ОГЛАВЛЕНИЕ

УДК 004.4’422

АВТОМАТИЗАЦИЯ РАСПАРАЛЛЕЛИВАНИЯ ПРОГРАММ ДЛЯ МНОГОЯДЕРНЫХ ПРОЦЕССОРОВ С РАСПРЕДЕЛЕННОЙ ЛОКАЛЬНОЙ ПАМЯТЬЮ

А. П. Баглий[1] [0000-0001-9089-4164], Н. М. Кривошеев[2] [0000-0002-1366-7037], Б. Я. Штейнберг[3] [0000-0001-8146-0479]

[1]*Южный федеральный университет, Институт математики, механики и компьютерных наук;*

[1]abagly@sfedu.ru, [2]krivosheev@sfedu.ru, [3]byshtyaynberg@sfedu.ru

Аннотация

В статье идет речь о создании распараллеливающих компиляторов для вычислительных систем с распределенной памятью. Промышленные распараллеливающие компиляторы распараллеливают программы для вычислительных систем с общей памятью. Преобразование последовательных программ для вычислительных систем с распределенной памятью требует разработки дополнительных функций. Это становится актуальным для перспективных систем на кристалле с сотнями и более ядер. В терминах графа информационных связей сформулировано условие распараллеливания программного цикла на вычислительную систему с распределенной памятью.

***Ключевые слова**: автоматизация распараллеливания, распределенная память, преобразования программ, размещение данных, пересылки данных*

ВВЕДЕНИЕ

Во многих публикациях рассматривается задача автоматического создания параллельного кода для систем с распределенной память, но предлагаются полуавтоматические инструменты, в которых пользователь должен вызывать специальные функции или писать директивы компилятору. В работе [4] отмечено, что автоматическая компиляция последовательной программы для параллельной архитектуры с распределенной памятью является очень сложной задачей, не имеющей в настоящее время эффективного решения. В этой же работе описана генерация параллельного кода с генерацией коммуникаций, основанных на MPI (для

кластера с мультипроцессорами). Об оптимизации размещения данных не говорится ничего – значит, пользователь компилятора должен это делать сам. Использована полиэдральная распараллеливающая система Pluto, которая позволяет в пространстве итераций находить подмножества точек, допускающих параллельное выполнение. Ничего не говорится о локализации данных, хотя на современных вычислительных архитектурах распараллеливание эффективно при параллельном выполнении относительно больших фрагментов кода [16], особенно при распределенной памяти.

Для вычислительной системы (ВС) с распределенной памятью самой длительной операцией является межпроцессорная пересылка данных. Как показано в [8], накладные расходы, связанные с пересылками данных, приходится учитывать при отображении программ на распределенную память. Но в последнее время появляются многоядерные процессоры, иногда называемые «суперкомпьютер на кристалле», с десятками, сотнями и тысячами ядер [9, 10, 13]. Пересылка данных между процессорными ядрами на одной микросхеме требует значительно меньше времени, чем на коммуникационной сети (Ethernet, Infiniband, PCI-express, …). Это означает расширение множества эффективно распараллеливаемых программ и делает целесообразным разработку распараллеливающих компиляторов. К этому списку можно добавить ориентированные на быструю работу с нейронными сетями процессоры, которые появляются в последние несколько лет [18–24].

В [14, 15] описаны блочно-аффинные размещения данных в распределенной памяти. Следует отметить работы группы DVM-System [5, 11] по генерации параллельного кода на ВС с распределенной памятью, где спецификации параллелизма (DVMH-указания) оформляются в виде специальных комментариев. Генерирующие автоматически параллельный код на ВС с распределенной памятью компилирующие системы DVM, Parawise и др. предполагают дописывание текста последовательной программы без предварительного преобразования. В работе [8] рассмотрена задача трансляции высокоуровневых описаний параллельной обработки данных в программе на уровень конструкций стандарта MPI для выполнения на системе с распределенной памятью. В [7] рассмотрена задача распарал-

леливания программного цикла на ВС с распределенной памятью с минимизацией межпроцессорных пересылок. Время, необходимое на пересылки данных, нелинейно зависит от объема пересылаемых данных. Значительное время отнимает инициализация пересылки. Метод размещения данных с перекрытиями [1, 3] существенно ускоряет параллельные итерационные алгоритмы за счет уменьшения количества пересылок при укрупнении множеств пересылаемых данных.

Для многих алгоритмов удобными являются циклические пересылки данных – это такие пересылки, которые для некоторой целой константы C из каждого процессорного элемента с номером k пересылают данные в процессорный элемент с номером $(p + C)\ mod\ p$, где p – количество процессорных элементов. Многие коммуникационные сети (например, кольцевая шина, mesh или тор) позволяют выполнять циклические пересылки для всех k за один такт. В [12] приведено много задач линейной алгебры и математической физики, для параллельного решения которых на ВС с распределенной памятью используются циклические пересылки.

В [2] приведены результаты экспериментов, показывающие, что современные оптимизирующие компиляторы плохо оптимизируют код и имеют большой неиспользованный потенциал оптимизирующих преобразований. Можно полагать, что для микросхем с сотнями вычислительных ядер этот потенциал оптимизаций больше, чем для процессоров, на которых проводились эксперименты.

Данная работа направлена на создание распараллеливающего компилятора, который автоматически анализирует высокоуровневый текст программы, находит размещение данных в распределенной памяти с минимизацией межпроцессорных пересылок, выполняемых коммуникационной сетью, и в итоге преобразует программу к виду, допускающему распараллеливание на ВС с распределенной памятью. Ниже приведен пример с генерацией параллельного MPI-кода, хотя можно использовать SHMEM или другие инструменты.

Настоящая работа отличается от других попыток автоматизировать отображение последовательных программ на вычислительные архитектуры с распределенной памятью тем, что в ней автоматически выполняется размещение массивов в распределенной памяти по тексту входной программы, написанному на высокоуровневом языке, а не директивам, дописанным к программе вручную. Это

удается сделать благодаря тому, что оптимальные размещения массивов ищутся среди разработанных ранее блочно-аффинных размещений массивов [14, 15], которые, с одной стороны, могут быть описаны малым количеством параметров (пропорционально размерности массива), а, с другой стороны, покрывают размещения, широко используемые на практике. Расширение множества распараллеливаемых программ может быть достигнуто с помощью оптимизирующих преобразований программ, которые имеются в используемой нами оптимизирующей распараллеливающей системе (ОРС) и известных оптимизирующих компиляторах LLVM, GCC, ICC, MS-compiler.

1. РАСПАРАЛЛЕЛИВАЕМЫЕ ПРОГРАММНЫЕ ЦИКЛЫ

Будем рассматривать задачу параллельного выполнения цикла. Как обычно, под распараллеливанием цикла понимаем одновременное выполнение его итераций – но это определение предполагает уточнения.

```
for (int j = 1; j < N; ++j) {
  Statement1(j);
  Statement2(j);
  Statement3(j);
}
```

Листинг 1: Простой цикл

Будем полагать, что цикл удовлетворяет следующим условиям.

1. В теле цикла счетчик цикла j не изменяет значения; имеется только один выход из цикла после завершения всех итераций, и переход на следующую итерацию возможен только после завершения предыдущей (т. е. в теле цикла нет операторов `break`, `continue` и `goto` с переходом за пределы цикла);
2. В цикле есть только вхождения одномерных массивов, индексное выражение которых имеет вид `(j + k)`, где `j` – счетчик цикла, k – некоторая константа или переменная, не изменяющая своего значения (в теле цикла);
3. В цикле есть только операторы присваивания.

Замечание. В последующих работах представленные ограничения будут ослаблены.

2. БЛОЧНО-АФФИНЫЕ РАЗМЕЩЕНИЯ МАССИВОВ

Основная особенность параллельного выполнения цикла на ВС с распределенной памятью состоит в том, что для каждой операции ее аргументы должны быть в одном модуле распределенной памяти.

Будем полагать, что ВС состоит из p процессорных элементов (ПЭ). Каждый ПЭ состоит из процессора и собственного модуля памяти, получение данных из которого происходит быстрее, чем из модулей памяти других ПЭ. Все ПЭ занумерованы, начиная с нуля. Размещение массива в памяти – это функция, которая для каждого элемента массива возвращает номер ПЭ, в котором этот элемент находится. При описании параллельных алгоритмов рассматриваются размещения матриц (двумерных массивов) «по строкам», «по столбцам», «по полосам строк», «по полосам столбцов», «по скошенным диагоналям». Эти описания, как и многие другие, могут быть описаны как блочно-аффинные размещения по модулю количества ПЭ.

Определение 1. Пусть натуральные (включая нуль) числа $p, d_1, d_2, \dots, d_m$ и целые константы $s_0, s_1, s_2, \dots, s_m$ зависят только от m-мерного массива X. Блочно-аффинное по модулю p размещение m-мерного массива X – это такое размещение, при котором элемент $X[i_1, i_2, \dots, i_m]$ находится в модуле памяти ПЭ с номером $u = \left(\left\lfloor\frac{i_1}{d_1}\right\rfloor * s_1 + \left\lfloor\frac{i_2}{d_2}\right\rfloor * s_2 + \dots + \left\lfloor\frac{i_m}{d_m}\right\rfloor * s_m + s_0\right) \ mod\ p.$

Число s_0 показывает номер модуля памяти, в котором размещается нулевой элемент $X[i_1, i_2, \dots, i_m]$. При описанном блочно-аффинном способе размещения m-мерный массив представляется как массив блоков размерности $d_1 * d_2 * \dots * d_m$, который размещается так, что у каждого блока все элементы оказываются в модуле памяти одного ПЭ. Числа $p, d_1, d_2, \dots, d_m, s_0, s_1, s_2, \dots, s_m$ будем называть параметрами размещения.

3. МЕЖПРОЦЕССОРНЫЕ ПЕРЕСЫЛКИ ДАННЫХ

Пересылка – это команда коммуникационной системы. Шина и кольцо позволяют выполнять межпроцессорные циклические пересылки. В данной статье будем рассматривать только циклические пересылки данных. К ним сводятся параллельные алгоритмы для многих задач линейной алгебры и численных мето-

дов решения задач математической физики [12]. Актуальность именно таких пересылок для ВС близкого будущего подчеркнута в [6].

Опишем условия, при которых можно в текст программного цикла вставить пересылку.

Рассмотрим последовательно выполняемый оператор цикла, содержащий два блока, между которыми планируется вставка пересылки

```
for (int j = 0; j < N; ++j) {
  B1(j);
  B2(j);
}
```

После вставки пересылки этот оператор цикла будет иметь вид

```
for (int j = 0; j < N; ++j) {
  B1(j);
  XX[j] ← X[j + k];
  B2(j);
}
```

Здесь пересылка в последовательной программе означает присваивание элементам нового массива элементов старого массива. Поскольку при переходе к параллельному выполнению цикла пересылка должна выполняться одновременно для всех значений счетчика цикла, этот код должен быть эквивалентен следующему

```
for (int j = 0; j < N; ++j) {
  B1(j);
}
for (int j = 0; j < N; ++j) {
  XX[j] ← X[j + k];
}
for (int j = 0; j < N; ++j) {
  B2(j);
}
```

В частности, отсюда вытекает, что к последовательному циклу можно применить преобразование «разрезание цикла». Это означает, что в исходном последовательном цикле не должно быть дуг графа информационных связей «снизу-вверх», точнее, из блока B1 в блок B2 [17]. Это требование можно ослабить, если вставку пересылки сопровождать переименованием некоторых вхождений пере-

менной X, заменив эту переменную новой переменной XX. Такое переименование позволяет иметь в исходном цикле дуги «снизу-вверх», которые ведут из вхождений переменной X нижнего блока B2 в верхний блок B1 и которые предполагается заменить новой переменной XX.

Заметим, в частности, что после вставки пересылки переименования вхождений переменной X возможны только в блоке B2, но не в B1.

Пример 1. В рассматриваемом цикле есть только одна дуга графа информационных связей, ведущая «снизу-вверх» от вхождения `X[j + 2]` к вхождению-генератору `X[j]`.

```
for  (int j = 0; j < N; ++j) {
  X[j] = A[j] + B[j] * X[j];
  Y[j] = X[j] + X[j + 2] * C[j + 1];
}
```

После вставки пересылки дуга графа информационных связей, ведущая «снизу-вверх», исчезает из-за переименования.

```
for (int j = 0; j < N; ++j) {
  X[j] = A[j + 3] + B[j] * X[j];
  XX[j] ← X[j + 2];
  Y[j] = X[j] + XX[j] * A[j + 1];
}
```

Условие распараллеливания программного цикла на вычислительную систему с распределенной памятью со вставкой межпроцессорных пересылок состоит в отсутствии дуг графа информационных связей «снизу-вверх», кроме дуг, ведущих из вхождений переменных, к которым применяется переименование (такие дуги исчезают после переименования). Такое переименование невозможно для дуг истинной или выходной зависимостей, а возможно только для антизависимостей.

Алгоритм распараллеливания программного цикла на вычислительную систему с распределенной памятью со вставкой межпроцессорных пересылок должен начинаться с приведения операторов тела цикла к виду, в котором все дуги графа информационных связей после соответствующих переименований будут направлены «сверху-вниз». Это возможно достичь в несколько этапов.

1) Избавляемся от выходных информационных зависимостей.

2) Располагаем операторы тела цикла так, чтобы все дуги истинной зависимости были направлены сверху вниз. Если это невозможно, то программный цикл имеет рекуррентно вычисляемые переменные и не может быть распараллелен.

3) Ищем наименьшее множество пересылок, которые необходимо вставить при распараллеливании.

4) Вставляем найденные пересылки в текст программного цикла.

5) Выполняем необходимые переименования вхождений переменных, которые копированы при пересылках.

Замечание. Переименования могут быть не связаны с удалением дуг графа информационных связей «снизу-вверх», например, переименования могут быть связаны с разрывом циклов ГОП (графа операторы-переменные).

Пример 2. В рассматриваемом цикле есть только одна дуга истинной зависимости графа информационных связей, ведущая «снизу-вверх» от вхождения `Y[j]` к вхождению-генератору `Y[j - 1]`.

```
for (int j = 1; j < N; ++j) {
  X[j] = A[j] * Y[j - 1] + Y[j + 1] * X[j];
  Y[j] = X[j] + X[j + 2] * C[j + 1];
  A[j] = Y[j + 1] * X[j];
}
```

В первую очередь переставим второй оператор присваивания на первое место, чтобы истинную зависимость по переменной Y, которая вела «снизу-вверх», направить в обратную сторону.

```
for (int j = 1; j < N; ++j) {
  Y[j] = X[j] + X[j + 2] * C[j + 1];
  X[j] = A[j] * Y[j - 1] + Y[j + 1] * X[j];
  A[j] = Y[j + 1] * X[j];
}
```

Теперь можно вставлять пересылки данных. Приведем один из вариантов таких вставок.

```
for (int j = 1; j < N; ++j) {
  XX[j] ← X[j + 2];
  Y[j] = X[j + 1] + XX[j] * C[j + 1];
  YY[j] ← Y[j + 1];
  AA[j] ← A[j + 3];
  X[j] = A[j] * Y[j - 1] + YY[j] * AA[j];
  A[j] = YY[j] * X[j];}
```

Определение 2. Размещение переменных будем называть согласованным для данного оператора, если для каждого значения счетчика цикла все вхождения переменных, входящих в данный оператор, расположены в одном и том же ПЭ.

Ясно, что для параллельного выполнения цикла на ВС с распределенной памятью оператор цикла и, в частности, все операторы цикла должны быть согласованы. Согласованность операторов может быть достигнута с помощью межпроцессорных пересылок и блочно-аффинных размещений данных.

4. ВЫБОР ОПТИМАЛЬНОГО РАЗМЕРА БЛОКА В БЛОЧНО-АФФИННОМ РАЗМЕЩЕНИИ ДАННЫХ

В данной работе рассматривается такой программный цикл, в котором шаг равен 1, тело содержит только операторы присваивания, правая часть каждого оператора присваивания является выражением, содержащим только вхождения массивов, левая часть каждого оператора присваивания является вхождением массива, а все вхождения массивов имеют вид a[i + c], где a – имя некоторого массива, i – счетчик цикла, c – некоторая целочисленная константа, известная на этапе компиляции.

```
{
  int i;
  for (…; …; ++i) {
    a_{j_1}[i + c_1] = f_1(a_{j_{1,1}}[i + c_{1,1}], …, a_{j_{1,n_1}}[i + c_{1,n_1}]);
    …
    a_{j_k}[i + c_k] = f_k(a_{j_{k,1}}[i + c_{k,1}], …, a_{j_{k,n_k}}[i + c_{k,n_k}]);
  }
}
```

Листинг 2: Вид программного цикла

Будем считать, что цикл выполняет N итераций, тогда можно выполнить гнездование цикла с размером блока d, то есть преобразовать его к виду

```
{
  int i, j;
  for (j = 0; j  < N; j += d) {
    for (i = j; i < min(N, j + d); ++i) {
      a_{j_1}[i + c_1] = f_1(a_{j_{1,1}}[i + c_{1,1}], …, a_{j_{1,n_1}}[i + c_{1,n_1}]);
```

…

$a_{j_k}[i + c_k] = f_k\left(a_{j_{k,1}}[i + c_{k,1}], \dots, a_{j_{k,n_k}}[i + c_{k,n_k}]\right);$

}

}

}

Листинг 3: Вид программного цикла после гнездования

Пусть каждый процессорный элемент целиком выполняет итерации внешнего цикла, приведенного на листинге 3, причем i-ую итерацию выполняет процессорный элемент с номером $\left\lfloor \frac{i}{d} \right\rfloor \ mod\ p$, где d – размер блока, p – число процессорных элементов. Пусть массивы размещены без дублирования, то есть каждый элемент массива размещен только в ПЭ.

Определение 3. Одинарная пересылка — это пересылка одного данного из одного ПЭ в другой ПЭ.

Следует отметить, что циклическая пересылка представляет собой множество одинарных пересылок, выполняемых параллельно.

В данной работе мы рассматриваем задачу поиска минимума одинарных пересылок. Минимум пересылок с учетом свойств коммуникационной сети не превосходит минимума одинарных пересылок.

