительности и частоты контактов. Например, сильная связь присуща друзьям, а слабая – соседям. Попробуем использовать понятие силы связи как аналогию для сайтов, хотя изначально понятно, что аналогия далеко не полная, поскольку Грановеттер считает, что связи симметричны.

Возьмем пару вершин *i*, *j* ориентированного графа. Если в графе существуют дуги (*i,j*) и (*j,i*), то будем говорить о сильной связи вершин *i, j*. Если же для вершин *i, j* существует только одна из дуг (*i,j*) или (*j,i*), то будем говорить о слабой связи вершин *i* и *j*.

Начальный научно-образовательный веб-граф, как и КСС веб-граф, содержит сильные и слабые дуги, в коммуникационном графе отсутствуют слабые дуги, а в «слабом» веб-графе и КСС «слабом» веб-графе отсутствуют сильные дуги. Из табл. 1 видно, что соотношение институты/университеты практически одно и то же для всех графов, имеющих слабые или сильные и слабые дуги, но резко изменяется для графа, содержащего только сильные дуги. Можно сделать вывод о том, что сильные связи более присущи институтам, чем университетам.

В [14, с. 1362] отмечается, что «… чем сильнее связи, тем больше похожи индивиды друг на друга в разных аспектах». Это подтверждается в случае коммуникационного графа научно-образовательного Веба. Можно сказать, что сильные связи способствуют появлению устойчивых тематически «похожих» сообществ. И хотя таких сообществ немного, всего 4 из 13, они имеют четкую содержательную интерпретацию.

Однако, если коммуникационный граф «нарастить» до КСС веб-графа (а тем более, до начального веб-графа), то оказывается, что ни одно из четырех сообществ коммуникационного графа не только не явилось основой для сообществ в веб-графах, но даже их участники попали в разные сообщества. Похоже, что слабые ссылки ведут к «размыванию» сообществ.

Правда, в КСС веб-графе мы обнаружили «сельскохозяйственно-агропромышленное» сообщество. Но вспомним, что в начальном веб-графе были найдены 73 изолированные вершины, также относящиеся к сельскохозяйственным институтам. Связывая эти два результата, можно дать следующее объяснение такого исключения: сайты бывшей РАСХН просто не успели установить ссылки с другими сайтами целевого множества. Поэтому те сайты РАСХН, которые были как-то связаны между собой, имели мало ссылок «вовне» и организовали сообщество, а остальные остались изолированными.

Какова же роль слабых ссылок? По Грановеттеру, в социологии слабые связи можно охарактеризовать как систему нерегулярных контактов, не охватывающих друзей индивида, а выходящих на представителей других тесно связан-

ных групп, в которых он не состоит. Из этого следует, что индивид может устанавливать связи с людьми из значительного числа взаимно непересекающихся групп, при этом не являясь членом каждой из них. В этом смысле слабые связи играют роль «мостов» в графе. Напомним, что мост – это ребро (неориентированного) графа, удаление которого увеличивает число компонент связности [15]. Грановеттер в [14] говорит, что сильные связи почти никогда не являются мостами, как правило, мостами являются именно слабые связи.

В нашем случае косвенным подтверждением этого факта являются диаметр и средняя длина пути, которые очень близки в КСС «слабом» веб-графе, КСС веб-графе и коммуникационном графе (табл. 1). Получается, что удаление сильных связей почти не сказывается на диаметре графов, значит, дуги, соответствующие им, чаще всего не являются мостами.

Отсюда следует вывод, что слабые связи фрагмента Веба служат для установления контактов сайтов из непересекающихся групп, что очень важно в смысле получения новой информации, не лежащей в круге интересов данной группы. Поэтому для поиска сообществ, характеризующихся сильными объединяющими признаками (география, сфера научной деятельности), следует использовать коммуникационный граф. Для поиска сообществ со слабыми признаками (вроде междисциплинарных сообществ) следует удалить коммуникационный граф и искать сообщества в полученном графе.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 18-07-00628-а.

СПИСОК ЛИТЕРАТУРЫ

1. *Newman M.E.J.* The structure of scientific collaboration networks // Proceedings of the National Academy of Sciences of the USA. 2001. No 98 (2). P. 404–409. https://doi.org/10.1073/pnas.98.2.404

2. *Thelwall M., Wilkinson D.* Graph structure in three national academic Webs: Power laws with anomalies // Journal of the American Society for Information Science and Technology. 2003. №54(8). P. 706–712.

3. *Ortega J.L., Aguillo I.F.* Visualization of the Nordic academic web: Link analysis using social network tools // Information Processing and Management. 2008. Vol. 44, Iss. 4. P. 1624–1633.

4. *Печников А.А.* Методы исследования регламентируемых тематических фрагментов Web // Труды Института системного анализа Российской академии наук. Серия: Прикладные проблемы управления макросистемами. 2010. Т. 59. С. 134–145.

5. *Watts D.J.; Strogatz S.H.* Collective dynamics of 'small-world' networks // Nature. No 393. P. 440–442.

6. *Ермолин Н.А., Мазалов В.В., Печников А.А.* Теоретико-игровые методы нахождения сообществ в академическом Вебе // Труды СПИИРАН. 2017. Вып. 55. С. 237–254.

7. *Pechnikov A.A., Chernobrovkin D.I.* Adaptive Crawler for External Hyperlinks Search and Acquisition // Automation and Remote Control. 2014. V. 75. No 3. P. 587–593.

8. *Головин А.С., Печников А.А.* База данных внешних гиперссылок для исследования фрагментов Веба // Информационная среда вуза XXI века: материалы VII Всероссийской научно-практической конференции (23–27 сентября 2013 г.). Петрозаводск, 2013. С. 55–57.

9. The Open Graph Viz Platform. URL: https://gephi.org

10. *Blondel V.D., Guillaume J-L., Lambiotte R., Lefebvre E.* Fast unfolding of communities in large networks // J. of Statistical Mechanics: Theory and Experiment. 2008. P. 10008.