Пример 3. Рассмотрим программный цикл

```
for (int i = 0; i < 10; ++i) {
  a[i] = b[i + 2];
  c[i] = c[i - 1];
}
```

Пусть параметр гнездования цикла равен 2, тогда после гнездования цикл примет вид

```
for (int j = 0; j < 10; j += 2) {
  for (int i = j; i < j + 2; ++i) {
    a[i] = b[i + 2];
    c[i] = c[i - 1];
  }
}
```

Пусть процессорных элементов $p = 3$, размер блока $d = 2$, элементы массива a размещены $P_a(i) = 2$ (все в одном ПЭ с номером 2), элементы массива b размещены $P_b(i) = (i\ mod\ p)$, элементы массива c размещены $P_c(i) =$

$\left(\left\lfloor\frac{i+1}{d}\right\rfloor\ mod\ p\right)$. Тогда вхождение a[i] требует одинарную пересылку на итерации I = 2 исходного цикла, но не требует одинарную пересылку на итерации 4, вхождение b[i + 2] требует 7 одинарных пересылок, а вхождение c[i] требует одинарную пересылку на каждой итерации вида $1 + 2\ *\ k$, где $k \in \mathbb{Z}$.

В примере 3 массив a размещен блочно-аффинно $P_a(i) = 2 = \left(\left\lfloor\frac{1}{d_1}\right\rfloor * i + 2\right)\ mod\ p$, где d_1 – большое число (не меньшее количества итераций цикла), $s_0 = 2, s_1 = 1;\ P_b(i) = (i\ mod\ p)$ определяет блочно-аффинное размещение с $s_b = 0, d = 1, P_c(i) = \left(\left\lfloor\frac{i+1}{d}\right\rfloor\ mod\ p\right)$ определяет блочно-аффинное размещение с $s_c = -1$.

Пусть массив a размещен блочно-аффинно. Пусть вхождение a[i + c] требует (не требует) пересылки на итерации i_0, тогда на любой итерации вида $i_0 + d * k$, где $k \in \mathbb{Z}$, это вхождение также требует (не требует) пересылки (здесь d – размер блока).

Теорема. *Пусть количество процессорных элементов* $p > 1 + \max\{|c_{i,j}|\}$, *где* $c_{i,j}$ *– константы в индексных выражениях цикла, приведенного в листинге 2, тогда минимальное количество одинарных пересылок достигается при размере блока* $d = \left\lfloor\frac{N}{p}\right\rfloor$.

Минимизация циклических пересылок для размещений массивов с блоками размера 1 рассмотрена в [7].

5. СОЗДАНИЕ ПАРАЛЛЕЛЬНОЙ ПРОГРАММЫ С ПОМОЩЬЮ РАЗМЕЩЕНИЙ И ПЕРЕСЫЛОК

При заданном размещении массивов для гнезда циклов с пересылками можно построить эквивалентную параллельную программу. Для параллельного выполнения можно использовать средства MPI. Размещения массивов по ПЭ, пересылки блоков элементов массивов и распределение итераций циклов между ПЭ отображаются в вызовы функций MPI, объявление вспомогательных массивов и эквивалентные преобразования циклов. Рассмотрим, с помощью каких конструкций и преобразований реализуется параллельная программа. Для описания

блочно-аффинного размещения массива A с заданными параметрами потребуются заведение дополнительного массива, описание пользовательского типа и вызов коллективной операции распределения элементов массива по всем ПЭ. Пользовательский тип служит для задания подмножества элементов массива, которые отправляются в один ПЭ. Тип создается с помощью функций MPI_Type_vector и MPI_Type_contiguous. Например, рассылка элементов в p узлов из одномерного массива X в размещенный на ПЭ массив XX проводится инструкцией

```
MPI_Scatter(X - 2, 1, T, XX + rank - 2, 1, T, 0, MPI_COMM_WORLD);
```

где T – пользовательский тип. Для удобства работы со сдвигами внутри массива здесь предполагается, что массив X – это указатель (на языке C), который указывает внутрь массива большего размера, в котором гарантированно помещаются все нужные элементы. После завершения цикла потребуется собрать элементы обратно в массив, размещенный в главном узле, с помощью операции MPI_Gather

```
MPI_Gather(YY + rank, 1, T, Y, 1, T, 0, MPI_COMM_WORLD);
```

Пересылки элементов массивов внутри цикла возможно реализовать с помощью вызова операции MPI_Sendrecv, в которой используются аналогичные пользовательские типы, построенные с учетом размещений двух массивов, участвующих в пересылке. Для организации пересылки по кольцу ПЭ ранги отправителя и получателя вычисляются с учетом расстояния пересылки.

Пример 4. Распределение итераций цикла между ПЭ можно описать как два последовательных преобразования:

- Гнездование цикла;
- Удаление заголовка внешнего цикла с установкой значения счетчика, равного рангу ПЭ.

Рассмотрим элементарный цикл, к которому можно применить этот подход.

```
for (int j = 0; j < N; ++j) {
  Y[j] = (XX[j - 1] + X1[j] + X2[j + 1]) / 3.0;
}
```

Для параллельного выполнения этот цикл преобразуется к виду

```
for (int j2 = 0; j2 < 11 / p; ++j2) {
  int j = j2 * p + j1 + padding;
  MPI_Sendrecv(XX + rank, 1, T, rank - 1, 0, X1 + rank - 1,
  1, T, MPI_ANY_SOURCE, 0, MPI_COMM_WORLD, &status);
  MPI_Sendrecv(XX + rank, 1, T, rank - 2, 0, X2 + rank, 1,
  T, MPI_ANY_SOURCE, 0, MPI_COMM_WORLD, &status);
  Y[j] = (XX[j - 1] + X1[j] + X2[j + 1]) / 3.0;
}
```

Листинг 4. Цикл после применения всех преобразований

В этом примере дополнительные массивы X1, X2 размещены аналогично с параметром $s_0 = 1$ и 0 соответственно, массивы Y и XX размещены с параметром $s_0 = 1$ и 2. Такое размещение согласовано. Пример служит для демонстрации генерации кода с использованием MPI.

6. ДАЛЬНЕЙШИЕ ИССЛЕДОВАНИЯ

В дальнейших работах предполагается рассмотреть многомерные циклы с многомерными массивами и с аппаратной поддержкой не только циклических сдвигов, но и операции широковещательной рассылки данных.

В данной статье для распараллеливания программного цикла использовалось преобразование «гнездование цикла». В последующих работах предполагается использовать и некоторые другие преобразования для расширения множества эффективно распараллеливаемых программ.

7. ЗАКЛЮЧЕНИЕ

Разработано преобразование программного цикла с автоматическим размещением данных в распределенной памяти и минимизацией межпроцессорных пересылок. Статья представляет собой шаг на пути к созданию оптимизирующих распараллеливающих компиляторов на высокопроизводительные системы на кристалле нового поколения типа «суперкомпьютер на кристалле».

Благодарности

Исследование выполнено при финансовой поддержке гранта Российского научного фонда № 22-21-00671, https://rscf.ru/project/22-21-00671/

СПИСОК ЛИТЕРАТУРЫ

1. *Ammaev S.G., Gervich L.R., Steinberg B.Y.* Combining parallelization with overlaps and optimization of cache memory usage // PaCT 2017: Parallel Computing Technologies, Lecture Notes in Computer Science. 2017. Vol. 10421. P. 257–264.

2. *Gong Z., Chen Z., Szaday Z., Wong D., Sura Z., Watkinson N., Maleki S., Padua D., Veidenbaum A., Nicolau A.* An empirical study of the effect of source-level loop transformations on compiler stability // Proceedings of the ACM on Programming Languages. 2018. P. 1–29.

3. *Гервич Л.Р., Кравченко Е.Н., Штейнберг Б.Я., Юрушкин М.В.* Автоматизация распараллеливания программ с блочным размещением данных // Сибирский журнал вычислительной математики. 2015. Т. 18. №1. С. 41–53.

4. *Bondhugula U.* Automatic distributed-memory parallelization and codegeneration using the polyhedral framework // Technical report ISc-CSA-TR-2011-3. 2011. 10 p.

5. DVM-система разработки параллельных программ. URL: http://dvm-system.org/ru/about/, дата обращения 26.03.2022.

6. *Корнеев В.В.* Параллельное программирование // Программная инженерия. 2022. Т. 13. № 1. С. 3–16.

7. *Krivosheev N.M., Steinberg B.Ya.* Algorithm for searching minimum internode data transfers. // «Procedia Computer Science», 10th International Young Scientist Conference on Computational Science. YSC 2021. 1–3 July 2021. P. 306–313.

8. *Kwon D., Han S., Kim H.* MPI backend for an automatic parallelizing compiler // Proceedings Fourth International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN’99). 06.1999. P. 152–157.
https://doi.org/10.1109/ISPAN.1999. 778932.

9. Epiphany-V: A 1024-core 64-bit RISC processor.
URL: https://parallella.org/2016/10/05/epiphany-v-a-1024-core-64-bit-risc-processor, дата обращения 26.03.2022.

10. SoC Esperanto. URL: https://www.esperanto.ai/technology, дата обращения 26.03.2022.

11. *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.

12. *Прангишвили И.В., Виленкин С.Я., Медведев И.Л.* Параллельные вычислительные системы с общим управлением. М.: Энергоатомиздат, 1983. 312 с.

13. Процессор НТЦ «Модуль».
URL: https://www.cnews.ru/news/top/2019-03-06_svet_uvidel_moshchnejshij_rossijskij_nejroprotsessor, дата обращения 26.03.2022.

14. *Штейнберг Б.Я.* Блочно-аффинные размещения данных в параллельной памяти // Информационные технологии. 2010. №6. С. 36–41.

15. *Штейнберг Б.Я.* Оптимизация размещения данных в параллельной памяти. Ростов-на-Дону: Изд-во Южного федерального университета, 2010. 255 с.

16. *Vasilenko A., Veselovskiy V., Metelitsa E., Zhivykh N., Steinberg B., Steinberg O.* Precompiler for the ACELAN-COMPOS Package Solvers // In: Malyshkin V. (eds). Parallel Computing Technologies. PaCT 2021. Lecture Notes in Computer Science. Vol. 12942. P. 103–116. Springer, Cham.
https://doi.org/10.1007/978-3-030-86359-3_8

17. *Штейнберг О.Б.* Минимизация количества временных массивов в задаче разбиения циклов // Известия ВУЗов. Северо-Кавказский регион. Естественные науки. 2011. №5. С. 31–35.

18. SambaNova Launches Second-Gen DataScale System.
URL: https://www.hpcwire.com/2022/09/14/sambanova-launches-second-gen-datascalesystem, дата обращения 20.01.2023.

19. *Елизаров Г.С., Конотопцев В.Н., Корнеев В.В.* Специализированные большие интегральные схемы для реализации нейросетевых выводов // XXII международная конференция «Харитоновские тематические научные чтения «Суперкомпьютерное моделирование и искусственный интеллект»: сб. трудов / Под ред. Р.М. Шагалиева. Саров: ФГУП «РФЯЦ-ВНИИЭФ», 2022. С. 181–184.

20. *Корнеев В.В.* Направления повышения производительности нейросетевых вычислений // Программная инженерия. 2020. Т. 11, № 1. С. 21–25.
https://doi.org/10.17587/prin.11.21–25

21. *Yen I.E., Xiao Zh., Xu D.* S4: a High-sparsity, High-performance AI Accelerator // arXiv:2207.08006v1 [cs.AR] 16 Jul 2022

22. *Gale T., Elsen E., Hooker S.* The state of sparsity in deep neural networks // arXiv preprint arXiv:1902.09574, 2019

23. Intelligence Processing Unit. URL: https://www.graphcore.ai/products/ipu. (accessed: 20.01.2023)

24. *Jia Zh., Tillman B., Maggioni M., Scarpazza D.P.* Dissecting the Graphcore IPU Architecture via Microbenchmarking // Technical Report. December 7, 2019. arXiv:1912.03413v1 [cs.DC] 7 Dec. 2019. 91 p.

AUTOMATION OF PROGRAM PARALLELIZATION FOR MULTICORE PROCESSORS WITH DISTRIBUTED LOCAL MEMORY

A. P. Bagliy[1] [0000-0001-9089-4164], N. M. Krivosheev[2] [0000-0002-1366-7037],
B. Ya. Steinberg[3] [0000-0001-8146-0479]

[1]*Southern federal university, Faculty of mathematics, mechanics and computer science*

[1]abagly@sfedu.ru, [2]krivosheev@sfedu.ru, [3]byshtyaynberg@sfedu.ru

Abstract

This paper is concerned with development of parallelizing compiler onto computer system with distributed memory. Industrial parallelizing compilers create programs for shared memory systems. Transformation of sequential programs onto systems with distributed memory requires development of new functions. This is becoming topical for future computer systems with hundreds and more cores. Conditions for program loop parallelization onto computer system with distributed memory is formulated in terms of information dependence graph.

Keywords*: automatic parallelization, distributed memory, program transformation, data distribution, data interchange*

REFERENCES

1. *Ammaev S.G., Gervich L R., Steinberg B.Y.* Combining parallelization with overlaps and optimization of cache memory usage // PaCT 2017: Parallel Computing Technologies, Lecture Notes in Computer Science. 2017. Vol. 10421. P. 257–264.

2. *Gong Z., Chen Z., Szaday Z., Wong D., Sura Z., Watkinson N., Maleki S.,*

Padua D., Veidenbaum A., Nicolau A. An empirical study of the effect of source-level loop transformations on compiler stability // Proceedings of the ACM on Programming Languages. 2018. P. 1–29.

3. *Gervich L.R., Kravchenko E.N., Steinberg B.Y., Yurushkin M.V.* Automatic program parallelization with block data distribution // Sibirskiy zhurnal vychislitelnoi matematiki. 2015. Vol. 18. No. 1. P. 41–53.

4. *Bondhugula U.* Automatic distributed-memory parallelization and codegeneration using the polyhedral framework // Technical report ISc-CSA-TR-2011-3. 2011. 10 p.

5. DVM-sistema razrabotki parallel'nyh program.
URL: http://dvm-system.org/ru/about, last accessed 26.03.2022.

6. *Korneev V.V.* Parallel'noe programmirovanie // Programmnaya Ingeneria. 2022. Vol. 13. No. 1. P. 3–16.

7. *Krivosheev N.M., Steinberg B.Ya.* Algorithm for searching minimum internode data transfers // «Procedia Computer Science», 10th International Young Scientist Conference on Computational Science. YSC 2021. 1–3 July 2021. P. 306–313.

8. *Kwon D., Han S., Kim H.* MPI backend for an automatic parallelizing compiler // Proceedings Fourth International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN'99). 06.1999. P. 152–157.
https://doi.org/10.1109/ISPAN.1999. 778932.

9. Epiphany-V: A 1024-core 64-bit RISC processor.
URL: https://parallella.org/2016/10/05/epiphany-v-a-1024-core-64-bit-risc-processor, last accessed 26.03.2022.

10. SoC Esperanto. URL: https://www.esperanto.ai/technology, last accessed 26.03.2022.

11. *Bahtin V.A., Zaharov D.A., Kolganov A.S., Kryukov V.A., Podderyugina N.V., Pritula M.N.* Reshenie prikladnyh zadach s ispol'zovaniem DVM-sistemy // Vestnik YUUrGU. Seriya: Vychislitel'naya matematika i informatika. 2019. T.. 8. № 1. S. 89–106.

12. *Prangishvili I.V., Vilenkin S.YA., Medvedev I.L.* Parallel'nye vychislitel'nye sistemy s obshchim upravleniem. M.: Energoatomizdat, 1983. 312 p.

13. Processor NTC "Modul". URL: https://www.cnews.ru/news/top/2019-03-06_svet_uvidel_moshchnejshij_rossijskij_nejroprotsessor, last accessed 26.03.2022.

14. *Shtejnberg B.Ya.* Blochno-affinnye razmeshcheniya dannyh v parallel'noj pamyati // Informacionnye tekhnologii. 2010. №6. S. 36–41.

15. *Shtejnberg B.Ya.* Optimizaciya razmeshcheniya dannyh v parallel'noj pamyati. Rostov-na-Donu: Izd-vo Yuzhnogo federal'nogo universiteta, 2010. 255 s.

16. *Vasilenko A., Veselovskiy V., Metelitsa E., Zhivykh N., Steinberg B., Steinberg O.* Precompiler for the ACELAN-COMPOS Package Solvers // In: Malyshkin V. (Ed.) Parallel Computing Technologies. PaCT 2021. Lecture Notes in Computer Science. Vol. 12942. P. 103-116. Springer, Cham. https://doi.org/10.1007/978-3-030-86359-3_8

17. *Shtejnberg O.B.* Minimizaciya kolichestva vrevtnnyh massivov v zadache raszbieniya ziklov // Izvestia VUZov. Severo-Kavkazsky region. Estestvennye nauki. 2011. №5. S. 31–35

18. SambaNova Launches Second-Gen DataScale System.
URL: https://www.hpcwire.com/2022/09/14/sambanova-launches-second-gen-datascalesystem, last accessed 20.01.2023.

19. *Elizarov G.S., Konotoptsev V.N., Korneev V.V.* Specialized large integrated circuits for the implementation of neural network inference // XXII International conference "Kharitonov thematic scientific readings "Supercomputer modeling and Artificial Intelligence": proceedings / Edited by R.M. Shagaliev. Sarov: FSUE "RFSC-VNIIEF", 2022. P. 181–184 (in Russian).

20. *Korneev V.V.* Approaches to improving the performance of neural network computing // Programmnaya ingeneria. 2020. Vol. 11. No. 1. P. 21–25.
https://doi.org/10.17587/prin.11.21-25 (in Russian).

21. *Yen I.E., Xiao Zh., Xu D.* S4: a High-sparsity, High-performance AI Accelerator // arXiv:2207.08006v1 [cs.AR] 16 Jul 2022

22. *Gale T., Elsen E., Hooker S.* The state of sparsity in deep neural networks // arXiv preprint arXiv:1902.09574, 2019

23. Intelligence Processing Unit. URL: https://www.graphcore.ai/products/ipu, last accessed: 20.01.2023.

24. *Jia Zh., Tillman B., Maggioni M., Scarpazza D.P.* Dissecting the Graphcore IPU Architecture via Microbenchmarking // Technical Report. December 7, 2019. arXiv:1912.03413v1 [cs.DC] 7 Dec 2019. 91 p.

СВЕДЕНИЯ ОБ АВТОРАХ

БАГЛИЙ Антон Павлович – ст. преподаватель Института математики, механики и компьютерных наук Южного федерального университета

Anton Pavlovich BAGLIY – senior teacher in department of Mathematics, mechanics and computer science of Southern federal university.

email: abagly@sfedu.ru

ORCID 0000-0001-9089-4164

КРИВОШЕЕВ Никита Максимович – студент 1 курса магистратуры Института математики, механики и компьютерных наук Южного федерального университета.

Nikita Maksimovich KRIVOSHEYEV – first-year master's student in department of Mathematics, mechanics and computer science of Southern federal university.

email: krivosheev@sfedu.ru

ORCID 0000-0002-1366-7037

ШТЕЙНБЕРГ Борис Яковлевич – д. т. н, зав. каф., с. н. с. Института математики, механики и компьютерных наук Южного федерального университета.

Boris Yakovlevich STEINBERG – doctor of computer science, head of chair in department of Mathematics, mechanics and computer science of Southern federal university.

email: borsteinb@mail.ru

ORCID: 0000-0001-8146-0479

Материал поступил в редакцию 20 января 2023 года

УДК 004.4

ЦИФРОВАЯ ЭКОСИСТЕМА OntoMath КАК ПОДХОД К ПОСТРОЕНИЮ ПРОСТРАНСТВА МАТЕМАТИЧЕСКИХ ЗНАНИЙ

А. М. Елизаров[1] [0000-0003-2546-6897], А. В. Кириллович[2] [0000-0001-9680-449X], Е. К. Липачёв[3] [0000-0001-7789-2332], О. А. Невзорова[4] [0000-0001-8116-9446]

[1–4]*Казанский федеральный университет, ул. Кремлевская, 35, г. Казань, 420008*
[1, 2]*Казанское отделение Межведомственного суперкомпьютерного центра Российской академии наук, ул. Лобачевского, 2, г. Казань, 420008*

[1]amelizarov@gmail.com, [2]alik.kirillovich@gmail.com, [3]elipachev@gmail.com, [4]onevzoro@gmail.com

Аннотация

Представлены результаты по созданию методов управления математическим знанием в контексте цифровых математических библиотек. Программные инструменты, разработанные на основе этих методов, являются частью цифровой экосистемы OntoMath, в рамках которой осуществляется их взаимодействие. Приведено краткое описание архитектуры экосистемы OntoMath, выделены уровни предметных онтологий и внешних онтологий, а также уровень программных инструментов и сервисов. В отдельную категорию выделены семантические сервисы. Этим термином обозначены программные инструменты, в функционале которых используются запросы к предметным онтологиям для обеспечения управления объектами знаний. Даны общие описания разрабатываемых предметных онтологий: образовательной математической онтологии $\text{OntoMath}^{\text{Edu}}$ и онтологии профессиональной математики $\text{OntoMath}^{\text{PRO}}$. Отражено развитие образовательной онтологии в направлении включения образовательных пререквизитных связей между классами. Среди программных инструментов цифровой экосистемы выделены сервисы поиска по математическим электронным коллекциям, сервис семантического аннотирования математических документов, инструменты семантической разметки образовательных математических документов, а также система автоматической генерации проверочных тестов по математическим образовательным дисциплинам.

В рамках цифровой экосистемы OntoMath развиваются рекомендательные системы специального назначения. В текущей версии экосистемы представлены рекомендательная система формирования списка близких статей, основанная на онтологии OntoMath^{PRO}, рекомендательная система назначения экспертов для поддержки процесса научного рецензирования и рекомендательные системы подбора предметных классификаторов УДК и кодов Mathematics Subject Classification для математических документов. Приведены также результаты, полученные в направлении создания фабрики метаданных цифровой библиотеки, включающей сервисы и инструменты извлечения, уточнения, пополнения и нормализации метаданных документов электронных математических коллекций. Отметим, что экосистема OntoMath разрабатывается как технологическая основа цифровой математической библиотеки Lobachevskii-DML.

***Ключевые слова**: Цифровая экосистема, экосистема OntoMath, цифровая математическая библиотека, Lobachevskii-DML, онтология, математическая онтология OntoMath^{PRO}, образовательная онтология OntoMath^{Edu}.*

ВВЕДЕНИЕ

Повсеместное применение компьютерных технологий в научных исследованиях, наблюдаемое в настоящее время, привело к качественным изменениям в процессах распространения научных знаний. Отметим из них, как наиболее заметные, переход к электронным формам научных документов и возникновение сетевого научного пространства. Совокупность этих изменений ряд авторов определяет как Вторую научную революцию, по аналогии с Первой научной революцией, которую связывают с профессионализацией создания знаний (см., например, [1]).

В настоящее время компьютерные технологии используются на всех этапах жизненного цикла научного документа. Научные издательства и отдельные журналы разрабатывают и используют в своей практике специализированные информационные системы (например, [2, 3]). Получили распространение новые формы научных публикаций, для поддержки жизненного цикла которых разрабатываются новые информационные среды (см. [4, 5]). Существенно изменилась и инфраструктура современных научных изданий – речь уже идет не только о формах и средствах использования информационных технологий, но, прежде всего, о создании

программных платформ, предлагающих систему сервисов для работы с электронным научным контентом (например, [6]).

Особенности, присущие математическим текстам, оказывают влияние на эффективность использования универсальных программных инструментов, и, как правило, с их помощью не удается достичь желаемых результатов. Формульная составляющая – наиболее заметная особенность математических документов. Научные результаты в статьях по математике во многих случаях выражаются именно в формулах, а текст играет вспомогательную роль. В качестве примера на Рис. 1 приведен фрагмент документа, целиком состоящий из формул и почти не содержащий текста. Отметим также, что совершенно одинаковые формулировки теорем могут иметь качественные различия по объявленным в них результатам, таковыми, например, являются теоремы об улучшении оценок приближений или справочники по специальным разделам математики. Поэтому для эффективной работы с математическими документами требуется разработка методов, использующих семантику не только текстов, но и формул [7, 8]. Также при обработке математических документов важно учитывать их логическую структуру. Она определяется строгой последовательностью объектов, таких как определения, леммы, теоремы, доказательства, следствия, примеры, в которых явно или латентно присутствуют связи с объектами других документов.

В рамках проекта построения Всемирной цифровой математической библиотеки (World Digital Mathematics Library – WDML) предложена парадигма управления математическими знаниями и представления математических документов на основе извлечения из них объектов знания и определения семантических связей между ними [9]. Сегодня на основе классификации математических объектов разрабатываются методы и интеллектуальные программные инструменты обработки математических документов. В частности, создаются сервисы, предназначенные для обработки математических формул (см., например, [10–13]). Имеется также ряд подходов к реализации поиска по формулам (например, [8, 14, 15]).

$$A_*^{(2)} = \begin{pmatrix} \Delta^6(\Delta+A)^3 & 0 & 0 & 0 \\ A_{*21}^{(2)} & \Delta^6(\Delta+A)^3 & 0 & 0 \\ A_{*31}^{(2)} & A_{*32}^{(2)} & \Delta^6(\Delta+A)^3 & 0 \\ A_{*41}^{(2)} & A_{*42}^{(2)} & A_{*43}^{(2)} & \Delta^6(\Delta+A)^3 \end{pmatrix}$$

where the elements of the matrices $A_*^{(1)}$ and $A_*^{(2)}$ have the view

$$A_{*11}^{(1)} = A_{*22}^{(1)} = A_{*33}^{(1)} = A_{*44}^{(1)} = A_{*55}^{(1)} = \Delta^8(\Delta+A)^4, \quad A_{*21}^{(1)} = -\frac{3}{h^2}\Delta^7(\Delta+A)^3(3\Delta+2A),$$

$$A_{*12}^{(1)} = A_{*13}^{(1)} = A_{*14}^{(1)} = A_{*15}^{(1)} = 0, \quad A_{*23}^{(1)} = A_{*24}^{(1)} = A_{*25}^{(1)} = A_{*34}^{(1)} = A_{*35}^{(1)} = A_{*45}^{(1)} = 0,$$

$$A_{*31}^{(1)} = -\frac{5}{h^2}\Delta^6(\Delta+A)^2\Big[2\Delta(\Delta+A)(3\Delta+2A) - \frac{21}{h^2}(6\Delta^2+8A\Delta+3A^2)\Big],$$

$$A_{*32}^{(1)} = -\frac{35}{h^2}\Delta^7(\Delta+A)^3(3\Delta+2A), \quad A_{*43}^{(1)} = -\frac{99}{h^2}\Delta^7(\Delta+A)^3(3\Delta+2A),$$

$$A_{*41}^{(1)} = -\frac{21}{h^2}\Delta^5(\Delta+A)\Big[\Delta^2(\Delta+A)^2(3\Delta+2A) - \frac{60}{h^2}\Delta(\Delta+A)(6\Delta^2+8A\Delta+3A^2)$$
$$+ \frac{495}{h^4}(10\Delta^3+20A\Delta^2+15A^2\Delta+4A^3)\Big], \quad A_{*54}^{(1)} = -\frac{195}{h^2}\Delta^7(\Delta+A)^3(3\Delta+2A),$$

$$A_{*42}^{(1)} = -\frac{45}{h^2}\Delta^6(\Delta+A)^2\Big[2\Delta(\Delta+A)(3\Delta+2A) - \frac{77}{h^2}(6\Delta^2+8\Delta A+3A^2)\Big],$$

$$A_{*51}^{(1)} = -\frac{9}{h^2}\Delta^4\Big[4\Delta^3(\Delta+A)^3(3\Delta+2A) - \frac{770}{h^2}\Delta^2(\Delta+A)^2(6\Delta^2+8A\Delta+3A^2)$$
$$+ \frac{30030}{h^4}\Delta(\Delta+A)(10\Delta^3+20A\Delta^2+15A^2\Delta+4A^3) - \frac{225225}{h^6}(15\Delta^4+40A\Delta^3$$
$$+ 45A^2\Delta^2+24A^3\Delta+5A^4)\Big], \quad A_{*52}^{(1)} = -\frac{165}{h^2}\Delta^5(\Delta+A)\Big[\Delta^2(\Delta+A)^2(3\Delta+2A)$$
$$- \frac{156}{h^2}\Delta(\Delta+A)(6\Delta^2+8A\Delta+3A^2) + \frac{4095}{h^4}(10\Delta^3+20A\Delta^2+15A^2\Delta+4A^3)\Big],$$

$$A_{*53}^{(1)} = -\frac{117}{h^2}\Delta^6(\Delta+A)^2\Big[2\Delta(\Delta+A)(3\Delta+2A) - \frac{165}{h^2}(6\Delta^2+8A\Delta+3A^2)\Big];$$

$$A_{*11}^{(2)} = A_{*22}^{(2)} = A_{*33}^{(2)} = A_{*44}^{(2)} = \Delta^6(\Delta+A)^3, \quad A_{*12}^{(2)} = A_{*13}^{(2)} = A_{*14}^{(2)} = 0,$$

$$A_{*21}^{(2)} = -\frac{15}{h^2}\Delta^5(3\Delta+2A)(\Delta+A)^2, \quad A_{*23}^{(2)} = A_{*24}^{(2)} = A_{*34}^{(2)} = 0,$$

$$A_{*31}^{(2)} = -\frac{21}{h^2}\Delta^4(\Delta+A)\Big[2\Delta(\Delta+A)(3\Delta+2A) - \frac{45}{h^2}(6\Delta^2+8\Delta A+3A^2)\Big],$$

$$A_{*32}^{(2)} = -\frac{63}{h^2}\Delta^5(\Delta+A)^2(3\Delta+2A), \quad A_{*41}^{(2)} = -\frac{27}{h^2}\Delta^3\Big[3\Delta^2(\Delta+A)^2(3\Delta+2A)$$
$$- \frac{308}{h^2}\Delta(\Delta+A)(6\Delta^2+8A\Delta+3A^2) + \frac{5005}{h^4}(10\Delta^3+20A\Delta^2+15A^2\Delta+4A^3)\Big];$$

$$A_{*42}^{(2)} = -\frac{77}{h^2}\Delta^4(\Delta+A)\Big[2\Delta(\Delta+A)(3\Delta+2A) - \frac{117}{h^2}(6\Delta^2+8A\Delta+3A^2)\Big];$$

$$A_{*43}^{(2)} = -\frac{143}{h^2}\Delta^5(3\Delta+2A)(\Delta+A)^2.$$

Applying the differential matrix operator $A_*^{(I)^T}$ from the left to the equations (2.27) and taking into account $A_*^{(I)^T}A^{(I)} = A^{(I)}A_*^{(I)^T} = E^{(I)}|A^{(I)}|$, $\langle I = 1, 2\rangle$, where $E^{(1)}$ and $E^{(2)}$ are the unit matrices of the fifth and fourth orders respectively, we will obtain decomposed systems of equations with respect to

Рис. 1. Фрагмент математического документа, содержание которого определяется формулами.

Методы выделения в Сети объектов научного знания, разрабатываемые в настоящее время, позволяют создавать новые структуры математических знаний,

в частности, графы научного сотрудничества, рекомендательные системы, автоматически формирующие списки «близких» (в определенном смысле) документов, выполняя при этом аннотирование как документов, так и объектов, извлеченных из них (например, [16–18]).

Важное направление в области семантического представления научных знаний связано с разработкой онтологий предметных областей, в частности, в области математического знания (см. [7, 19, 20]). Наиболее важные задачи в управлении математическими знаниями выделены в работах [7, 21, 22]. Как отмечено в ряде работ, определяющая часть этих задач может быть решена с помощью цифровых математических библиотек, построенных с использованием семантических технологий (см., например, [10, 23, 24]).

В [25, 26] введен термин Big Math для обозначения области создания методов и разработки программных систем поддержки математических исследований; проведена аналогия с широко известным термином Big Data, а также предложено рассматривать пять основных направлений разработки методов Big Math: *Выводимость* (*Inference*), *Вычисления (Computation), Табулирование* (*Tabulation*), *Нарратив* (*Narration*), *Организация* (*Organization*). Направление *Выводимость* включает методы вывода утверждений путем дедукции, направление *Вычисления* объединяет алгоритмические преобразования представлений математических объектов в формы, более легкие для понимания. *Табулирование* обозначает создание статических, конкретных данных, относящихся к математическим объектам и структурам, которые можно легко хранить, запрашивать и совместно использовать. *Нарратив* включает методы приведения результатов в форму, которая может быть усвоена людьми, а *Организация* – методы модульной организации математических знаний.

Термин «экосистема» широко используется в научных публикациях для обозначения систем, имеющих признаки самоорганизации. Этот термин введен А. Тенсли в 1935 году в статье [27] для обозначения биологических систем, которые обладают способностью к изменениям, включая масштабируемость своей архитектуры, и не теряют при этом свойства надежности и способности решать сложные

динамические проблемы. В публикациях по информационным технологиям термин «цифровая экосистема» применяют к информационным системам как цифровой аналог биологических экосистем (см., например, [28–31]).

В работах [32, 33] описаны назначение и принципы организации цифровой математической библиотеки Lobachevskii Digital Mathematical Library (Lobachevskii-DML, http://www.lobachevskii-dml.ru/). Эта цифровая библиотека через систему сформированных метаданных и семантических отношений объединяет электронные коллекции математических документов и предоставляет сервисы навигации по понятиям и объектам, извлеченным из этих документов. Названные сервисы базируются на цифровой экосистеме OntoMath, которая в свою очередь обеспечивает взаимодействие онтологий, инструментов текстовой аналитики и приложений для управления объектами математического знания. Эта экосистема впервые была представлена в [34–36], в настоящей работе описано её текущее состояние.

1. СТРУКТУРА ЦИФРОВОЙ ЭКОСИСТЕМЫ OntoMath

Цифровая экосистема OntoMath — это экосистема онтологий, инструментов текстовой аналитики и сервисов для управления математическим знанием [36, 37]. Эта экосистема является технологической платформой цифровой математической библиотеки Lobachevskii-DML. Архитектура современной версии цифровой экосистемы OntoMath представлена на Рис. 2.

Рис. 2. Архитектура цифровой экосистемы OntoMath. Синим цветом обозначены онтологии, разработанные в рамках проекта OntoMath; серым цветом – внешние онтологии; зеленым цветом – платформа семантической публикации, коричневым – сервисы экосистемы. Компоненты, по уровню расположенные выше, базируются на нижестоящих.

Перечислим компоненты, составляющие основу цифровой экосистемы OntoMath.

- Внешние онтологии:
  - онтология AKT Portal Ontology (https://www.w3.org/archive/www.aktors.org/ontology/) используется для представления метаданных научных статей, включая такие классы, как организации, университеты, исследователи и публикации, в формате Открытых связанных данных (Linked Open Data, LOD) (см., например, [38]);
  - онтология SALT Document Ontology (SALT – Semantically Annotated LaTeX), средствами которой определяется семантика отдельных сегментов научных документов, представленных в нотации LaTeX (см. [39–41]);
- Онтология логической структуры математических документов Mocassin (https://code.google.com/archive/p/mocassin/) [42, 43];
- Онтология профессиональной математики $\text{OntoMath}^{\text{PRO}}$ (https://github.com/CLLKazan/OntoMathPro/) [44];
- Математическая образовательная онтология $\text{OntoMath}^{\text{Edu}}$ (https://github.com/CLLKazan/OntoMathEdu) [45–47];
- Платформа семантической публикации [48];
- Сервис семантического поиска по математическим формулам (https://lobachevskii-dml.ru/mathsearch) [8];
- Рекомендательная система для коллекций физико-математических документов [17];
- Сервисы категоризации и классификации математических документов, построенные с использованием предметных классификаторов УДК и кодов Mathematics Subject Classification 2020, а также основанные на применении этих классификаторов сервисы автоматизированного подбора экспертов в информационной журнальной системе (см. [49–54]);
- Сервис семантической разметки математических учебных материалов и справочная база данных [47, 55];
- Параллельный формальный/неформальный корпус математических утверждений [56];
- Сервис автоматической генерации тестовых вопросов [57];

- Фабрика метаданных цифровой библиотеки, содержащая инструменты автоматизации формирования метаданных документов электронных математических коллекций [58–64].