11. *Newman M.E., Girvan M.* Finding and evaluating community structure in networks // Physical Review E. 2004. V. 69(2). P. 026113.

12. *Левкина Л.И.* Социально-историческая роль сообществ. М.: Русайнс, 2016. 216 с.

13. *Malliaros F.D., Vazirgiannis M.* Clustering and community detection in directed networks: A survey // Physics Reports. 2013. V. 533. Issue 4. P. 95–142.

14. *Granovetter M.S.* The Strength of Weak Ties // The American J. of Sociology. 1973. No 78 (6). P. 1360–1380.

15. *Харари Ф.* Теория графов. М.: Мир, 1973. 300 с.

STRONG AND WEAK RELATIONS IN THE ACADEMIC WEB

A.A. Pechnikov

Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Petrozavodsk

pechnikov@krc.karelia.ru

Abstract

The web graph is the most popular model of real Web fragments used in Web science. The study of communities in the web graph contributes to a better understanding of the organization of the fragment of the Web and the processes occurring in it. It is proposed to allocate a communication graph in a web graph containing only those vertices (and arcs between them) that have counter arcs, and in it to investigate the problem of splitting into communities. By analogy with social studies, connections realized through edges in a communication graph are proposed to be called "strong" and all others "weak". Thematic communities with meaningful interpretations are built on strong connections. At the same time, weak links facilitate communication between sites that do not have common features in the field of activity, geography, subordination, etc., and basically preserve the coherence of the fragments of the Web even in the absence of strong links. Experiments conducted for a fragment of the scientific and educational Web of Russia show the possibility of meaningful interpretation of the results and the prospects of such an approach.

Keywords*: web graph, communication graph, community in graph, strength of the linkages*

REFERENCES

1. *Newman M.E.J.* The structure of scientific collaboration networks // Proceedings of the National Academy of Sciences of the USA. 2001. No 98 (2). P. 404–409. https://doi.org/10.1073/pnas.98.2.404

2. *Thelwall M., Wilkinson D.* Graph structure in three national academic Webs: Power laws with anomalies // J. of the American Society for Information Science and Technology. 2003. No 54(8). P. 706–712.

3. *Ortega J.L., Aguillo I.F.* Visualization of the Nordic academic web: Link analysis using social network tools // Information Processing and Management. 2008. Vol. 44. Iss. 4. P. 1624–1633.

4. *Pechnikov A.A.* Metody issledovanija reglamentiruemyh tematicheskih fragmentov Web // Trudy Instituta sistemnogo analiza Rossiiskoi akademii nauk. Serija: Prikladnye problem upravlenija makrosistemami. 2010. T. 59. S. 134–145 (in Russian).

5. *Watts D.J.; Strogatz S.H.* Collective dynamics of 'small-world' networks // Nature. No 393. P. 440–442.

6. *Ermolin N.A., Mazalov V.V., Pechnikov A.A.* Teoretiko-igrovye metody nahojdenija soobschestv v akademicheskom Webe // Trudy SPIIRAN. 2017. Vyp. 55. S. 237–254 (in Russian).

7. *Pechnikov A.A., Chernobrovkin D.I.* Adaptive Crawler for External Hyperlinks Search and Acquisition // Automation and Remote Control. 2014. V. 75. No 3. P. 587–593.

8. *Golovin A.S., Pechnikov A.A.* Baza dannyh vneshnih giperssylok dlja issledovanija fragmentov Weba // Informacionnaja sreda vuza XXI veka: materialy VII Vserossiiskoi nauchno-prakticheskoi konferencii (23–27 sentjabrja 2013). Petrozavodsk, 2013. S. 55–57 (in Russian).

9. The Open Graph Viz Platform. URL: https://gephi.org

10. *Blondel V.D., Guillaume J-L., Lambiotte R., Lefebvre E.* Fast unfolding of communities in large networks // J. of Statistical Mechanics: Theory and Experiment. 2008. P. 10008.

11. *Newman M.E., Girvan M.* Finding and evaluating community structure in networks // Physical Review E. 2004. Vol. 69(2). P. 026113.

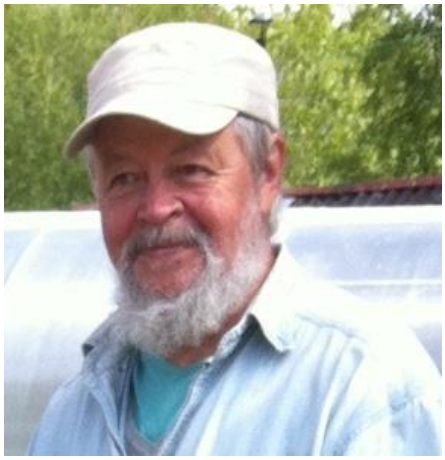
12. *Levkina L.I.* Social'no-istoricheskaja rol' soobschestv. M.: Rusains, 2016. 216 s. (in Russian).

13. *Malliaros F.D., Vazirgiannis M.* Clustering and community detection in directed networks: A survey // Physics Reports. 2013. V. 533. Issue 4. P. 95–142.

14. *Granovetter M.S.* The Strength of Weak Ties // The American J. of Sociology. 1973. No 78 (6). P. 1360–1380.

15. *Harary F.* Teorija grafov. M.: Mir, 1973. 300 s. (in Russian)

СВЕДЕНИЯ ОБ АВТОРЕ



ПЕЧНИКОВ Андрей Анатольевич – главный научный сотрудник Института прикладных математических исследований — обособленное подразделение ФИЦ «Карельский научный центр Российской академии наук». Сфера научных интересов – математическое моделирование, дискретная оптимизация, вебометрика.

Andrey Anatolievich PECHNIKOV – Chief Research Associate of the Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences. Research interests include mathematical modeling, discrete optimization, webometrics.

email: pechnikov@krc.karelia.ru

Материал поступил в редакцию 25 октября 2019 года

УДК 001

РИНЦ КАК ЗЕРКАЛО ПУБЛИКАЦИОННОЙ АКТИВНОСТИ ЧЛЕНОВ РАО

Ю. Е. Поляк

Федеральное государственное бюджетное учреждение науки Центральный экономико-математический институт Российской академии наук, г. Москва

polak@cemi.rssi.ru

Аннотация

На основе информации из открытых источников составлена таблица, отражающая показатели 128 действительных членов Российской академии образования в Российском индексе научного цитирования (РИНЦ). Основные результаты даны в сжатом виде и сопоставлены с итогами аналогичного исследования, выполненного несколькими годами ранее. Обсуждены сделанные выводы и особенности РИНЦ как аналитического инструмента.

***Ключевые слова**: Российская академия образования, Российский индекс научного цитирования, публикационная активность*

ВВЕДЕНИЕ

Одним из показателей эффективности деятельности научных работников является их публикационная активность. В 2009–2010 гг. сотрудниками Лаборатории вебометрики Института научной информации и мониторинга Российской академии образования (ИНИМ РАО) и Научной педагогической библиотеки им. К.Д. Ушинского проводились исследования публикационной активности членов РАО. В них участвовал и автор в качестве ведущего научного сотрудника ИНИМ. В 2019 г. он по собственной инициативе выполнил подобную работу в части данных, регистрируемых Российским индексом научного цитирования (РИНЦ). Результаты этой работы позволяют проследить динамику изменения показателей за 10 лет, а также обсудить возможности РИНЦ для оценки научной продуктивности.

О ПУБЛИКАЦИОННОЙ АКТИВНОСТИ ЧЛЕНОВ РАО 10 ЛЕТ НАЗАД

Методика и результаты исследования 2009–2010 гг. подробно описаны и в 2011 г. опубликованы в [2]. Объектами исследования стали работы здравство-

вавших на тот момент членов Российской академии образования – действительных членов и членов-корреспондентов. Всего были исследованы показатели 279 персон; их краткие биографические данные содержатся на сайте РАО[1] и в [9].

В качестве международных источников информации были взяты базы данных Scopus и Web of Science (Science Citation Index и Social Science Citation Index). Выяснилось, что международные БД довольно скупо отражают публикации членов РАО: публикации большинства ученых в них вообще не отражены (в Scopus не было зарегистрировано трудов у 152 человек, а в WoS – у 230; более 10 публикаций имели соответственно 40 и 22 человека). Причины очевидны: журналы, в которых публиковались эти авторы, не представлены в соответствующих БД, а доля англоязычных работ незначительна. Примерно такая же картина наблюдалась с числом цитирований.

Нормативные документы для оценки результативности отечественных учёных и научных организаций недвусмысленно предписывают наряду с международными БД использовать Российский индекс научного цитирования (РИНЦ). Подробно это будет рассмотрено ниже.

Авторы исследования также посчитали полезным включить в изучение публикационной активности данные из российского сектора интернета – посвящённые учёным веб-страницы, их персональные сайты, упоминания в блогах. Интернет-публикации и другие формы коммуникации с помощью интернета если и не вытесняют традиционные научные коммуникации, то существенно их дополняют. Соответствующие данные частично описаны в [2]; в полном объёме они представлены в отчёте об исследовании.

В качестве отраслевого источника информации для оценки публикационной активности был использован электронный каталог НПБ им. К.Д. Ушинского. Именно в отраслевом каталоге представлены практически все члены РАО (265 человек из 279).

Предполагалось, что подобные исследования будут повторяться. Для этого имелись все предпосылки. Однако жизнь распорядилась иначе.

Научный коллектив ИНИМ формировался с 1969 г. на базе лаборатории НИИ содержания и методов обучения в г. Черноголовка. В 1989 г. был создан

[1] http://rusacademedu.ru

Центр комплексного формирования личности АПН СССР, который в 2003 г. переименован в Центр экспериментальной психодидактики РАО, а в 2008 г. - в Институт научной информации и мониторинга. В 2012 г. после очередного переименования он получил название Институт научной и педагогической информации (ФГБНУ ИНИПИ РАО). В институте были созданы и поддерживались, в частности, такие информационные ресурсы, как Открытый архив по педагогике, психологии и образованию[2], Объединенный фонд электронных ресурсов «Наука и образование» ОФЭРНиО[3]. А потом началась академическая реформа. Согласно Каталогу организаций России [13], «деятельность юридического лица прекращена путем реорганизации в форме присоединения с 19 мая 2015 года». Правопреемник – Институт управления образованием (ФГБНУ ИУО РАО).

Здесь уместно напомнить, что Российская академия образования ведёт свою историю с 6 октября 1943 г., когда Совнарком СССР постановлением № 1092 утвердил проект об организации Академии педагогических наук [14]. В 1967 г. численность АПН СССР была установлена в количестве 50 действительных членов и 80 членов-корреспондентов. Преемницей союзной академии в 1992 году стала РАО. 27 сентября 2013 г. был принят Федеральный закон № 253-ФЗ «О Российской академии наук, реорганизации государственных академий наук и внесении изменений в отдельные законодательные акты Российской Федерации». Согласно постановлению № 1290 от 26 декабря 2013 года, научно-исследовательские институты, подведомственные РАО, были отнесены к ведению Минобрнауки России [15]. В декабре 2014 года была проведена реорганизация институтов, входивших в систему РАО, — из 22 было создано 10 научных организаций. В октябре 2015 года к Академии в качестве структурного подразделения была присоединена Научная педагогическая библиотека имени К.Д. Ушинского. Но в октябре 2014 г. в библиотеке сменилось руководство, после чего резко сократилась её научная активность.

Таким образом, из приведённой информации следует, что к настоящему времени ни коллектив лаборатории, выполнявший исследования десятью годами ранее, ни сам институт ИНИМ более не существуют. НПБ им. К.Д. Ушинского

[2] http://ушинский.рф

[3] http://www.ofernio.ru

потеряла интерес к проекту. Поэтому, решив повторить исследование в 2019 году, автор с учётом своих физических возможностей ограничился показателями действительных членов РАО по состоянию на начало апреля 2019 г. и информацией РИНЦ.