2. ОНТОЛОГИИ ЦИФРОВОЙ ЭКОСИСТЕМЫ OntoMath

Опишем подробнее разработанные математические онтологии, включенные в цифровую экосистему OntoMath.

Онтология OntoMath^{PRO} – онтология профессионального математического знания [44, 65], организованная в виде двух иерархий:

- иерархии разделов математики (*Математическая логика*, *Теория множеств*, *Алгебра*, *Геометрия*, *Топология* и т. д.);
- иерархии элементов математического знания (*Множество*, *Функция*, *Интеграл*, *Элементарное событие*, *Многочлен Лагранжа* и т. д.).

Фрагмент иерархии элементов математического знания представлен на Рис. 3. На Рис. 4 представлен один из разделов этой иерархии – таксономия «Элемент теории дифференциальных уравнений».

Онтология определяет пять типов отношений между концептами:

- Класс → Подкласс (*Уравнение смешанного типа → Уравнение Трикоми*);
- Область математики → Математический объект *(Теория дифференциальных уравнений → Уравнение Гельмгольца*);
- Определяется с помощью (*Символ Кристоффеля → Связность*);
- Ассоциативная связь (*Циклический итерационный метод Чебышева → Численное решение системы линейных уравнений*);
- Задача → Метод решения (*Система линейных уравнений → Метод Гаусса*).

Описание концепта содержит: название на русском и английском языках; определение; связи с другими концептами и ссылки на внешние ресурсы из наборов облака Открытых связанных данных (LOD).

В настоящее время идет работа над новой версией онтологии, основанной на новой архитектуре [65]. Архитектура новой версии онтологии была протестирована на практике при создании образовательной онтологии OntoMath^{Edu} (см. следующий раздел).

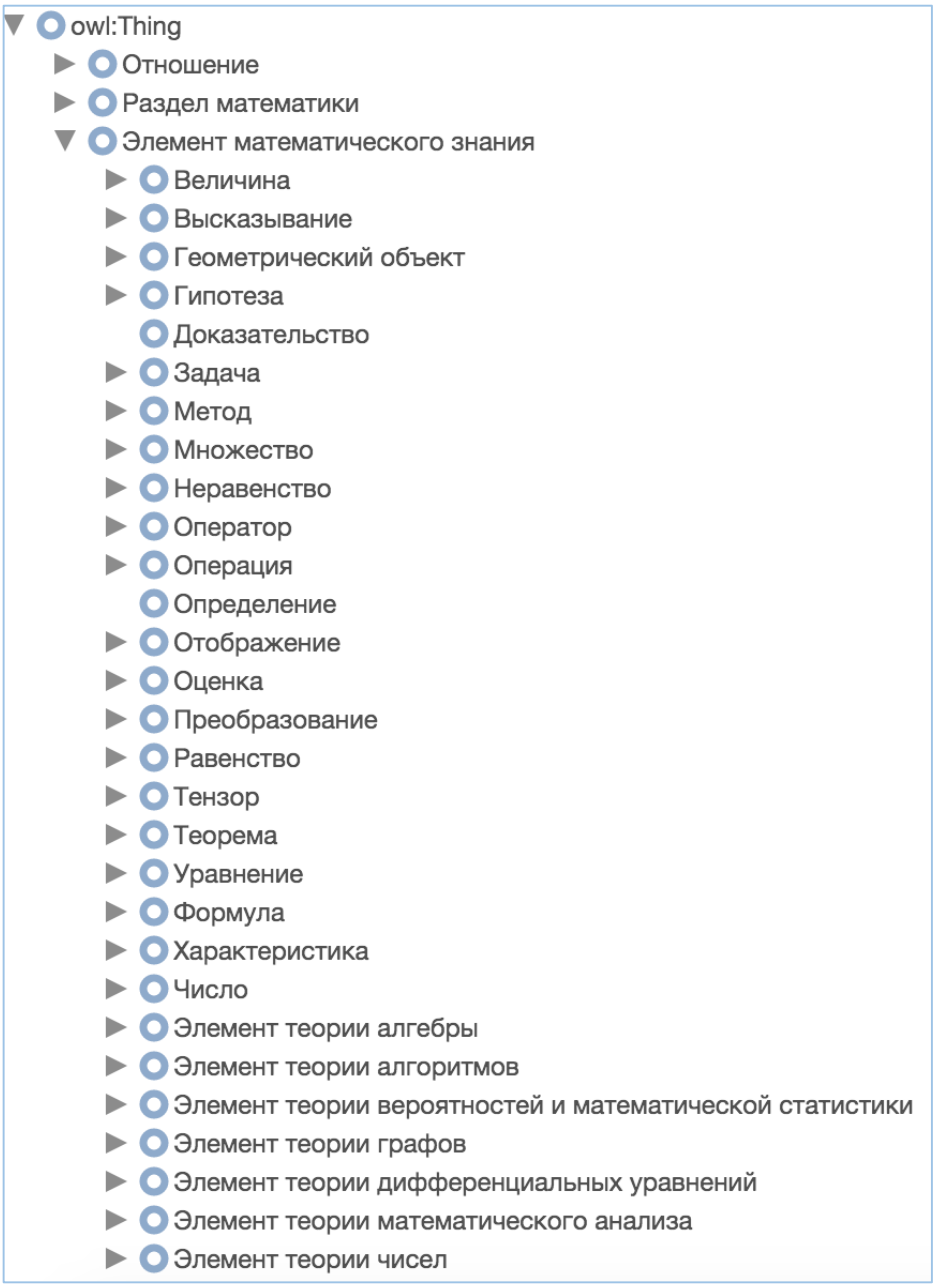


Рис. 3. Фрагмент иерархии элементов математического знания в онтологии OntoMath^{PRO}

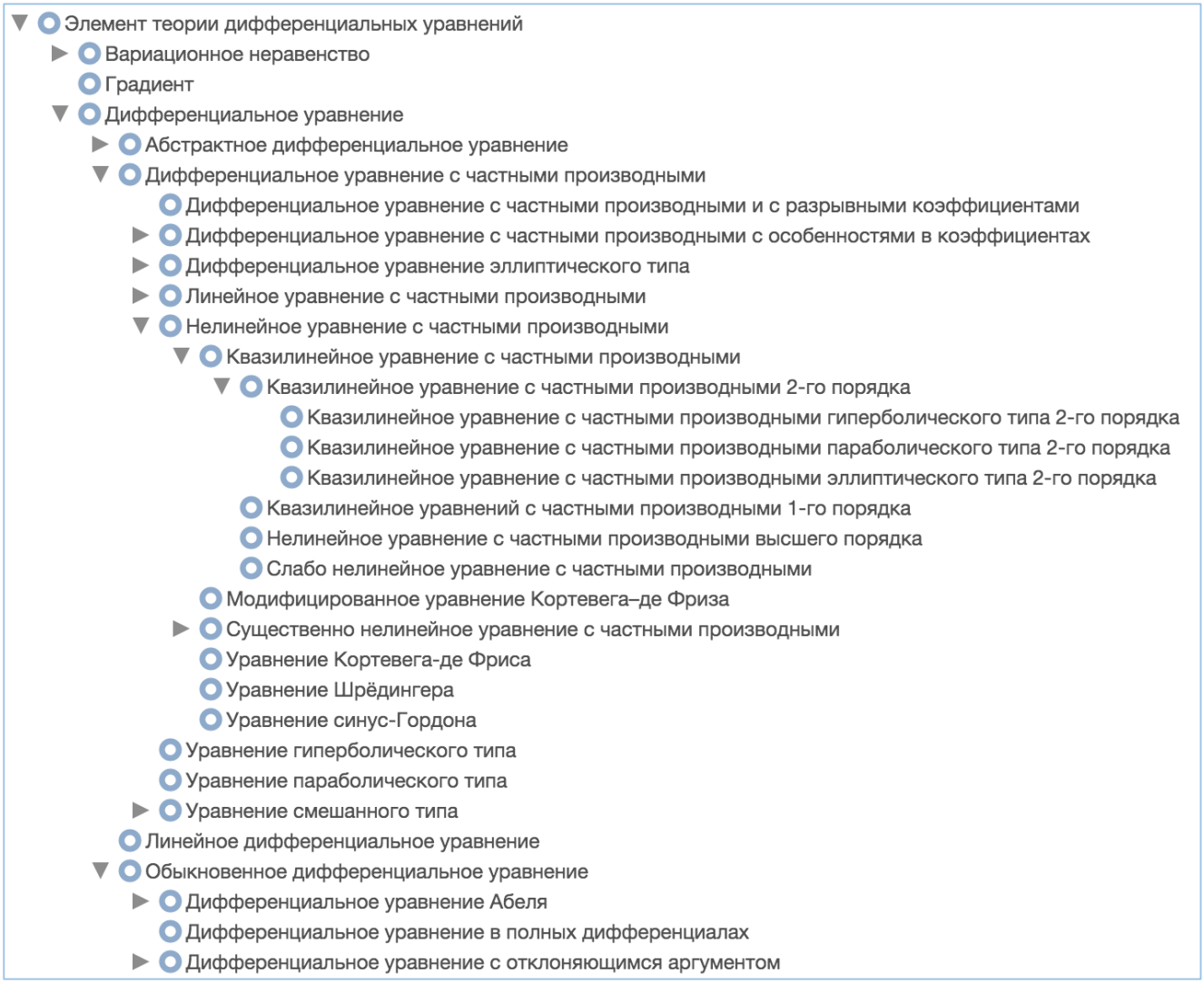


Рис. 4. Фрагмент одного из разделов иерархии элементов математического знания в онтологии OntoMath^{PRO}

Онтология OntoMath^{Edu} — образовательная математическая онтология [45–47]. Она организована в виде трех уровней: уровень предметной онтологии, лингвистический уровень и мета-онтологический уровень. Уровень предметной онтологии содержит независимые от языка концепты, относящиеся к школьной математике. Лингвистический уровень состоит из многоязычных лексиконов, которые содержат информацию о том, как концепты онтологии выражаются в естественном языке (русском, английском, татарском и испанском). Мета-онтологический уровень снабжает концепты мета-онтологическими аннотациями, определенными в онтологии верхнего уровня UFO [66]. Общая архитектура онтологии приведена на Рис. 5.

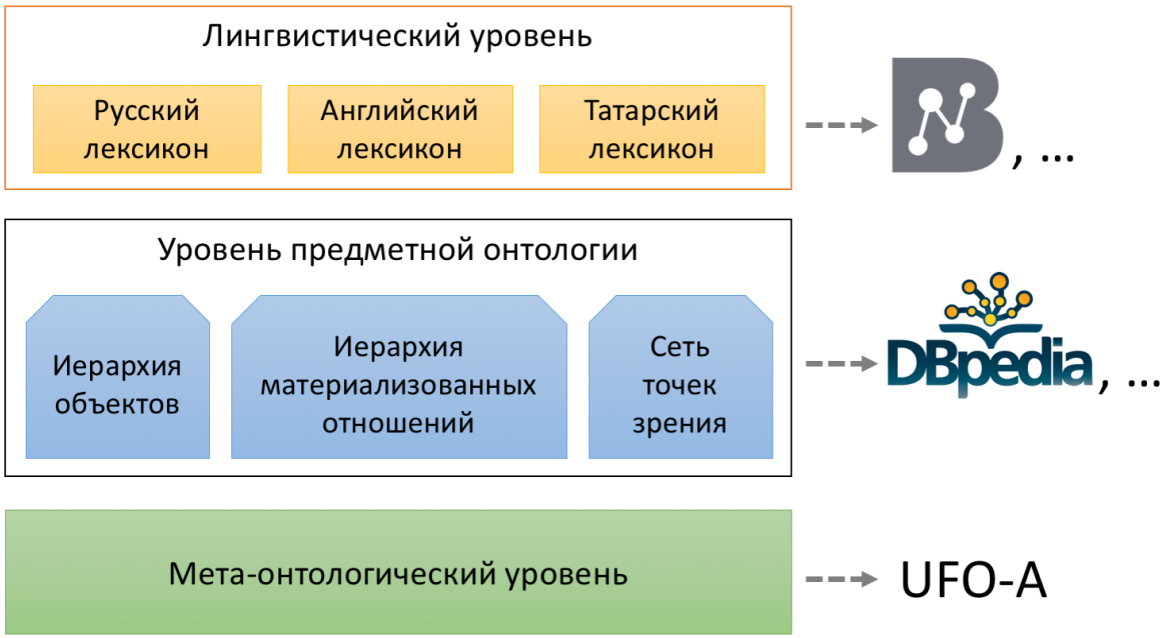


Рис. 5. Общая архитектура онтологии OntoMath^{Edu}

Концепты уровня предметной онтологии организованы в две иерархии: иерархия объектов и иерархия материализованных отношений. На Рис. 6 приведен фрагмент иерархии объектов. На Рис. 7 приведен фрагмент иерархии материализованных отношений.

Чтобы эта онтология могла быть использована в образовательных приложениях, таксономические отношения между концептами были дополнены пререквизитными отношениями (см., например, [47, 67, 68]). Эти отношения отражают, каким образом концепты изучаются в актуальном образовательном процессе. Концепт *A* является пререквизитом концепта *B*, если для того, чтобы изучить концепт *B,* необходимо сначала изучить концепт *A*. Пререквизитные отношения являются независимыми по отношению к таксономическим отношениям и образуют независимую иерархию онтологии. Так, например, концепт *Натуральное число* является пререквизитом как для нижестоящего концепта *Простое число*, так и для вышестоящего концепта *Действительное число*.

На Рис. 8 представлен фрагмент сети пререквизитных отношений в курсе планиметрии российской средней школы.

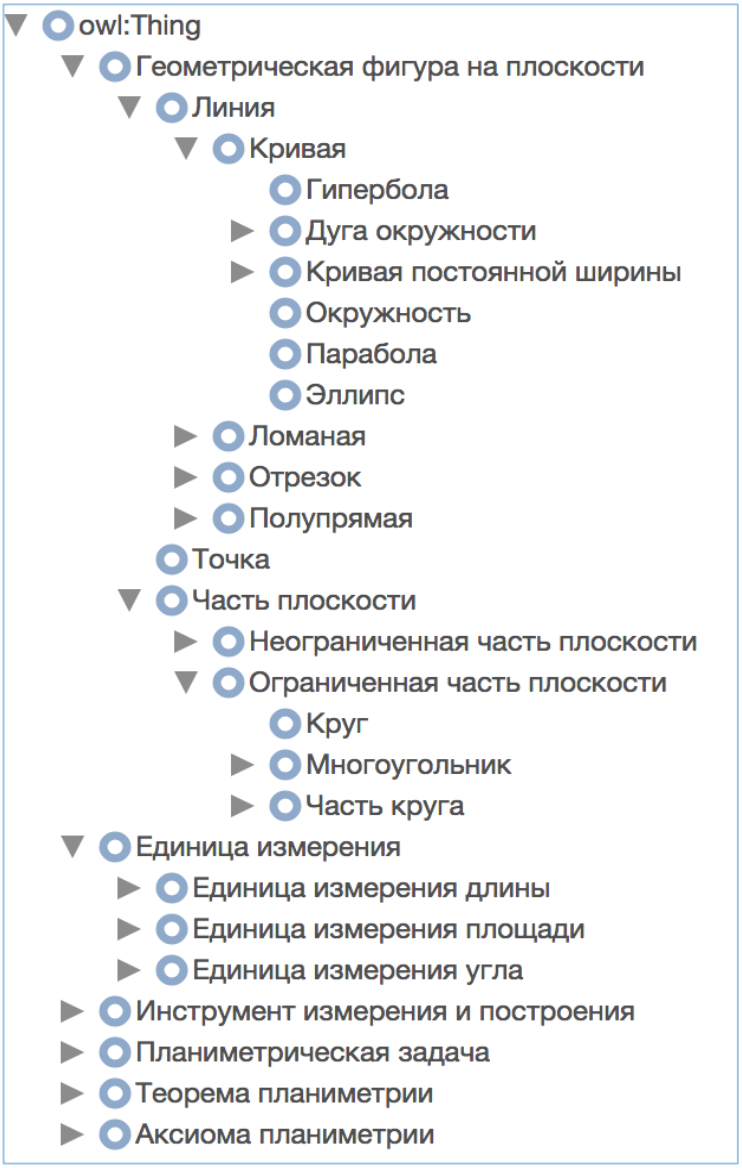


Рис. 6. Фрагмент иерархии объектов образовательной математической онтологии OntoMath^{Edu}

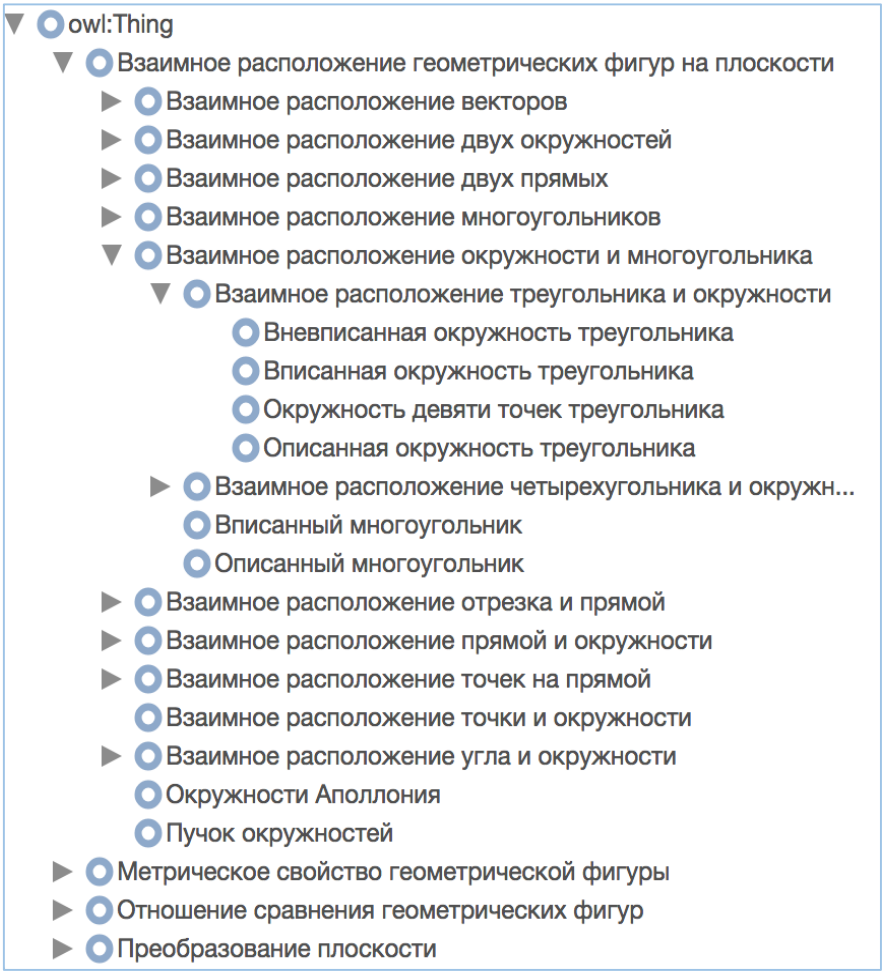


Рис. 7. Фрагмент иерархии материализованных отношений онтологии OntoMath[Edu]