О ПУБЛИКАЦИОННОЙ АКТИВНОСТИ ЧЛЕНОВ РАО В 2019 Г. (ПО ДАННЫМ РИНЦ)

Как известно, Российский индекс научного цитирования разрабатывается с 2005 года компанией «Научная электронная библиотека»[4]. Заявленная цель РИНЦ состоит в обеспечении научных исследований актуальной справочно-библиографической информацией и оценивании результативности и эффективности деятельности научно-исследовательских организаций, учёных, уровня научных журналов и т. д. К настоящему времени он стал информационно-аналитической системой национального уровня, содержащей более 12 миллионов публикаций российских ученых, а также информацию о цитировании этих публикаций из более 6000 журналов. РИНЦ позволяет оценивать результативность исследовательской работы и детально исследовать статистику публикационной активности более 600 тысяч российских ученых и 11 тысяч научных организаций, относящихся ко всем отраслям знаний. В России база данных РИНЦ является одним из основных источников информации для оценки эффективности организаций, занимающихся НИР. Так, постановление президиума РАН № 201 от 12.10.2010 [16] предписывает использовать для оценки научного потенциала и эффективности научных исследований такие показатели, как число публикаций и цитируемость работников научной организации в РИНЦ, отнесённое к численности исследователей.

Согласно данным открытых источников, в первую очередь сайтов РАО и РИНЦ, на апрель 2019 года Академия объединяет 128 действительных членов, на долю которых приходятся 17953 зарегистрированные публикации (в среднем по 140.26 на человека) и 397230 цитирований (3103.36). В приложении содержится информация об индексе Хирша, числе публикаций и числе цитирований для каждого академика. В публикации 2011 года приводились материалы по 123 академикам РАО. За прошедшее время академия пополнилась 49 новыми чле-

[4] http://elibrary.ru

нами, в то же время 44 человека выбыли в силу естественных причин. Таким образом, 79 персон присутствуют в обоих списках.

Сайт РАО содержит сведения о датах рождения членов Академии. Несложные вычисления показывают, что средний возраст академиков равен 75.5 годам, при этом достигли 70 лет 97 человек (75.8%), а 90 лет – 10 (7.8%).

Приведём результаты исследования публикационной активности с разбиением по возрастным группам. В таблице 1 в первом столбце указан возрастной диапазон, N обозначает численность соответствующей группы, P – среднее число публикаций, C – среднее число цитирований, H – усреднённый индекс Хирша.

Таблица 1. Публикационная активность в зависимости от возраста

	N	P	C	H
49-59	9	103	938	13
60-69	22	158	3268	18
70-79	52	151	2744	15
80-89	35	131	4472	16
90+	10	109	1771	10
Всего	128	140	3103	15

Как отмечалось выше, для 79 человек имеются данные РИНЦ из обоих исследований. Представляет определённый интерес сравнение их показателей прежде и теперь.

Таблица 2. Изменение средних показателей

	Колич	Публ (ср)	Цит (ср)	H
2019	128	140	3103	15
2010	123	11	48	
По пересечению списков				
2019	79	140	3472	
2010	79	10	50	

И ещё несколько фактов о первом исследовании (значения индекса Хирша в нём не фиксировались). Тогда 38 учёных не имели публикаций, зафиксированных в РИНЦ, а у 51 человека не было цитирований. При этом не менее 50 публикаций имели 6 академиков, а 13 – не менее 100 цитат. Заметим, что в 2019 г. только у 3 членов академии нули в соответствующих графах.

Последние показатели требуют комментариев. Обращает внимание, что при прочих равных условиях учёные за последние годы имеют в 13 раз больше публикаций, чем за всю предшествующую жизнь, а цитирований – почти в 70 раз. Предложим в качестве объяснения следующие соображения. Во-первых, как указано выше, база данных РИНЦ начала формироваться в 2005 г. и поначалу пополнялась довольно медленно. Но после того, как президиум ВАК назвал наличие научных периодических изданий в системе РИНЦ необходимым условием для включения их в Перечень ВАК [4], рост заметно ускорился. Информационная база РИНЦ существенно расширилась после включения в базу сведений по отечественным журналам, извлеченных из БД Scopus.

Во-вторых, свою роль сыграло упомянутое постановление № 201 о методике оценки результативности деятельности научных организаций. Назовём в этой связи и правительственное постановление №312 от 08.04.2009 [17], предписывающие разделить организации на три категории в зависимости от их достижений. Институты, стремящиеся улучшить своё положение (и получить увеличенное финансирование), стремятся повысить свои показатели. РИНЦ идёт им навстречу: заключив договор и уплатив соответствующую сумму, организация получает доступ к базам данных и может исправлять ошибки в описаниях работ и списках литературы, добавлять отсутствующие в базе цитирования и публикации, в том числе монографии и сборники трудов конференций, вносить другие изменения и дополнения. После внесения добавлений или исправлений публикация проверяется в ручном режиме сотрудниками РИНЦ и может быть возвращена на доработку или отклонена. В 2011 г. получили возможность корректировать свои записи в РИНЦ и учёные – авторы публикаций. Академики РАО, в большинстве являющиеся руководителями научных подразделений и организаций, имеют достаточно ресурсов для увеличения своего «научного веса».

Автор (не являющийся академиком РАО) может проиллюстрировать динамику роста показателей в РИНЦ на примере собственной статистики. В 2014 году

система фиксировала у него 34 публикации с 70 цитатами при индексе Хирша H=2. Сейчас эти цифры выглядят так: 185, 749, 13. При этом рост достигнут в основном за счёт включения публикаций прошлых лет, ранее отсутствовавших в РИНЦ.