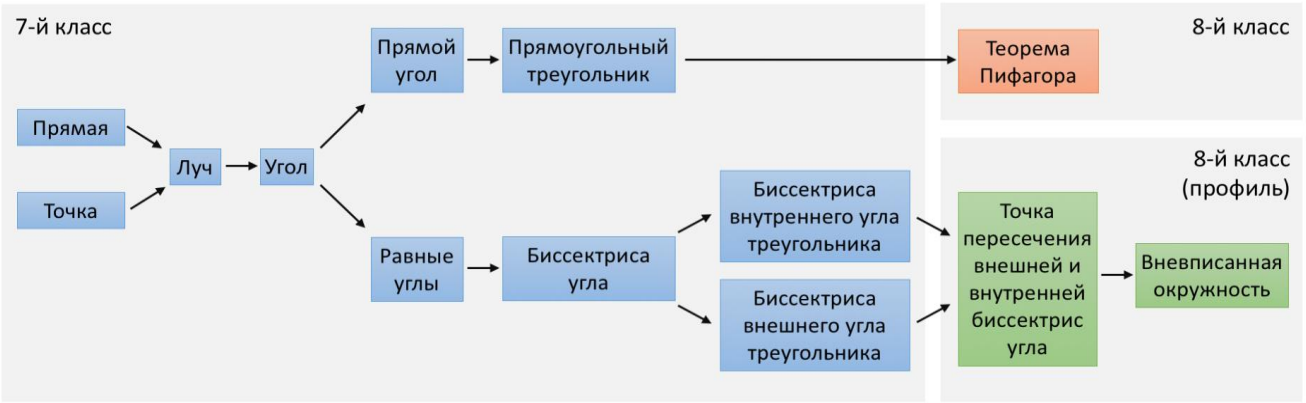


Рис. 8. Фрагмент сети пререквизитных отношений онтологии OntoMath[Edu]. Концепты изображены в виде цветных прямоугольников; прямые пререквизитные отношения между концептами — в виде стрелок; образовательные уровни (7-й, 8-й и 8-й профильный классы) — в виде серых прямоугольников на заднем плане

В отличие от таксономических отношений пререквизитные отношения являются не универсальными, а привязанными к определенной образовательной программе. Так, согласно одной программе, концепт *A* может выступать в качестве пререквизита для концепта *B*, а согласно другой программе, наоборот, концепт *B* может выступать как пререквизит концепта *A*.

В текущей версии онтологии описаны сети пререквизитных отношений, соответствующие образовательным программам средней школы России и Великобритании.

В онтологии OntoMath^{Edu} существуют два подхода для определения пререквизитных отношений: прямой и опосредованный. В соответствии с прямым подходом пререквизитные отношения устанавливаются напрямую между двумя концептами. В соответствии с опосредованным подходом пререквизитные отношения между концептами устанавливаются посредством распределения концептов онтологии по образовательным уровням.

Образовательные уровни — это упорядоченные сегменты образовательной программы, такие как классы в программе общеобразовательной школы, курсы в программе механико-математического факультета и т. д. Как и концепты, образовательные уровни связаны отношениями пререквизитов. Образовательный уровень *L1* является пререквизитом для образовательного уровня *L2*, если обучающийся должен изучить содержание уровня *L1* прежде чем приступать к изучению содержания уровня *L2*.

На Рис. 9. изображены некоторые образовательные уровни российской образовательной программы средней школы и пререквизитные отношения между ними.

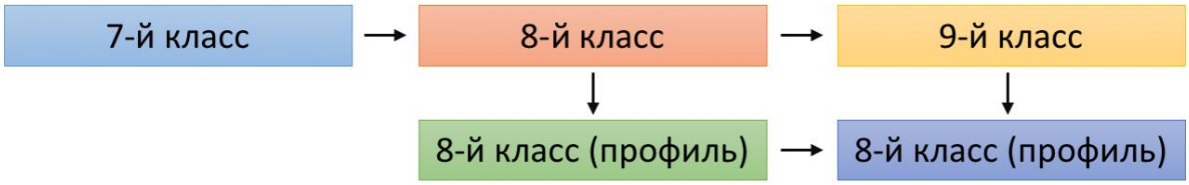


Рис. 9. Фрагмент сети образовательных уровней российской образовательной программы

Образовательные системы и образовательные уровни определены в онтологии как виды точек зрения, которые используются для релятивизации утверждений, не являющихся универсальными. Примерами точек зрения являются системы определений, т. к. один и тот же концепт может быть определен разными способами.

Онтология OntoMath^{Edu} лежит в основе математической образовательной платформы Казанского федерального университета и используется в ряде сервисов экосистемы OntoMath, описанных ниже.

3. ПЛАТФОРМА СЕМАНТИЧЕСКОЙ ПУБЛИКАЦИИ

Центральным компонентом экосистемы OntoMath является программная платформа семантической публикации [48]. Эта платформа предназначена для построения RDF-набора Открытых связанных данных (Linked Open Data, https://lod-cloud.net/) по заданной коллекции математических научных статей в формате LaTeX. Построенный RDF-набор включает:

- метаданные документов, представленные на основе онтологии AKT Portal [69];
- структурные элементы математических документов, представленные на основе онтологии Mocassin (https://code.google.com/archive/p/mocassin/);
- терминологию, представленную с помощью онтологии OntoMath^{PRO};
- формулы, привязанные к терминам.

Построение RDF-набора состоит из следующих шагов:

1) Конвертация исходных документов из формата LaTeX в формат XML.

2) Извлечение метаданных. Для формирования метаданных научных документов в терминах онтологии AKT Portal разработан специализированный программный модуль, с помощью которого решаются следующие задачи: извлечение метаданных из заголовков статей (название, имена авторов и их места работы, название журнала, год публикации и номер журнала); создание идентификаторов для опубликованных статей; обработка библиографических описаний статей с использованием построенных идентификаторов.

3) Аннотирование текста. Данный шаг включает решение стандартных лингвистических задач, таких как токенизация, разделение предложений, морфологический анализ и извлечение именных групп.

4) Извлечение именованных математических сущностей. Для извлечения математической терминологии разработаны специальные программные инструменты, которые используют синтаксические модели именных групп для извлечения из научных статей математических именованных сущностей, привязанных к концептам онтологии OntoMath^{PRO}.

5) Извлечение логической структуры документа. Элементы этой логической структуры размечаются в терминах онтологии Mocassin, которая описывает семантику типичных для научных математических статей структурных элементов, таких как теоремы, леммы, доказательства, определения, следствия и др. Каждый структурный элемент характеризуется своим расположением в тексте, текстовым и формульным содержанием, а также уникальной функциональной нагрузкой (см., например, [42, 43]).

6) Связывание именованных математических сущностей с формулами. Для обработки математических формул разработан программный сервис, с помощью которого осуществляются разбор математических формул в LaTeX-нотации, выделение переменных формулы и связывание выделенных переменных с математическими именованными сущностями, которые данные переменные обозначают.

7) Генерация RDF-набора данных.

8) Связывание RDF-набора с внешними ресурсами из облака Открытых связанных данных (LOD).

Сгенерированный RDF-набор имеет самостоятельную ценность и может использоваться для навигации [70], поиска, агрегирования данных и решения других задач. Кроме того, этот набор лежит в основе ряда сервисов экосистемы OntoMath, описанных далее.

4. СИСТЕМА СЕРВИСОВ ЦИФРОВОЙ ЭКОСИСТЕМЫ OntoMath

Приведем краткое описание семантических сервисов, входящих в настоящее время в состав цифровой экосистемы OntoMath.

Сервис семантического поиска по математическим формулам. Сервис позволяет находить формулы, релевантные заданному математическому понятию, вне зависимости от его символического представления [8]. В качестве поискового

запроса пользователь вводит название интересующего его математического понятия (например, *Угол*, *Граф* или *Простое число*). Сервис возвращает список формул, которые содержат переменные, обозначающие данное понятие.

Пример применения названного поискового сервиса представлен в виде таблицы, в которой каждая строка соответствует найденной формуле (см. Рис. 10). Первая колонка таблицы содержит переменную, которая обозначает искомое понятие в найденной формуле. Вторая колонка содержит саму найденную формулу. Третья колонка содержит элемент логической структуры документа, в котором находится данная формула (например, теорема, доказательство, лемма, утверждение, и т. д.). Четвертая колонка содержит кнопку для отображения окна с метаданными формулы.

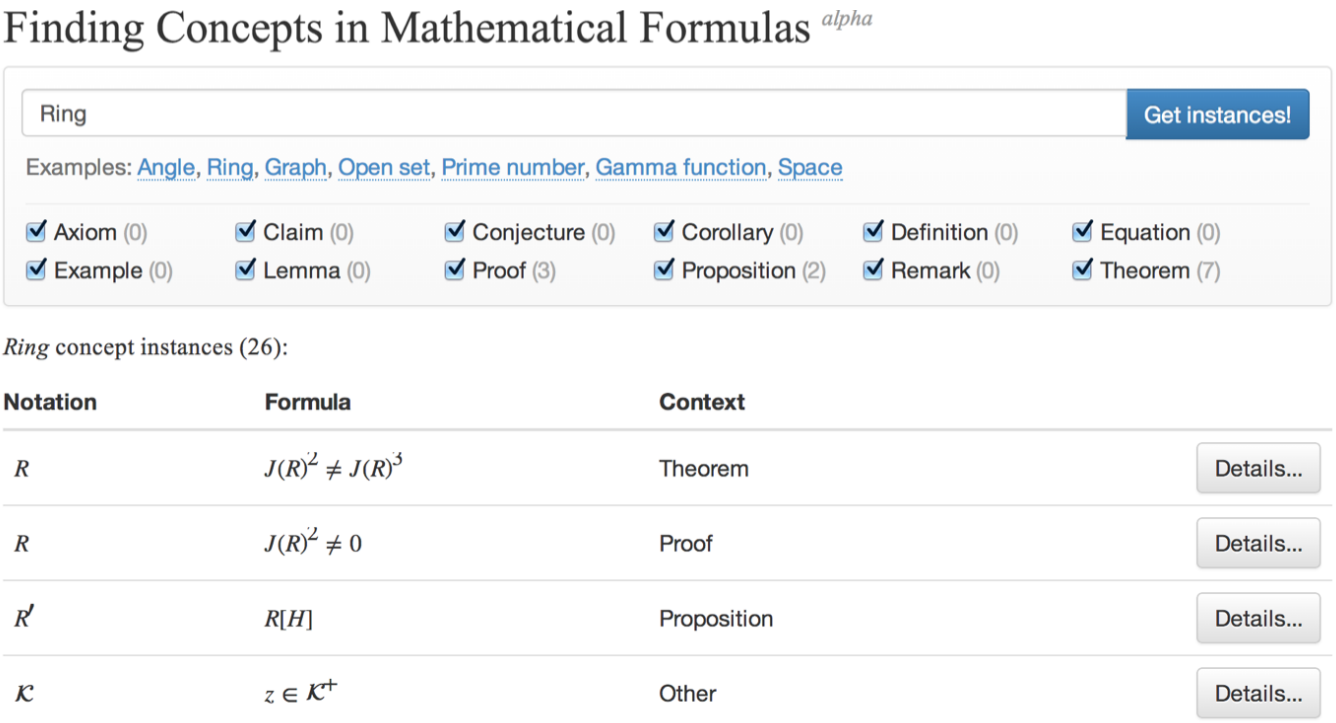


Рис. 10. Результаты семантического поиска по запросу *Ring* – включают обозначения переменной, связанной с концептом *Ring*, формулы и контекст формул.

Пользовательский интерфейс сервиса позволяет фильтровать результаты поиска на основе элемента логической структуры документа, в котором они нахо-

дятся (например, отображать только те формулы, которые находятся в определениях и формулировках теорем). Сервис работает на базе RDF-набора, построенного платформой семантической публикации.

Рекомендательная система для коллекций физико-математических документов. Этот сервис работает с коллекцией математических документов и для каждого документа коллекции формирует список близких статей [17]. Список близких статей помогает пользователю, заинтересовавшемуся некоторой публикацией, найти другие публикации по схожей теме.

Построение списка близких статей происходит в несколько этапов. На первом этапе сервис извлекает из документов термины, определенные в онтологии OntoMath^{PRO}. На втором этапе сервис размечает элементы логической структуры документов, определенные в онтологии Mocassin (такие, как теорема, доказательство, определение и т. д.). На третьем этапе сервис строит векторное представление документа, учитывающее терминологический состав, положение терминов в логической структуре и связи терминов в графе онтологии OntoMath^{PRO}. Далее сервис вычисляет меру близости между векторами документов и на ее основе строит для каждого документа коллекции список близких к нему документов.

Сервис работает на базе RDF-набора, построенного платформой семантической публикации.

Рекомендательная система назначения классификаторов УДК математическим статьям [49]. Классификация документов с присвоением кодов-классификаторов является традиционным способом систематизации и поиска документов по определенной тематике. Универсальная десятичная классификация (УДК) лежит в основе систематизации знаний, представленных в библиотеках, базах данных и других хранилищах информации. В России УДК является обязательным реквизитом всей книжной продукции и информации по естественным и техническим наукам. Выбор классификационных кодов связан с анализом структуры дерева классификатора и традиционно выполняется автором научной статьи. Разработанная рекомендательная система автоматически выполняет подбор классификационного кода УДК для математической статьи на основе онтологии OntoMath^{PRO} с помощью создания «кодовых карт» для каждого классифицирующего кода в дереве УДК в области математики. Под «кодовой картой» понимается взвешенный набор всех

математических именованных сущностей, извлеченных с помощью онтологии OntoMath^{PRO} из коллекции статей с заданным кодом УДК. Создание «кодовых карт» основано на гипотезе о том, что выбор кода УДК обусловлен определённым набором классифицирующих признаков, которые можно представить классами из онтологии OntoMath^{PRO}. Названная гипотеза проверена и подтверждена в ряде экспериментов, проведенных на коллекции математических статей, опубликованных в журнале «Известия ВУЗов. Математика» за 1999–2009 гг.

Система сервисов уточнения предметных классификаторов. Для реализации этого сервиса создан словарь терминов, ассоциированных с классификаторами УДК, относящимися к физико-математическим областям знания. Словарь содержит как классификаторы УДК, так и наборы ключевых терминов, по которым производятся систематизация и классификация материала (см. [71–73]). Большая часть этих терминов была получена путем автоматизированной обработки физико-математических коллекций Общероссийского математического портала Math-Net.Ru (https://www.mathnet.ru/). Извлечение терминов из документов электронных коллекций производится с помощью разработанных программных инструментов, учитывающих стилевые и структурные особенности документов (см., например, [62, 63, 74, 75]).

Автоматизированный подбор экспертов в информационной журнальной системе. В редакции журнала Lobachevskii Journal of Mathematics (https://ljm.kpfu.ru/) на протяжении ряда лет разрабатывается система автоматического подбора экспертов для проведения научного рецензирования статей, поступающих в редакцию журнала (см. [50–54, 76]). Текущая версия сервиса автоматического поиска рецензентов инкапсулирована в информационную журнальную систему Open Journal System (см., например, [77]). Алгоритм работы сервиса основан на использовании таксономии Mathematics Subject Classification 2020 с возможностью преобразования кодов из предыдущих версий этой математической системы классификации (см., например, [78 – 80]).

Сервис семантического аннотирования учебных материалов и справочная база данных. Этот сервис находит в тексте учебных материалов упоминания математических концептов из онтологии OntoMath^{Edu} и связывает найденные упомина-

ния ссылками на соответствующие страницы справочной базы данных. Пользователь, заинтересовавшийся некоторым концептом, может перейти по ссылке и открыть соответствующую страницу [55]. Страница справочной базы данных содержит подробную информацию о концепте, в том числе название концепта, его определение, положение в иерархии концептов и связи с другими концептами [47]. На Рис. 11 изображена страница концепта *Параллелограмм*.