РИНЦ ГЛАЗАМИ ИССЛЕДОВАТЕЛЕЙ

РИНЦ не свободен от внутренних недостатков и уязвим для внешних манипуляций. В статьях [5, 6] профессор Н.Е. Калёнов приводит многочисленные примеры некорректной работы алгоритмического и программного обеспечения РИНЦ. Демонстрируя скриншоты, он убедительно обосновывает свои претензии к полноте, актуальности, точности алгоритмов обработки данных в РИНЦ. В частности, на материалах собственных публикаций он выяснил, что при определённых запросах система отображает у него больше работ, чем у его организации в целом. В итоге делается вывод: «В том виде, в котором РИНЦ представлен в настоящее время на сайте НЭБ, использовать систему как инструмент, позволяющий осуществлять оценку результативности и эффективности деятельности научно-исследовательских организаций, ученых, уровень научных журналов и т. д., ни в коем случае нельзя» [5, с. 12].

Несколько лет спустя профессор А.Л. Фрадков высказывает аналогичное суждение: «РИНЦ продолжает искажать наукометрические данные учёных и не пытается их системно исправлять. Использовать эти данные для оценки ученых, журналов и организаций нельзя» [10, с. 5]. И причинами этого он называет как объективные трудности («проблема однофамильцев непроста, если решать её без помощи авторов»), так и сознательные действия руководства («РИНЦ готов свою систему превратить в помойку и вписывать туда всё, лишь бы платили»).

Те же оценки встречаем у профессора Р.М. Хантемирова: «РИНЦ приносит только вред. Этот вред связан, во-первых, с тем, что база журналов РИНЦ напоминает огромную помойку, в которой непросто отыскать что-либо стоящее. И, во-вторых, с принявшим угрожающие масштабы жульничеством отдельных авторов и журналов при накручивании своих библиометрических показателей, которому руководство РИНЦ потворствует своим принципиальным нежеланием противодействовать» [12, с. 6]. Так, по его словам, Сибирский педагогический журнал увеличил свой импакт-фактор с помощью примитивного мошенничества.

Схема проста: «проверенные» авторы в тексты объёмом в несколько страниц вставляют десятки ссылок на нужные журналы. Заметим, подобные предложения поступали и автору [18]. Проплаченные и не проверяемые на плагиат статьи невысокого качества часто используются для искусственного повышения цитируемости. «Мусорные» журналы делают на этом прибыльный бизнес. Это дискредитирует не только использование наукометрических показателей, но и саму научную деятельность в России.

Справедливости ради нужно констатировать: требования Минобрнауки по росту числа и цитируемости публикаций ставит учёных в условия, когда стремление соблюдать нормы этики приходит в противоречие с материальной заинтересованностью. «Часто само начальство в погоне за рейтингами заставляет нарушать этические нормы под угрозой понижения в должности, сокращения занятости, увольнения, расформирования кафедр и лабораторий и т. п. Поэтому нормы этики нарушаются фактически под давлением сверху» [11, с. 5]. Для исправления ситуации требуются усилия и научной общественности, и журнальных редакций, и политическая воля руководящих органов.

ЗАКЛЮЧЕНИЕ

Что касается РИНЦ, в последнее время очевидны положительные сдвиги. Реагируя на справедливую критику, специалисты РИНЦ начали следить за публикационной деятельностью журналов и сборников тезисов конференций. В апреле 2017 г. генеральный директор eLibrary.Ru Г.О. Еременко сообщил на конференции «Научное издание международного уровня мировая практика подготовки и продвижения публикаций» об исключении из РИНЦ 344 «мусорных» журналов [3]. На очереди – материалы «заочных» мультидисциплинарных конференций.

Не раз отмечалось, что количественные наукометрические показатели не должны использоваться для оценки эффективности научных работников [1, 7, 8]: они уязвимы для манипулирования, допускают неоднозначную интерпретацию. Пожалуй, наиболее искажённую картину формализация даёт в гуманитарных науках (заметим, в РАО преобладают именно гуманитарии). Они имеют, как правило, низкие показатели в библиометрических базах WoS и Scopus. В гуманитарных науках принято представлять результаты исследований в виде монографий

и статей в тематических сборниках, которые выпадают из поля зрения этих баз. Кроме того, национальная специфика объекта исследования далеко не всегда интересна зарубежной аудитории, а многие ведущие журналы не имеют английской версии.

Тем не менее, значение РИНЦ не следует недооценивать. Он фактически стал национальной информационно-аналитической системой с данными о публикациях и цитировании этих публикаций. Созданный аналитический аппарат даёт подробное и наглядное представление информации. Упомянутым гуманитариям он предоставляет более полную и объективную картину, чем WoS и Scopus. И относиться к нему следует не как к главному критерию качества научных работ, а как к инструменту анализа для исследователей и экспертов.

В статье использованы материалы доклада на XXI Всероссийской научной конференции «Научный сервис в сети интернет» (сентябрь 2019).

Приложение. Показатели публикационной активности членов РАО

(Р – число публикаций, С – число цитирований, Н –индекс Хирша)

Name	H	2019		2011	
		P	C	P	C
АБУЛЬХАНОВА Ксения Александровна	38	154	19877	6	45
АЛАШКЕВИЧ Юрий Давыдович	7	197	387		
АМОНАШВИЛИ Шалва Александрович	9	111	5596	4	0
АНТОНОВА Ирина Александровна	0	1	0	0	0
АНТОНОВА Лидия Николаевна	9	73	369		
АСМОЛОВ Александр Григорьевич	32	336	19442	51	830
БАЕВА Ирина Александровна	19	172	2470		
БАШМАКОВ Марк Иванович	8	199	754	1	6
БЕЗРУКИХ Марьям Моисеевна	21	215	4987	30	274
БЕЛОУСОВ Лев Сергеевич	8	143	183		
БЕРУЛАВА Галина Алексеевна	21	67	2283		
БЕРУЛАВА Михаил Николаевич	17	98	2921	1	0
БЕСПАЛЬКО Владимир Павлович	18	97	12069	18	7
БИМ-БАД Борис Михайлович	17	152	3194	6	4
БОЛОТОВ Виктор Александрович	19	152	4979		
БОНДЫРЕВА Светлана Константиновна	20	95	2156	4	101