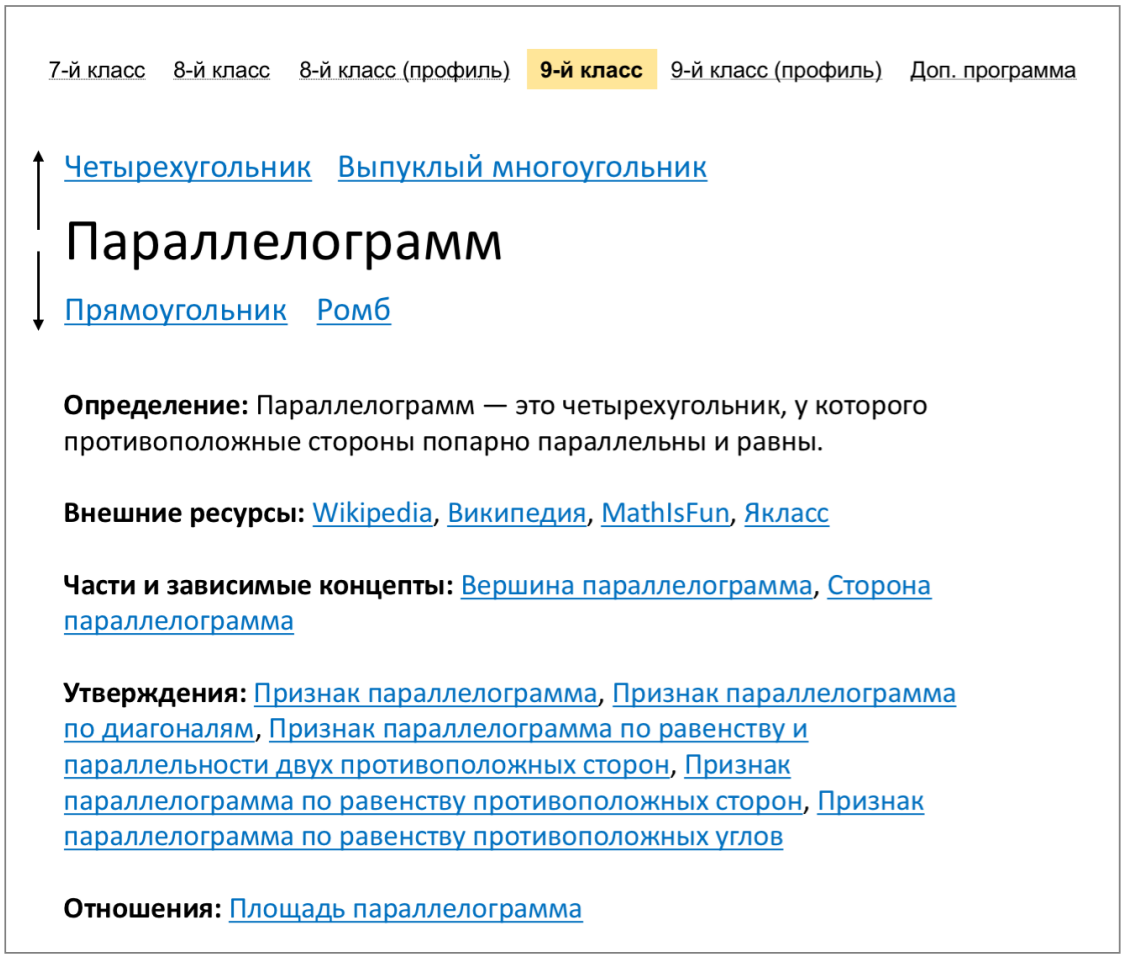


Рис. 11. Страница концепта *Параллелограмм* в справочной базе данных. Представление концепта соответствует образовательному уровню *9-й класс*

По умолчанию, на странице концепта отображается не вся информация о нем, а только та информация, которая соответствует текущему образовательному уровню пользователя. Например, на Рис. 11 страница концепта *Параллелограмм* содержит информацию об этом концепте, соответствующую образовательному уровню *9-й класс*. В частности, список подклассов этого концепта содержит концепты *Прямоугольник* и *Ромб*, которые изучаются в *9-м классе* или ранее, но не

содержит концепт *Параллелограмм Вариньона*, который изучается только в *9-м профильном классе*. С помощью меню, размещенному наверху страницы, пользователь может переключиться на представления концепта, относящиеся к другим образовательным уровням.

В качестве источника информации о концепте и образовательных уровнях справочная база данных использует онтологию OntoMath^{Edu}.

Параллельный формальный/неформальный корпус математических утверждений. Данный ресурс содержит математические утверждения, представленные одновременно с тремя разными степенями формализации [56]. Корпус представляет собой коллекцию записей, каждая из которых содержит следующие поля:

1. математическое утверждение на естественном языке, извлеченное из учебных математических текстов;
2. представление этого утверждения в виде формулы в формате LaTeX;
3. формализация формулы в формате представления семантики математических объектов OpenMath (https://openmath.org/) [81], где в качестве контентных словарей OpenMath (OpenMath content dictionaries) использована онтология OntoMath^{Edu}.

На Рис. 12. представлен пример записи «Сумма углов треугольника равна 180°» из указанного корпуса.

Утверждение на естественном языке	Представление утверждения в виде формулы (в визуальном виде)	Представление утверждения в виде формулы (LaTeX-код)	Формализация формулы в формате OpenMath (в визуальном виде)
Сумма углов треугольника равна 180°.	$\angle A + \angle B + \angle C = 180°$, где ABC – треугольник; $\angle A$, $\angle B$ и $\angle C$ – углы треугольника.	$\angle A +\angle B + \angle C = 180^{\circ}$, где $\angle A$, $\angle B$ и $\angle C$ – углы треугольника.	$\forall\ ABC_{type:Triangle}, A_{type:Angle}, B_{type:Angle}, C_{type:Angle}.$ isAngleOf (A, ABC) ∧ isAngleOf (B, ABC) ∧ isAngleOf (C, ABC) ∧ A ≠ B ∧ B ≠ C ∧ A ≠ C → degreeMesure (A) + degreeMesure (B) + degreeMesure (C) = 180.

Рис. 12. Пример записи из параллельного формального/неформального корпуса образовательных математических текстов.

В качестве исходных математических текстов были использованы учебные материалы по курсу планиметрии средней школы. Представления утверждений в форматах LaTeX и OpenMath были построены вручную. Сформированный корпус математических утверждений может быть использован в качестве тестовой коллекции при разработке методов автоматической формализации математических документов на естественном языке. Кроме того, на базе этого корпуса работает сервис для автоматической генерации тестовых вопросов, описанный ниже.

Сервис автоматической генерации тестовых вопросов для проверки математических знаний. Этот сервис базируется на параллельном формальном-неформальном корпусе математических учебных материалов [57]. Для генерации тестового вопроса инструмент извлекает из корпуса OpenMath-утверждение, выражающее функциональную зависимость между переменными, и заменяет независимые переменные сгенерированными значениями. Область определения переменных определяется на основе символов, представленных в онтологии OntoMath^{Edu}. Задача обучающегося при прохождении теста состоит в том, чтобы на основе предъявленных значений независимых переменных найти значение зависимой. Сервис автоматически проверяет правильность ответа путем выполнения OpenMath-выражения.

Фабрика метаданных цифровой библиотеки. Управление контентом цифровых библиотек основано на использовании метаданных документов (см., например, [82]). Программные инструменты, обеспечивающие основные операции с метаданными, объединяются в систему, которую называют фабрикой метаданных по аналогии с привычным термином из промышленного производства, основанного на применении машин (например, [58, 83]).

В цифровой математической библиотеке Lobachevskii-DML разрабатывается фабрика метаданных, включающая программные инструменты извлечения метаданных из документов с помощью анализа их структурных и стилевых особенностей, а также с использованием методов обработки естественного языка [59]. Созданы сервисы уточнения метаданных, а также, сервисы пополнения обязательного набора метаданных в случаях отсутствия необходимой информации. Для реализации сервисов пополнения метаданных создана система SPARQL-запросов к Wikidata и другим открытым внешним сетевым ресурсам (см., например, [84–89]). Представление метаданных документов в цифровой библиотеке Lobachevskii-DML

основано на схемах NISO JATS (например, [90]). Разработаны сервисы нормализации метаданных, обеспечивающие их преобразование по xml-схемам агрегирующих научных библиотек (см. [60, 91–93]).

ЗАКЛЮЧЕНИЕ

Представлены разработанный подход и новые направления развития цифровой экосистемы OntoMath, которая является технологической основой цифровой математической библиотеки Lobachevskii-DML. Сервисы, входящие в состав этой экосистемы, основаны на использовании онтологий OntoMath^{PRO} и OntoMath^{Edu}, что позволяет учитывать семантику предметных областей. Разработанные сервисы применены в практике работы редколлегий математических журналов, издаваемых в Казанском федеральном университете (КФУ). На основе образовательной онтологии OntoMath^{Edu} в КФУ созданы также учебные курсы в системе дистанционного обучения.

Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 21-11-00105).

СПИСОК ЛИТЕРАТУРЫ

1. *Bartling S., Friesike S.* Towards Another Scientific Revolution // S. Bartling and S. Friesike (Eds.) Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer, Cham, 2014. P. 3–15. https://doi.org/10.1007/978-3-319-00026-8_1.

2. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Управление жизненным циклом электронных публикаций в информационной системе научного журнала // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2014. № 4. С. 81–88.

3. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Сервисы поддержки жизненного цикла электронных научных публикаций // Научный сервис в сети Интернет: многообразие суперкомпьютерных миров. Труды Международной суперкомпьютерной конференции. Российская академия наук. Суперкомпьютерный консорциум университетов России. 2014. С. 436–438.

4. *Heller L., The R., Bartling S.* Dynamic Publication Formats and Collaborative

Authoring // S. Bartling and S. Friesike (Eds.) Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer, Cham, 2014. P. 191–211. https://doi.org/10.1007/978-3-319-00026-8_13.

5. *Горбунов-Посадов М.* Живая публикация // Открытые системы. СУБД. 2011. № 4. С. 48.

6. *Елизаров А.М., Липачёв Е.К.* Цифровые платформы и цифровые научные библиотеки // International Journal of Open Information Technologies. 2020. Vol. 8. No. 11. P. 80–90.

7. *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 (4). P. 347–353. https://doi.org/10.1134/S1995080214040143.

8. *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), Vladivostok, Russia, 25–29 September, 2017. IEEE, 2017. P. 39–43. https://doi.org/10.1109/RPC.2017.8168063.

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

10. SearchOnMath Site. URL: https://www.searchonmath.com/.

11. MathWebSearch: Searching Math on the Web. URL: https://search.mathweb.org/.

12. The zbMATH Open formula search. URL: https://zbmath.org/formulae/.

13. *Berčič K, Carette J., Farmer W.M., Kohlhase M., Dennis Müller D., Rabe F., Sharoda Y.* The Space of Mathematical Software Systems – A Survey of Paradigmatic Systems // arXiv: 2002.04955v1 [cs.MS] 12 Feb 2020.

14. *Kohlhase M., Sucan I.* A Search Engine for Mathematical Formulae // J. Calmet et al. (Eds.). Proceedings of the 8th International Conference on Artificial Intelligence and Symbolic Computation (AISC 2006), Beijing, China, September 20–22, 2006. Lecture Notes in Computer Science, Vol. 4120. Springer, Berlin, Heidelberg, 2006. P. 241–253. https://doi.org/10.1007/11856290_21.

15. *Guidi F., Sacerdoti Coen C.* A Survey on Retrieval of Mathematical Knowledge // Math. Comput. Sci. 2016. Vol. 10. P. 409–427.

16. *Pechnikov A., Chebukov D., Nwohiri A.* Communication of Scientists Through Scientific Publications: Math-Net.Ru as a Case Study // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-2020), Novorossiysk–Abrau, Russia, September 21–25, 2020. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2784. P. 234–244. https://ceur-ws.org/Vol-2784/rpaper19.pdf.

17. *Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К.* Онтология математического знания и рекомендательная система для коллекций физико-математических документов // Доклады Академии наук. 2016. Т. 467 (4). P. 392–395 (2016). https://doi.org/10.7868/S0869565216100042.

18. *Kozicyn A.S., Afonin S.A., Shachnev D.A.* The Use of Thematic Analysis Methods in Scientometric Systems // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-2020), Novorossiysk–Abrau, Russia, September 21–25, 2020. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2784. P. 178–188. https://ceur-ws.org/Vol-2784/rpaper14.pdf.

19. *Abecker A., van Elst L.* Ontologies for Knowledge Management // S. Staab and R. Studer (Eds.) Handbook on Ontologies. Springer, Berlin, Heidelberg, 2009. P. 713–734. https://doi.org/10.1007/978-3-540-92673-3_32.

20. *Lange Ch.* Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web // Semantic Web Journal. 2013. Vol. 4 (2). P. 119–158. https://doi.org/10.3233/SW-2012-0059.

21. *Hazewinkel M.* Mathematical Knowledge Management: Mathematical knowledge management is needed // arXiv:cs/0410055 [cs.IR] Oct 2004.

22. *Carette J., Farmer W.M.* A Review of Mathematical Knowledge Management // J. Carette et al. (Eds.). Proceedings of the International Conference on Intelligent Computer Mathematics (CICM 2009), Grand Bend, Canada, July 6-12, 2009. Lecture Notes in Computer Science. Springer, 2009. Vol. 5625. P. 233–246. https://doi.org/10.1007/978-3-642-02614-0_21.

23. *Elizarov A.M., Lipachev E.K., Zuev D.S.* Digital Mathematical Libraries: Overview of implementations and content management services // L. Kalinichenko et al. (Eds.) Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, Russia, October 9–

13, 2017. CEUR Workshop Proceedings. CEUR-WS, 2017. Vol. 2022. P. 317–325. https://ceur-ws.org/Vol-2022/paper49.pdf.

24. *Borwein J., Rocha E.M., Rodrigues J.F.* Communicating Mathematics in the Digital Era. A K Peters/CRC Press, 2008.

25. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the One-Brain Barrier. A Position Paper and Architecture Proposal // arXiv:1904.10405v1 [cs.MS] 23 Apr 2019.

26. *Carette J., Farmer W.M., Kohlhase M. Rabe F.* Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge. Math Intelligencer. 2021. Vol. 43. P. 78–87. https://doi.org/10.1007/s00283-020-10006-0.

27. *Tansley A.G.* The Use and Abuse of Vegetational Concepts and Terms // Ecology. 1935. Vol. 16 (3). P. 284–307. https://doi.org/10.2307/1930070. https://www.jstor.org/stable/1930070.

28. *Briscoe G., De Wilde P.* Digital ecosystems: self-organisation of evolving agent populations // Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES '09), Lyon, France, October 27–30, 2009. ACM, 2009. P. 44–48. https://doi.org/10.1145/1643823.1643832.

29. *Kurz T., Eder R., and Heistracher T.* Knowledge Resources – A Knowledge Management Approach for Digital Ecosystems // F.A. Basile Colugnati et al. (Eds.) Revised Selected Papers of the 3rd International Conference on Digital Eco-Systems (OPAALS 2010), Aracujú, Sergipe, Brazil, March 22–23, 2010. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, Berlin, Heidelberg, 2010. Vol. 67. P. 131–145. https://doi.org/10.1007/978-3-642-14859-0_11.

30. *Bosch J.* Speed, Data, and Ecosystems. Excelling in a Software-Driven World. CRC Press. Taylor & Francis Group, 2017.

31. *Szoniecky S., Bouhai N. (Eds.)* Collective Intelligence and Digital Archives: Towards Knowledge Ecosystems. ISTE Ltd and John Wiley & Sons, Inc., 2017.

32. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // L. Kalinichenko et al. (Eds.) Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, Russia, October 9–13, 2017. CEUR Workshop Proceedings. CEUR-WS, 2017. Vol. 2022. P. 326–333.

https://ceur-ws.org/Vol-2022/paper50.pdf.

33. *Elizarov A., Lipachev E.* Big Math Methods in Lobachevskii-DML Digital Library // A. Elizarov et al. (Eds.) Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019), Kazan, Russia, October 15–18, 2019. CEUR-WS, 2019. Vol. 2523. P. 59–72. https://ceur-ws.org/Vol-2523/invited08.pdf.

34. *Елизаров А.М., Кириллович А.В., Липачёв Е.К., Невзорова О.А.* Управление математическими знаниями: онтологические модели и цифровые технологии // Сборник статей XVIII международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/RCDL'2016). М.: ФИЦ ИУ РАН, 2016. С. 95–101.

35. *Elizarov A.M., Kirilovich A.V., Lipachev E.K., Nevzorova O.A.* Mathematical Knowledge Management: Ontological Models and Digital Technology // L. Kalinichenko, et al. (Eds.) Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow, Russia, October 11–14, 2016. CEUR Workshop Proceedings. CEUR-WS, 2016. Vol. 1752. P. 44–50. https://ceur-ws.org/Vol-1752/paper08.pdf.

36. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O.* Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management // L. Kalinichenko, S. Kuznetsov, and Y. Manolopoulos (Eds.) Revised Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow, Russia, October 11–14, 2016. Communications in Computer and Information Science. Springer, 2017. Vol. 706. P. 33–46. https://doi.org/10.1007/978-3-319-57135-5_3.

37. *Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., and Solovyev V.D.* The OntoMath ecosystem: Ontologies and applications for math knowledge management // Semantic Representation of Mathematical Knowledge Workshop, Fields Institute, Toronto, Canada, February 5, 2016. URL: https://video-archive.fields.utoronto.ca/view/4698.

38. *d'Aquin M., Motta E.* Visualizing consensus with online ontologies to support quality in ontology development // EKAW 2010 Workshop on Ontology Quality, 15 Oct 2010, Lisbon, Portugal, 2010.

URL: https://www.researchgate.net/publication/267562537_Visualizing_Consensus_with_Online_Ontologies_to_Support_Quality_in_Ontology_Development.

39. *Groza T., Handschuh S., Möller K., Decker S.* SALT – Semantically Annotated LaTeX for Scientific Publications // E. Franconi et al. (Eds.). Proceedings of the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, June 3–7, 2007. Lecture Notes in Computer Science. Springer, 2007. Vol 4519. P. 518–532. https://doi.org/10.1007/978-3-540-72667-8_37.

40. *Groza T., Handschuh S.* SALT Document Ontology. DERI, 2009. URL: https://web.archive.org/web/20100516153736/http://salt.semanticauthoring.org/ontologies/sdo.

41. *Невзорова О.А, Биряльцев Е.В., Жильцов Н.Г.* Коллекции математических текстов: аннотирование и применение в поисковых задачах // Искусственный интеллект и принятие решений. 2012. № 3. С. 51–62.

42. *Solovyev V., Zhiltsov N.* Logical structure analysis of scientific publications in mathematics // R. Akerkar (Ed.). Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS 2011), Sogndal, Norway May 25–27, 2011. ACM, 2011. Article No. 21. https://doi.org/10.1145/1988688.1988713.

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

44. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMath^{PRO} Ontology: A Linked Data Hub for Mathematics // P. Klinov and D. Mouromstev (Eds.). Proceedings of the 5th International Conference on Knowledge Engineering and Semantic Web (KESW 2014), Kazan, Russia, September 29–October 1, 2014. Communications in Computer and Information Science. Springer, Cham, 2014. Vol. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4_9.