БОРДОВСКАЯ Нина Валентиновна	19	149	5500	0	0
БОРДОВСКИЙ Геннадий Алексеевич	20	555	3458	63	221
БОРИСЕНКОВ Владимир Пантелей-монович	11	77	934	18	38
БУЕВА Людмила Пантелеевна	9	69	2692	0	0
ВЕРБИЦКАЯ Людмила Алексеевна	11	130	1597	5	21
ВЕРБИЦКИЙ Андрей Александрович	32	354	16869		
ГАЙДАМАШКО Игорь Вячеславович	8	60	347		
ГАЛАЖИНСКИЙ Эдуард Владимиро-вич	16	103	1885		
ГАРАДЖА Виктор Иванович	7	28	1027	1	2
ГАФУРОВ Ильшат Рафкатович	13	87	688		
ГЕВОРКЯН Елена Николаевна	15	81	948		
ГЛЕЙЗЕР Григорий Давыдович	6	31	352	0	0
ГРАНИК Генриэтта Григорьевна	7	74	733	0	0
ДАРМОДЕХИН Сергей Владимиро-вич	9	62	476	9	23
ДЕДЕГКАЕВ Виктор Хасанбиевич	3	27	31		
ДЁМИН Вадим Петрович	1	13	6		
ДЕРКАЧ Анатолий Алексеевич	25	162	9961	20	349
ДЖУРИНСКИЙ Александр Наумович	27	218	4391		
ДОНЦОВ Александр Иванович	20	130	3324	1	0
ДРОНОВ Виктор Павлович	9	87	390		

ДУБРОВИНА Ирина Владимировна	17	188	2480	0	0
ЕРМАКОВ Павел Николаевич	14	178	1204		
ЖУРАВЛЁВ Анатолий Лактионович	68	792	16298		
ЖУРАКОВСКИЙ Василий Максимилианович	15	100	1271	5	5
ЗАГВЯЗИНСКИЙ Владимир Ильич	45	263	11625	33	142
ЗАПЕСОЦКИЙ Александр Сергеевич	28	370	3925	29	62
ЗАХЛЕБНЫЙ Анатолий Никифорович	13	96	1501		
ЗИМНЯЯ Ирина Алексеевна	24	101	18378	42	482
ЗИНЧЕНКО Юрий Петрович	23	208	2031		
ИВАННИКОВ Вячеслав Андреевич	13	61	1308		
КАНДЫБОВИЧ Сергей Львович	8	68	566		
КАРАМУРЗОВ Барасби Сулейманович	10	171	734		
КЕЗИНА Любовь Петровна	1	3	45	1	0
КИНЕЛЁВ Владимир Георгиевич	13	77	1556	2	6
КИСЕЛЕВ Александр Федотович	12	119	580	31	9
КОРОЛЬКОВ Александр Аркадьевич	10	177	1118	4	8
КОСТОМАРОВ Виталий Григорьевич	19	222	13176	4	1
КУЗНЕЦОВ Александр Андреевич	23	221	2734	0	0
КУКУШКИНА Ольга Ильинична	12	99	1200		
КУРАКОВ Лев Пантелеймонович	8	92	1684	5	15
КУЦЕВ Геннадий Филиппович	17	122	1079		

ЛАЗАРЕВ Валерий Семёнович	29	143	4339	4	5
ЛАПТЕВ Владимир Валентинович	19	239	1195	10	58
ЛАПЧИК Михаил Павлович	22	100	2297	1	18
ЛЕБЕДЕВ Юрий Александрович	5	38	89	0	0
ЛЕВИЦКИЙ Михаил Львович	9	82	309	0	0
ЛЕКТОРСКИЙ Владислав Александрович	42	282	9966	4	14
ЛИФЕРОВ Анатолий Петрович	15	123	1127	5	5
ЛИХАНОВ Альберт Анатольевич	2	22	68	1	0
ЛОМОВ Станислав Петрович	2	66	582	0	0
МАКСИМОВИЧ Валентина Фёдоровна	6	23	170	0	0
МАЛОФЕЕВ Николай Николаевич	21	128	2836	29	83
МАЛЫХ Сергей Борисович	22	241	1937		
МАЛЫШЕВ Владимир Сергеевич	2	23	28		
МАНУШИН Эдуард Анатольевич	16	74	852	0	0
МАРТИРОСЯН Борис Пастерович	7	22	530	0	0
МЕДВЕДЕВ Леонид Георгиевич	9	23	248		
МИНДИАШВИЛИ Дмитрий Георгиевич	11	21	571		
МИХАЙЛОВА Евгения Исаевна	6	41	173		
МИХАЙЛОВА Наталья Ивановна	3	20	90	0	0
МУХИНА Валерия Сергеевна	21	281	7814	26	66

МЯСНИКОВ Владимир Афанасьевич	8	81	463	7	6
НЕВЕРКОВИЧ Сергей Дмитриевич	18	163	1940		
НЕМЕНСКИЙ Борис Михайлович	5	26	676	0	0
НЕЧАЕВ Николай Николаевич	10	81	1715	3	13
НИКАНДРОВ Николай Дмитриевич	18	220	4937	15	9
НИКИТИН Александр Александрович	5	91	469	5	1
ОМАРОВ Омар Алиевич	7	155	532	26	23
ОРЛОВ Александр Андреевич	19	128	2179		
ПАТОВ Николай Александрович	6	22	98		
ПОДДЬЯКОВ Николай Николаевич	9	32	748	0	0
ПОДУФАЛОВ Николай Дмитриевич	4	43	181	7	9
ПОНОМАРЕНКО Владимир Александрович	18	332	5345	9	18
ПОПКОВ Владимир Андреевич	23	306	3017	14	11
ПОТАШНИК Марк Матусович	15	224	4021	0	0
РЕАН Артур Александрович	22	165	11794		
РОБЕРТ Ирэна Веньяминовна	24	208	7353	10	15
РУБЦОВ Виталий Владимирович	25	223	3524	8	10
РЫЖАКОВ Михаил Викторович	14	110	1095	38	77
СЕЙРАНОВ Сергей Германович	20	157	1419		
СЕМЁНОВ Алексей Львович	9	119	1037		
СЕНЬКО Юрий Васильевич	26	150	4346	25	59