45. *Kirillovich A., Nevzorova O., Falileeva M., Lipachev E., and Shakirova L.* OntoMath^{Edu}: A Linguistically Grounded Educational Mathematical Ontology // C. Benzmüller and B. Miller (Eds.). Proceedings of the 13th International Conference on Intelligent Computer Mathematics (CICM 2020), Bertinoro, Italy, July 26–31, 2020. Lecture Notes in Computer Science. Springer, 2020. Vol. 12236. P. 157–172. https://doi.org/10.1007/978-3-030-53518-6_10.

46. *Kirillovich A., Nevzorova O., Falileeva M., Lipachev E., and Shakirova L*. OntoMath^{Edu}: Towards an Educational Mathematical Ontology // E. Brady et al. (Eds.). Workshop Papers at 12th Conference on Intelligent Computer Mathematics (CICM-WS 2019), Prague, Czech Republic, 8–12 July 2019. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2634. URL: https://ceur-ws.org/Vol-2634/WiP1.pdf.

47. *Kirillovich A., Nevzorova O., Falileeva M., Lipachev E., Dyupina A., Shakirova L.* Prerequisite Relationships of the OntoMath^{Edu} Educational Mathematical Ontology // J.C. Figueroa-García et al. (Eds.). Proceedings of the 8th Workshop on Engineering Applications (WEA 2021), Medellín, Colombia, October 6–8, 2021. Communications in Computer and Information Science. Springer, 2021. Vol. 1431. P. 517–524. https://doi.org/10.1007/978-3-030-86702-7_44.

48. *Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.* Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics // Harith Alani et al. (Eds.). Proceedings of the 12th International Semantic Web Conference (ISWC 2013), Sydney, NSW, Australia, October 21–25, 2013. Lecture Notes in Computer Science. Springer, 2013. Vol. 8218. P. 379–394. https://doi.org/10.1007/978-3-642-41335-3_24.

49. *Nevzorova O., Almukhametov D.* Towards a Recommender System for the Choice of UDC Code for Mathematical Articles // A. Pozanenko et al. (Eds.). Supplementary Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021), Moscow, Russia, October 26–29, 2021. CEUR Workshop Proceedings. CEUR-WS, 2021. Vol. 3036. P. 54–62. URL: https://ceur-ws.org/Vol-3036/paper04.pdf.

50. *Глухов В.А.*, *Елизаров А.М., Липачёв Е.К., Малахальцев М.А.* Электронные научные издания: переход на технологии семантического веба // Электронные библиотеки. 2007. Т. 10. № 1. С. 2.

51. *Елизаров А.М., Липачёв Е.К., Малахальцев М.А.* Веб-технологии в работе электронного математического журнала Lobachevskii Journal of Mathematics // Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ. Труды Всероссийской научной конференции. Московский государственный университет им. М.В. Ломоносова, Южный федеральный университет, Институт вычислительной математики РАН. 2007. С. 355–356.

52. *Ахметов Д.Ю., Елизаров А.М., Липачев Е.К.* Автоматизация редакционных процессов в информационной системе управления электронными научными журналами // Электронные библиотеки. 2015. Т. 18. № 1–2. С. 32–45.

53. *Elizarov A.M., Khaydarov S.M., Lipachev E.K.* The Formation Method of Recommendations in the Process of Scientific Peer Review of Mathematical Papers // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019), Novorossiysk–brau, Russia, September 23–28, 2019. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2543. P. 126–135.
URL: https://ceur-ws.org/Vol-2543/rpaper12.pdf.

54. *Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Рекомендательная система поиска экспертов для проведения научного рецензирования в математическом журнале // Электронные библиотеки. 2020. Т. 23. № 4. С. 708–732.
https://doi.org/ 10.26907/1562-5419-2020-23-4-708-732.

55. *Николаев К.С., Невзорова О.А.* Метод автоматической семантической разметки математических образовательных текстов // Информационные технологии в образовании и науке (ИТОН–2022) и II International Workshop “Digital Technologies for Teaching and Learning” (DTTL). Материалы III Международного форума по математическому образованию: Международной научно-практической конференции и II Международного научного семинара. Казань, 2022. С. 181–190.

56. *Kirillovich A., Nevzorova O., Nikolaev K., and Galiaskarova K.* Towards a Parallel Informal-Formal Corpus of Educational Mathematical Texts in Russian // Zhengbing Hu et al. (Eds.). Proceedings of the 2019 International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS 2019), Moscow, Russia, on 4–6 October 2019. Advances in Intelligent Systems and Computing. Springer, 2020. Vol. 1127. P. 325–334. https://doi.org/10.1007/978-3-030-39216-1_29.

57. *Nikolaev K., Kirillovich A., and Nevzorova O.*A Corpus-Based Approach to Elementary Geometry Knowledge Test Generation // L. Gómez Chova et al. (Eds.). Proceedings of the 14th International Technology, Education and Development Conference (INTED 2020), Valencia, Spain, 2–4 March 2020. IATED, 2020. P. 6342–6348.

58. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // R.V. Bolgov et al. (Eds.). Proceedings of the International Conference on Internet and Modern Society (IMS-2020), St. Petersburg, Russia, 17–20 June 2020. CEUR Workshop Proceedings. CEUR-WS, 2021. Vol. 2813. P. 13–21. URL: https://ceur-ws.org/Vol-2813/rpaper01.pdf.

59. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Базовые сервисы фабрики метаданных цифровой математической библиотеки Lobachevskii-DML // Электронные библиотеки. 2020. Т. 23. № 3. С. 336–381.
https://doi.org/ 10.26907/1562-5419-2020-23-3-336-381.

60. *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.

61. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Алгоритмы формирования метаданных математических ретро-коллекций на основе анализа структурных особенностей документов // Электронные библиотеки. 2021. Т. 24. № 2. С. 238–271. https://doi.org/10.26907/1562-5419-2021-24-2-238-270.

62. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Automated System of Services for Processing of Large Collections of Scientific Documents // L. Kalinichenko, S. Kuznetsov, and Y. Manolopoulos (Eds.). Revised Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow, Russia, October 11–14, 2016. Communications in Computer and Information Science. Springer, 2017. Vol. 706. P. 58–64.
URL: https://ceur-ws.org/Vol-1752/paper10.pdf.

63. *Elizarov A., Khaydarov S., Lipachev E.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017), Vladivostok, Russia, 25–29 September, 2017. IEEE, 2017. P. 1–5.
https://doi.org/ 10.1109/RPC.2017.8168064.

64. *Elizarov A.M., Lipachev E.K.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019), Novorossiysk–Abrau, Russia, September 23-28, 2019. CEUR Workshop Proceedings. SSI 2019 – Proceedings of the 21st Conference on Scientific Services and Internet. CEUR-WS, 2020. Vol. 2543. P. 354–360.
URL: https://ceur-ws.org/Vol-2543/spaper05.pdf.

65. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* OntoMath^{PRO}: An Ontology of Mathematical Knowledge // Doklady Mathematics. 2022. Vol. 106 (3).

P. 29–35. https://doi.org/10.1134/S1064562422700016.

66. *Guizzardi G., Botti Benevides A., Fonseca C.M., Porello D., Almeida J.P.A., Prince Sales T*. UFO: Unified Foundational Ontology // Applied Ontology. 2022. 17 (1), P. 167–210. https://doi.org/10.3233/AO-210256.

67. *Фалилеева М.В., Кириллович А.В., Невзорова О.А., Шакирова Л.Р., Липачёв Е.К., Дюпина А.Э.* Системы образовательных проекций, уровней и пререквизитов математической онтологии OntoMath^{Edu} // Электронные библиотеки. 2021. Т. 24. № 3. С. 505–530. https://doi.org/10.26907/1562-5419-2021-24-3-505-530.

68. *Муромцев Д.И.* Модели и методы индивидуализации электронного обучения в контексте онтологического подхода // Онтологии проектирования. 2020. Т. 10, № 1. С. 34–49. https://doi.org/10.18287/2223-9537-2020-10-1-34-49.

69. *Schraefel M., Shadbolt N., Gibbins N.* CS AKTive Space: Representing Computer Science on the Semantic Web // Proceedings of the 13th international conference on World Wide Web (WWW 2004), New York, USA, May 17–20, 2004. N.Y.: ACM Press New York, 2004. P. 384–392. https://doi.org/10.1145/988672.988724.

70. *Kirillovich A. and Nikolaev K.* Adapting the LodView RDF Browser for Navigation over the Multilingual Linguistic Linked Open Data Cloud // Proceedings of the 9th IEEE International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT 2022), Genova, Italy & Sfax, Tunisia, 28–30 May 2022. IEEE, 2022. P. 143–149. https://doi.org/10.1109/SETIT54465.2022.9875628.

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

72. *Khaydarov S.M., Yamalutdinova G.S.* Recommender system of physical and mathematical documents classification // V. Voevodin et al. (Eds.). Proceedings of the 20th Conference Scientific Services & Internet (SSI-2018), Novorossiysk–Abrau, Russia, September 17–22, 2018. CEUR Workshop Proceedings. CEUR-WS, 2018. Vol. 2260. P. 480–486. URL: https://ceur-ws.org/Vol-2260/57_480-486.pdf.

73. *Хайдаров Ш.М., Ямалутдинова Г.Ш.* Рекомендательная система классификации физико-математических документов // Труды XX Всероссийской научной конференции «Научный сервис в сети Интернет», 17–22 сентября 2018, г. Новороссийск. М.: ИПМ им. М.В. Келдыша, 2018. С. 480–486.

74. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2014. № 4. С. 12–17.

75. *Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V.* Methods for Analyzing Semantic Data of Electronic Collections in Mathematics // Automatic Documentation and Mathematical Linguistics. 2014. V. 48. No. 2. P. 81–85.

76. *Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К., Хайдаров Ш.М.* Программный комплекс формирования рекомендаций по подбору рецензентов для научных документов в информационных издательских системах // Свидетельство о регистрации программы для ЭВМ RU 2018611617, 02.02.2018.

77. *MacGregor J., Stranack K. and Willinsky J.* The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication // S. Bartling and S. Friesike (Eds.) Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer, Cham, 2014. P. 165–175. https://doi.org/ 10.1007/978-3-319-00026-8_11.

78. Mathematics Subject Classification (MSC2010). https://mathscinet.ams.org/mathscinet/msc/pdfs/classifications2010.pdf.

79. MSC2020-Mathematics Subject Classification System. https://mathscinet.ams.org/msnhtml/msc2020.pdf.

80. MSC Conversion Table. https://mathscinet.ams.org/mathscinet/msc/conv.html?from=2010.

81. *Buswell S. et al.* (Eds.) The OpenMath Standard. Version: 2.0r2. The OpenMath Society, July 2019.
URL: https://openmath.org/standard/om20-2019-07-01/omstd20.html.

82. *Xie I, Matusiak K.* Discover Digital Libraries: Theory and Practice. Elsevier, 2016.

83. *Bouche T., Labbe O.* The New Numdam Platform // H. Geuvers et al. (Eds.). Proceedings of the 10th International Conference on Intelligent Computer Mathematics (CICM 2017), Edinburgh, UK, July 17–21, 2017. Lecture Notes in Computer Science. Springer, Cham, 2017. Vol. 10383. P. 70–82.
https://doi.org/10.1007/978-3-319-62075-6_6.

URL: https://zenodo.org/record/581405.

84. *Erxleben F., Günther M., Krötzsch M., Mendez J., Vrandečić D*. Introducing Wikidata to the Linked Data Web // P. Mika et al. (Eds.). Proceedings of the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19–23, 2014, Part I. Lecture Notes in Computer Science. Springer, Cham, 2014. Vol. 8796. P. 50–65. https://doi.org/10.1007/978-3-319-11964-9_4.

85. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase. Communications of the ACM. 2014. Vol. 57, Issue 10, October 2014. P. 78–85. https://doi.org/10.1145/2629489.

86. *Scharpf Ph., Schubotz M., Gipp B.* Mathematics in Wikidata // L.-A. Kaffee et al. (Eds.). Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), October 24, 2021. CEUR Workshop Proceedings. CEUR-WS, 2021. Vol. 2982.
URL: http://ceur-ws.org/Vol-2982/paper-1.pdf.

87. *Андреичев М.Д., Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Пополнение метаданных документов математических цифровых ретро-коллекций методом семантических сетей // Труды XXIII Всероссийской научной конференции «Научный сервис в сети Интернет». М.: ИПМ им. М.В. Келдыша, 2021. С. 22–33. https://doi.org/10.20948/abrau-2021-22.

88. *Harris S. et al. (Eds.)* SPARQL 1.1 Query Language. W3C Recommendation, 21 March 2013. URL: https://www.w3.org/TR/sparql11-query/.

89. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Извлечение знаний из Wikidata для формирования метаданных документов электронных математических коллекций // Электронные библиотеки. 2021. Т. 24. № 6. С. 1023–1059. https://doi.org/10.26907/1562-5419-2021-24-6-1023-1059.

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

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

92. *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.

93. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Metadata Normalization Methods in the Digital Mathematical Library // M. Gorbunov-Posadov et

al. (Eds.). Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019), Novorossiysk–Abrau, Russia, September 23–28, 2019. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2543. P. 136–148.
URL: https://ceur-ws.org/Vol-2543/rpaper13.pdf.

DIGITAL ECOSYSTEM OntoMath AS AN APPROACH TO BUILDING THE SPACE OF MATHEMATICAL KNOWLEDGE

A. M. Elizarov[1] [0000-0003-2546-6897], A. V. Kirillovich [2] [0000-0001-9680-449X], E. K. Lipachev[3] [0000-0001-7789-2332], O. A. Nevzorova[4] [0000-0001-8116-9446]

[1–4] *Kazan Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[1, 2] *Joint Supercomputer Center of the Russian Academy of Sciences, Kazan, Russia*

[1]amelizarov@gmail.com, [2]alik.kirillovich@gmail.com, [3]elipachev@gmail.com, [4]onevzoro@gmail.com

Abstract

The results on the creation of methods for managing mathematical knowledge in the context of digital mathematical libraries are presented. The software tools developed on the basis of these methods are part of the OntoMath digital ecosystem, within which they interact. A brief description of the architecture of the OntoMath ecosystem is given, the levels of subject ontologies and external ontologies are highlighted, as well as the level of software tools and services. Semantic services are separated into a separate category. This term denotes software tools, in the functionality of which queries to subject ontologies are used to ensure the management of knowledge objects. General descriptions of developed subject ontologies are given: educational mathematical ontology OntoMath^{Edu} and ontology of professional mathematics OntoMath^{PRO}. The development of educational ontology is reflected in the direction of including educational prerequisite links between classes. Among the software tools of the digital ecosystem, search services for mathematical electronic collections, a service for semantic annotation of mathematical documents, tools for semantic marking of educational mathematical documents, as well as a system for automatically generating testing tests in mathematical educational disciplines are

highlighted. As part of the OntoMath digital ecosystem, special-purpose recommender systems are being developed. The current version of the ecosystem includes a recommender system for generating a list of related articles based on the OntoMath^{PRO} ontology, a recommender system for appointing experts to support the scientific review process, and recommender systems for selecting subject classifiers UDC and Mathematics Subject Classification codes for mathematical documents. The results are also presented in the direction of creating a digital library metadata factory, which includes services and tools for extracting, refining, replenishing and normalizing the metadata of electronic mathematical collections. Note that the OntoMath ecosystem is being developed as the technological basis for the Lobachevskii Digital Mathematical Library.

Keywords*: Digital Ecosystem, OntoMath Ecosystem, Digital Mathematical Library, Lobachevskii-DML, Ontology, OntoMath^{PRO}, OntoMath^{Edu}.*

REFERENCES

1. *Bartling S., Friesike S.* Towards Another Scientific Revolution // S. Bartling and S. Friesike (Eds.) Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer, Cham, 2014. P. 3–15. https://doi.org/10.1007/978-3-319-00026-8_1.

2. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Lifecycle management of Electronic Publications in Information Systems Scientific Journal // Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyj analiz i informacionnye tekhnologii. 2014. № 4. S. 81–88.

3. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Servisy podderzhki zhiznennogo cikla elektronnyh nauchnyh publikacij // Nauchnyj servis v seti Internet: mnogoobrazie superkomp'yuternyh mirov. Trudy Mezhdunarodnoj superkomp'yuternoj konferencii. Rossijskaya akademiya nauk Superkomp'yuternyj konsorcium universitetov Rossii. 2014. S. 436–438.

4. *Heller L., The R., Bartling S.* Dynamic Publication Formats and Collaborative Authoring // S. Bartling and S. Friesike (Eds.) Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer, Cham, 2014. P. 191–211. https://doi.org/10.1007/978-3-319-00026-8_13.

5. *Gorbunov-Posadov M.* Zhivaya publikaciya // Otkrytye sistemy. SUBD. 2011.

№ 4. S. 48.

6. *Elizarov A.M., Lipachev E.K.* Digital Platforms and Digital Science Libraries // International Journal of Open Information Technologies. 2020. Vol. 8. No. 11. P. 80–90.

7. *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 (4). P. 347–353. https://doi.org/10.1134/S1995080214040143.

8. *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), Vladivostok, Russia, 25–29 September, 2017. IEEE, 2017. P. 39–43. https://doi.org/10.1109/RPC.2017.8168063.

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

10. SearchOnMath Site. URL: https://www.searchonmath.com/.

11. MathWebSearch: Searching Math on the Web. URL: https://search.mathweb.org/.

12. The zbMATH Open formula search. URL: https://zbmath.org/formulae/.

13. *Berčič K, Carette J., Farmer W.M., Kohlhase M., Dennis Müller D., Rabe F., Sharoda Y.* The Space of Mathematical Software Systems – A Survey of Paradigmatic Systems // arXiv: 2002.04955v1 [cs.MS] 12 Feb 2020.

14. *Kohlhase M., Sucan I.* A Search Engine for Mathematical Formulae // J. Calmet et al. (Eds.). Proceedings of the 8th International Conference on Artificial Intelligence and Symbolic Computation (AISC 2006), Beijing, China, September 20–22, 2006. Lecture Notes in Computer Science, Vol. 4120. Springer, Berlin, Heidelberg, 2006. P. 241–253. https://doi.org/10.1007/11856290_21.