СЕРГЕЕВ Николай Константинович	10	93	1174		
СИНЕНКО Василий Яковлевич	9	69	502		
СЛОНИМСКИЙ Сергей Михайлович	10	79	686	0	0
СМОЛИН Олег Николаевич	12	196	1223		
СМОЛЯНИНОВА Ольга Георгиевна	13	132	1239		
СОБКИН Владимир Самуилович	26	413	4394	13	95
СОВЕТОВ Борис Яковлевич	15	110	2754	4	16
СОЛОМИН Юрий Мефодьевич	0	0	0		
СТРИХАНОВ Михаил Николаевич	64	534	18346		
ТАЮРСКИЙ Анатолий Иванович	5	71	128	8	0
ТИКТИНСКИЙ-ШКЛОВСКИЙ Виктор Маркович	10	139	920	0	0
ТРЯПИЦЫНА Алла Прокофьевна	36	434	5058		
ТХАКУШИНОВ Асланчерий Китович	7	37	206		
УСАНОВ Владимир Евгеньевич	9	71	356		
УШАКОВА Татьяна Николаевна	22	129	3566	15	49
ФАРБЕР Дебора Ароновна	21	131	4513	46	476
ФИЛИППОВ Владимир Михайлович	16	166	1979	51	97
ФОХТ-БАБУШКИН Юрий Ульрихович	2	3	193	0	0
ХАЛЕЕВА Ирина Ивановна	7	26	2654	7	0
ЦВЕТКОВА Лариса Александровна	9	85	378		
ЦИРУЛЬНИКОВ Анатолий Маркович	6	45	539		

ЧЕБЫШЕВ Николай Васильевич	6	83	636	4	22
ЧИСТЯКОВА Светлана Николаевна	17	160	2986		
ШАДРИКОВ Владимир Дмитриевич	34	257	14528	21	59
ШКОЛЯР Людмила Валентиновна	6	95	665	1	0
ЩЕТИНИН Михаил Петрович	0	0	0	0	0
ЩУКИН Евгений Дмитриевич	21	530	5172	0	0
ЭРДНИЕВ Пюрвя Мучкаевич	11	61	1227	0	0
ЭСКИНДАРОВ Мухадин Абдурахманович	44	215	6350		
ЯМБУРГ Евгений Шоломович	9	74	979		

СПИСОК ЛИТЕРАТУРЫ

1. *Антопольский А.Б.* Принципы системы научной информации в Российской академии образования // Информационные ресурсы России. 2009. № 4 (110). С. 16–22.

2. *Антопольский А.Б., Поляк Ю.Е.* Об исследовании публикационной активности ученых (на примере членов Российской академии образования) // Информационные ресурсы России. 2011. № 1 (119). С. 26–30.

3. *Еременко Г.О.* Актуальные проблемы современной научной периодики: мусорные журналы и ретракция статей. URL: https://conf.neicon.ru/materials/26-Domestic0417/170419-06-Eremenko.pdf

4. *Информационное сообщение №45.1-132 от 14.10.2008 о порядке формирования Перечня ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени доктора и кандидата наук.* URL: https://elibrary.ru/projects/events/vak/infletter-14-10-2008.doc

5. *Каленов Н.Е.* Еще раз о РИНЦ (письмо Министру образования и науки РФ Фурсенко А.А.) // Троицкий вариант-Наука. 2011. № 71. С. 4.

6. *Каленов Н.Е., Селюцкая О.В.* Некоторые оценки качества Российского индекса научного цитирования на примере журнала «Информационные ресурсы России» // Информационные ресурсы России. 2010. № 6. С. 2–13.

7. *Поляк Ю.Е.* Наукометрические показатели в оценке деятельности ученых и организаций // Дистанционное и виртуальное обучение. 2014. №8. С.101–106

8. *Поляк Ю.Е.* Оценивание и ранжирование веб-сайтов. Вебометрические рейтинги // Научный редактор и издатель. 2017. Т. 2. № 1. С. 19–29.

9. Российская академия образования. Персональный состав, 1943—2013. М.: НПБ им. К.Д. Ушинского, 2013. 436 с.

10. *Фрадков А.Л.* РИНЦ продолжает врать // Троицкий вариант-Наука.

11. *Фрадков А.Л.* РИНЦ учит врать? // Троицкий вариант-Наука. 2015. № 189. С. 5.

12. *Хантемиров Р.М.* РИНЦ: от примитивного мошенничества до растления малолетних // Троицкий вариант-Наука. 2014. № 163. С. 6.

13. *Организация ФГБНУ «ИНИПИ РАО».* URL: http://www.list-org.com/company/824439

14. *Об организации Академии педагогических наук РСФСР.* URL: http://rusacademedu.ru/wp-content/uploads/2018/10/ postanovlenie_1943_ 1092_.pdf

15. *О федеральных органах исполнительной власти, уполномоченных осуществлять функции и полномочия учредителя и собственника имущества организаций, находившихся в ведении Российской академии образования.* URL: http://government.ru/docs/9585

16. *Об утверждении Положения о Комиссии по оценке результативности деятельности научных организаций Российской академии наук и Методики оценки результативности деятельности научных организаций Российской академии наук.* URL: http://www.ras.ru/presidium/documents/directions.aspx?ID=9767952e-4821-4510-89d6-5f678677066d

17. *Об оценке и о мониторинге результативности деятельности научных организаций, выполняющих научно-исследовательские, опытно-конструкторские и технологические работы гражданского назначения* URL: http://www.pravo.gov.ru/proxy/ips/?docbody= &nd=102128788

18. *Поляк Ю.Е.* О методах повышения импакт-фактора // Телематика-2014. Труды XXI Всероссийской научно-методической конференции. СПб, 2014. С. 49–51.

RSCI AS A MIRROR OF PUBLICATION ACTIVITY OF RAE MEMBERS

Yu.E. Polyak

Central Economics and Mathematics Institute of the Russian Academy of Sciences, Moscow

polak@cemi.rssi.ru

Abstract

Based on information from open sources, a table was compiled reflecting the indicators of 128 full members of the Russian Academy of Education (RAE) in the Russian Science Citation Index (RSCI). The main results are given in a condensed form and compared with the results of a similar study performed several years earlier. The conclusions and features of the RSCI as an analytical tool are discussed.

Keywords*: Russian Academy of Education, Russian Science Citation Index, publication activity*

REFERENCES

1. *Antopol'skij A.B.* Principy sistemy nauchnoj informacii v Rossijskoj akademii obrazovaniya // Informacionnye resursy Rossii. 2009. № 4 (110). S. 16–22.

2. *Antopol'skij A.B., Polyak YU.E.* Ob issledovanii publikacionnoj aktivnosti uchenyh (na primere chlenov Rossijskoj akademii obrazovaniya) // Informacionnye resursy Rossii. 2011. № 1 (119). S. 26–30.

3. *Eremenko G.O.* Aktual'nye problemy sovremennoj nauchnoj periodiki: musornye zhurnaly i retrakciya statej. URL: https://conf.neicon.ru/materials/26-Domestic0417/170419-06-Eremenko.pdf

4*. Informacionnoe soobshchenie №45.1-132 ot 14.10.2008 o poryadke formirovaniya Perechnya vedushchih recenziruemyh nauchnyh zhurnalov i izdanij, v kotoryh dolzhny byt' opublikovany osnovnye nauchnye rezul'taty dissertacij na soiskanie*

uchenoj stepeni doktora i kandidata nauk. URL: https://elibrary.ru/ projects/ events/ vak/infletter-14-10-2008.doc

5. *Kalenov N.E.* Eshche raz o RINC (pis'mo Ministru obrazovaniya i nauki RF Fursenko A.A.) // Troickij variant-Nauka. 2011. № 71. S. 4.

6. *Kalenov N.E., Selyuckaya O.V.* Nekotorye ocenki kachestva Rossijskogo indeksa nauchnogo citirovaniya na primere zhurnala «Informacionnye resursy Rossii» // Informacionnye resursy Rossii. 2010. № 6. S. 2–13.

7. *Polyak YU.E.* Naukometricheskie pokazateli v ocenke deyatel'nosti uchenyh i organizacij // Distancionnoe i virtual'noe obuchenie. – 2014. No 8. S. 101–106.

8*. Polyak YU.E.* Ocenivanie i ranzhirovanie veb-sajtov. Vebometricheskie rejtingi // Nauchnyj redaktor i izdatel'. 2017. T. 2. No 1. S. 19–29.

9. Rossijskaya akademiya obrazovaniya. Personal'nyj sostav, 1943–2013. M.: NPB im. K.D. Ushinskogo, 2013. 436 s.

10. *Fradkov A.L.* RINC prodolzhaet vrat' // Troickij variant-Nauka. 2015. No 187. S. 5.

11. *Fradkov A.L.* RINC uchit vrat'? // Troickij variant-Nauka. 2015. No 189. S. 5.

12. *Hantemirov R.M.* RINC: ot primitivnogo moshennichestva do rastleniya maloletnih // Troickij variant-Nauka. 2014. No 163. S. 6.

13. *Organizaciya FGBNU «INIPI RAO».* URL: http://www.list-org.com/ company/824439

14. *Ob organizacii Akademii pedagogicheskih nauk RSFSR.* URL: http://rusacademedu.ru/wp-content/uploads/2018/10/postanovlenie_1943_ 1092_.pdf

15. *O federal'nyh organah ispolnitel'noj vlasti, upolnomochennyh osushchestvlyat' funkcii i polnomochiya uchreditelya i sobstvennika imushchestva organizacij, nahodivshihsya v vedenii Rossijskoj akademii obrazovaniya.* URL: http://government. ru/docs/9585

16. *Ob utverzhdenii Polozheniya o Komissii po ocenke rezul'tativnosti deyatel'nosti nauchnyh organizacij Rossijskoj akademii nauk i Metodiki ocenki rezul'tativnosti deyatel'nosti nauchnyh organizacij Rossijskoj akademii nauk.* URL: http://www.ras.ru/presidium/documents/directions.aspx?ID=9767952e-4821-4510-89d6-5f678677066d

17. *Ob ocenke i o monitoringe rezul'tativnosti deyatel'nosti nauchnyh organizacij, vypolnyayushchih nauchno-issledovatel'skie, opytno-konstruktorskie i tekhnologicheskie raboty grazhdanskogo naznacheniya.* URL: http://www.pravo.gov.ru/proxy/ips/?docbody= &nd=102128788

18. *Polyak Yu.E.* O metodah povysheniya impakt-faktora // Telematika-2014. Trudy XXI Vserossijskoj nauchno-metodicheskoj konferencii. SPb, 2014. S. 49–51.

СВЕДЕНИЯ ОБ АВТОРЕ

ПОЛЯК Юрий Евгеньевич – ведущий научный сотрудник Центрального экономико-математического института РАН (Москва). Подробнее: http://computer-museum.ru/articles/sovet-muzeya/561/

Yuri Evgenievich POLYAK – Candidate of Economic Sciences, Leading Researcher, Central Economics and Mathematics Institute. Moscow, Russia. More detailed: http://computer-museum.ru/articles/sovet-muzeya/561/

email: polak@cemi.rssi.ru

Материал поступил в редакцию 12 ноября 2019 года