15. *Guidi F., Sacerdoti Coen C.* A Survey on Retrieval of Mathematical Knowledge // Math. Comput. Sci. 2016. Vol. 10. P. 409–427.

16. *Pechnikov A., Chebukov D., Nwohiri A.* Communication of Scientists Through Scientific Publications: Math-Net.Ru as a Case Study // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-

2020), Novorossiysk–Abrau, Russia, September 21–25, 2020. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2784. P. 234–244.
https://ceur-ws.org/Vol-2784/rpaper19.pdf.

17. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Zhizhchenko A.B., and Zhil'tsov N.G.* Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics // Doklady Mathematics. 2016. Vol. 93 (2). P. 231–233. https://doi.org/ 10.1134/S1064562416020174.

18. *Kozicyn A.S., Afonin S.A., Shachnev D.A.* The Use of Thematic Analysis Methods in Scientometric Systems // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-2020), Novorossiysk–Abrau, Russia, September 21–25, 2020. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2784. P. 178–188. https://ceur-ws.org/Vol-2784/rpaper14.pdf.

19. *Abecker A., van Elst L.* Ontologies for Knowledge Management // S. Staab and R. Studer (Eds.) Handbook on Ontologies. Springer, Berlin, Heidelberg, 2009. P. 713–734. https://doi.org/10.1007/978-3-540-92673-3_32.

20. *Lange Ch.* Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web // Semantic Web Journal. 2013. Vol. 4 (2). P. 119–158. https://doi.org/10.3233/SW-2012-0059.

21. *Hazewinkel M.* Mathematical Knowledge Management: Mathematical knowledge management is needed // arXiv:cs/0410055 [cs.IR] Oct 2004.

22. *Carette J., Farmer W.M.* A Review of Mathematical Knowledge Management // J. Carette et al. (Eds.). Proceedings of the International Conference on Intelligent Computer Mathematics (CICM 2009), Grand Bend, Canada, July 6-12, 2009. Lecture Notes in Computer Science. Springer, 2009. Vol. 5625. P. 233–246.
https://doi.org/10.1007/978-3-642-02614-0_21.

23. *Elizarov A.M., Lipachev E.K., Zuev D.S.* Digital Mathematical Libraries: Overview of implementations and content management services // L. Kalinichenko et al. (Eds.) Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, Russia, October 9–13, 2017. CEUR Workshop Proceedings. CEUR-WS, 2017. Vol. 2022. P. 317–325.
https://ceur-ws.org/Vol-2022/paper49.pdf.

24. *Borwein J., Rocha E.M., Rodrigues J.F.* Communicating Mathematics in the Digital Era. A K Peters/CRC Press, 2008.

25. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the One-Brain Barrier. A Position Paper and Architecture Proposal // arXiv:1904.10405v1 [cs.MS] 23 Apr 2019.

26. *Carette J., Farmer W.M., Kohlhase M. Rabe F.* Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge. Math Intelligencer. 2021. Vol. 43. P. 78–87. https://doi.org/10.1007/s00283-020-10006-0.

27. *Tansley A.G.* The Use and Abuse of Vegetational Concepts and Terms // Ecology. 1935. Vol. 16 (3). P. 284–307. https://doi.org/10.2307/1930070. https://www.jstor.org/stable/1930070.

28. *Briscoe G., De Wilde P.* Digital ecosystems: self-organisation of evolving agent populations // Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES '09), Lyon, France, October 27–30, 2009. ACM, 2009. P. 44–48. https://doi.org/10.1145/1643823.1643832.

29. *Kurz T., Eder R., and Heistracher T.* Knowledge Resources – A Knowledge Management Approach for Digital Ecosystems // F.A. Basile Colugnati et al. (Eds.) Revised Selected Papers of the 3rd International Conference on Digital Eco-Systems (OPAALS 2010), Aracujú, Sergipe, Brazil, March 22–23, 2010. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, Berlin, Heidelberg, 2010. Vol. 67. P. 131–145. https://doi.org/10.1007/978-3-642-14859-0_11.

30. *Bosch J.* Speed, Data, and Ecosystems. Excelling in a Software-Driven World. CRC Press. Taylor & Francis Group, 2017.

31. *Szoniecky S., Bouhai N. (Eds.)* Collective Intelligence and Digital Archives: Towards Knowledge Ecosystems. ISTE Ltd and John Wiley & Sons, Inc., 2017.

32. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // L. Kalinichenko et al. (Eds.) Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, Russia, October 9–13, 2017. CEUR Workshop Proceedings. CEUR-WS, 2017. Vol. 2022. P. 326–333. https://ceur-ws.org/Vol-2022/paper50.pdf.

33. *Elizarov A., Lipachev E.* Big Math Methods in Lobachevskii-DML Digital Library // A. Elizarov et al. (Eds.) Selected Papers of the XXI International Conference on

Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019), Kazan, Russia, October 15–18, 2019. CEUR-WS, 2019. Vol. 2523. P. 59–72. https://ceur-ws.org/Vol-2523/invited08.pdf.

34. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Upravlenie matematicheskimi znaniyami: ontologicheskie modeli i cifrovye tekhnologii // Analitika i upravlenie dannymi v oblastyah s intensivnym ispol'zovaniem dannyh. XVIII mezhdunarodnaya konferenciya. 2016. S. 95–101.

35. *Elizarov A.M., Kirilovich A.V., Lipachev E.K., Nevzorova O.A.* Mathematical Knowledge Management: Ontological Models and Digital Technology // L. Kalinichenko, et al. (Eds.) Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow, Russia, October 11–14, 2016. CEUR Workshop Proceedings. CEUR-WS, 2016. Vol. 1752. P. 44–50. https://ceur-ws.org/Vol-1752/paper08.pdf.

36. *Elizarov A., Kirillovich A., Lipachev E., Nevzorova O.* Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management // L. Kalinichenko, S. Kuznetsov, and Y. Manolopoulos (Eds.) Revised Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow, Russia, October 11–14, 2016. Communications in Computer and Information Science. Springer, 2017. Vol. 706. P. 33–46. https://doi.org/10.1007/978-3-319-57135-5_3.

37. *Elizarov A.M., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., and Solovyev V.D.* The OntoMath ecosystem: Ontologies and applications for math knowledge management // Semantic Representation of Mathematical Knowledge Workshop, Fields Institute, Toronto, Canada, February 5, 2016.
URL: https://video-archive.fields.utoronto.ca/view/4698.

38. *d'Aquin M., Motta E.* Visualizing consensus with online ontologies to support quality in ontology development // EKAW 2010 Workshop on Ontology Quality, 15 Oct 2010, Lisbon, Portugal, 2010.
URL: https://www.researchgate.net/publication/267562537_Visualizing_Consensus_with_Online_Ontologies_to_Support_Quality_in_Ontology_Development.

39. *Groza T., Handschuh S., Möller K., Decker S.* SALT – Semantically Annotated LaTeX for Scientific Publications // E. Franconi et al. (Eds.). Proceedings of the 4th Euro-

pean Semantic Web Conference (ESWC 2007), Innsbruck, Austria, June 3–7, 2007. Lecture Notes in Computer Science. Springer, 2007. Vol 4519. P. 518–532. https://doi.org/10.1007/978-3-540-72667-8_37.

40. *Groza T., Handschuh S.* SALT Document Ontology. DERI, 2009. URL: https://web.archive.org/web/20100516153736/http://salt.semanticauthoring.org/ontologies/sdo.

41. *Nevzorova O.A, Biryal'cev E.V., Zhil'cov N.G.* Kollekcii matematicheskih tekstov: annotirovanie i primenenie v poiskovyh zadachah // Iskusstvennyj intellekt i prinyatie reshenij. 2012. № 3. S. 51–62.

42. *Solovyev V., Zhiltsov N.* Logical structure analysis of scientific publications in mathematics // R. Akerkar (Ed.). Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS 2011), Sogndal, Norway May 25–27, 2011. ACM, 2011. Article No. 21. https://doi.org/10.1145/1988688.1988713.

43. *Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solovyev V.D.* Methods and means for semantic structuring of electronic mathematical documents// Doklady Mathematics. 2014. Vol. 90, No. 1. P. 521–524. https://doi.org/0.1134/S1064562414050275.

44. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMath^{PRO} Ontology: A Linked Data Hub for Mathematics // P. Klinov and D. Mouromstev (Eds.). Proceedings of the 5th International Conference on Knowledge Engineering and Semantic Web (KESW 2014), Kazan, Russia, September 29–October 1, 2014. Communications in Computer and Information Science. Springer, Cham, 2014. Vol. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4_9.

45. *Kirillovich A., Nevzorova O., Falileeva M., Lipachev E., and Shakirova L.* OntoMath^{Edu}: A Linguistically Grounded Educational Mathematical Ontology // C. Benzmüller and B. Miller (Eds.). Proceedings of the 13th International Conference on Intelligent Computer Mathematics (CICM 2020), Bertinoro, Italy, July 26–31, 2020. Lecture Notes in Computer Science. Springer, 2020. Vol. 12236. P. 157–172. https://doi.org/10.1007/978-3-030-53518-6_10.

46. *Kirillovich A., Nevzorova O., Falileeva M., Lipachev E., and Shakirova L*. OntoMath^{Edu}: Towards an Educational Mathematical Ontology // E. Brady et al. (Eds.). Workshop Papers at 12th Conference on Intelligent Computer Mathematics (CICM-WS 2019), Prague, Czech Republic, 8–12 July 2019. CEUR Workshop Proceedings. CEUR-WS,

2020. Vol. 2634. URL: https://ceur-ws.org/Vol-2634/WiP1.pdf.

47. *Kirillovich A., Nevzorova O., Falileeva M., Lipachev E., Dyupina A., Shakirova˚L.* Prerequisite Relationships of the OntoMath^{Edu} Educational Mathematical Ontology // J.C. Figueroa-García et al. (Eds.). Proceedings of the 8th Workshop on Engineering Applications (WEA 2021), Medellín, Colombia, October 6–8, 2021. Communications in Computer and Information Science. Springer, 2021. Vol. 1431. P. 517–524. https://doi.org/10.1007/978-3-030-86702-7_44.

48. *Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.* Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics // Harith Alani et al. (Eds.). Proceedings of the 12th International Semantic Web Conference (ISWC 2013), Sydney, NSW, Australia, October 21–25, 2013. Lecture Notes in Computer Science. Springer, 2013. Vol. 8218. P. 379–394. https://doi.org/10.1007/978-3-642-41335-3_24.

49. *Nevzorova O., Almukhametov D.* Towards a Recommender System for the Choice of UDC Code for Mathematical Articles // A. Pozanenko et al. (Eds.). Supplementary Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021), Moscow, Russia, October 26–29, 2021. CEUR Workshop Proceedings. CEUR-WS, 2021. Vol. 3036. P. 54–62. URL: https://ceur-ws.org/Vol-3036/paper04.pdf.

50. *Gluhov V.A., Elizarov A.M., Lipachev E.K., Malahal'cev M.A.* Elektronnye nauchnye izdaniya: perekhod na tekhnologii semanticheskogo veba // Elektronnye biblioteki. 2007. T. 10. № 1. S. 2.

51. *Elizarov A.M., Lipachev E.K., Malahal'cev M.A.* Veb-tekhnologii v rabote elektronnogo matematicheskogo zhurnala Lobachevskii Journal of Mathematics // Nauchnyj servis v seti Internet: mnogoyadernyj komp'yuternyj mir. 15 let RFFI. Trudy Vserossijskoj nauchnoj konferencii. Moskovskij gosudarstvennyj universitet im. M.V. Lomonosova, Yuzhnyj federal'nyj universitet, Institut vychislitel'noj matematiki RAN. 2007. S. 355–356.

52. *Akhmetov D.Yu., Elizarov A.M., Lipachev E.K.* Information systems of electronic scientific journals and editorial process automation // Russian Digital Libraries Journal. 2015. Vol. 18. No 1-2. P. 32-45.

53. *Elizarov A.M., Khaydarov S.M., Lipachev E.K.* The Formation Method of Recommendations in the Process of Scientific Peer Review of Mathematical Papers //

M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019), Novorossiysk–brau, Russia, September 23–28, 2019. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2543. P. 126–135.
URL: https://ceur-ws.org/Vol-2543/rpaper12.pdf.

54. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Recommender system in the process of scientific peer review in mathematical journal // Russian Digital Libraries Journal. 2020. Vol. 23. No 4. P. 708–732.
https://doi.org/ 10.26907/1562-5419-2020-23-4-708–732.

55. *Nikolaev K.S., Nevzorova O.A.* Metod avtomaticheskoj semanticheskoj razmetki matematicheskih obrazovatel'nyh tekstov // Informacionnye tekhnologii v obrazovanii i nauke (ITON – 2022) i II International Workshop "Digital Technologies for Teaching and Learning (DTTL)". Materialy III Mezhdunarodnogo foruma po matematicheskomu obrazovaniyu: Mezhdunarodnoj nauchno-prakticheskoj konferencii i II Mezhdunarodnogo nauchnogo seminara. Kazan', 2022. S. 181–190.

56. *Kirillovich A., Nevzorova O., Nikolaev K., and Galiaskarova K.* Towards a Parallel Informal-Formal Corpus of Educational Mathematical Texts in Russian // Zhengbing Hu et al. (Eds.). Proceedings of the 2019 International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS 2019), Moscow, Russia, on 4–6 October 2019. Advances in Intelligent Systems and Computing. Springer, 2020. Vol. 1127. P. 325–334. https://doi.org/10.1007/978-3-030-39216-1_29.

57. *Nikolaev K., Kirillovich A., and Nevzorova O.*A Corpus-Based Approach to Elementary Geometry Knowledge Test Generation // L. Gómez Chova et al. (Eds.). Proceedings of the 14th International Technology, Education and Development Conference (INTED 2020), Valencia, Spain, 2–4 March 2020. IATED, 2020. P. 6342–6348.

58. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // R.V. Bolgov et al. (Eds.). Proceedings of the International Conference on Internet and Modern Society (IMS-2020), St. Petersburg, Russia, 17–20 June 2020. CEUR Workshop Proceedings. CEUR-WS, 2021. Vol. 2813. P. 13–21. URL: https://ceur-ws.org/Vol-2813/rpaper01.pdf.

59. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Basic services of factory metadata digital mathematical library Lobachevskii-DML // Russian Digital Libraries Journal. 2020. Vol. 23. No 3. P. 336–381.
https://doi.org/ 10.26907/1562-5419-2020-23-3-336-381.

60. *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. Vol. 18. No 1–2. P. 6–31.

61. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Algorithms for formation of metadata mathematical retro collections based on analysis of structural features of documents // Russian Digital Libraries Journal. 2021. Vol. 24. No 2. P. 238–271. https://doi.org/10.26907/1562-5419-2021-24-2-238-270.

62. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Automated System of Services for Processing of Large Collections of Scientific Documents // L. Kalinichenko, S. Kuznetsov, and Y. Manolopoulos (Eds.). Revised Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow, Russia, October 11–14, 2016. Communications in Computer and Information Science. Springer, 2017. Vol. 706. P. 58–64. URL: https://ceur-ws.org/Vol-1752/paper10.pdf.

63. *Elizarov A., Khaydarov S., Lipachev E.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017), Vladivostok, Russia, 25–29 September, 2017. IEEE, 2017. P. 1–5. https://doi.org/ 10.1109/RPC.2017.8168064.

64. *Elizarov A.M., Lipachev E.K.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019), Novorossiysk–Abrau, Russia, September 23-28, 2019. CEUR Workshop Proceedings. SSI 2019 – Proceedings of the 21st Conference on Scientific Services and Internet. CEUR-WS, 2020. Vol. 2543. P. 354–360. URL: https://ceur-ws.org/Vol-2543/spaper05.pdf.

65. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* OntoMath^{PRO}: An Ontology of Mathematical Knowledge // Doklady Mathematics. 2022. Vol. 106 (3). P. 29–35. https://doi.org/10.1134/S1064562422700016.

66. *Guizzardi G., Botti Benevides A., Fonseca C.M., Porello D., Almeida J.P.A., Prince Sales T*. UFO: Unified Foundational Ontology // Applied Ontology. 2022. 17 (1), P. 167–210. https://doi.org/10.3233/AO-210256.

67. *Falileeva M.V., Kirillovich A.V., Nevzorova O.A., Shakirova L.R.,*

Lipachev E.K., Dyupina A.E. Educational projection systems, levels and prerequisites of mathematical ontology OntoMath^{Edu} // Russian Digital Libraries Journal. 2021. Vol. 24. No 3. P. 505–530. https://doi.org/10.26907/1562-5419-2021-24-3-505-530.

68. *Mouromtsev D.* Models and methods of e-learning individualization in the context of ontological approach // Ontology of Designing. 2020. Vol. 10, No 1. P. 34–49. https://doi.org/10.18287/2223-9537-2020-10-1-34-49.

69. *Schraefel M., Shadbolt N., Gibbins N.* CS AKTive Space: Representing Computer Science on the Semantic Web // Proceedings of the 13th international conference on World Wide Web (WWW 2004), New York, USA, May 17–20, 2004. N.Y.: ACM Press New York, 2004. P. 384–392. https://doi.org/10.1145/988672.988724.

70. *Kirillovich A. and Nikolaev K.* Adapting the LodView RDF Browser for Navigation over the Multilingual Linguistic Linked Open Data Cloud // Proceedings of the 9th IEEE International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT 2022), Genova, Italy & Sfax, Tunisia, 28–30 May 2022. IEEE, 2022. P. 143–149. https://doi.org/10.1109/SETIT54465.2022.9875628.

71. *Khaydarov S.M., Yamalutdinova G.S.* Algoritm formirovaniya slovarej rekomenduyushchej sistemy podbora klassifikatorov nauchnoj informacii // Uchenye zapiski ISGZ. 2017. T. 15. № 1. S. 552–557.

72. *Khaydarov S.M., Yamalutdinova G.S.* Recommender system of physical and mathematical documents classification // V. Voevodin et al. (Eds.). Proceedings of the 20th Conference Scientific Services & Internet (SSI-2018), Novorossiysk–Abrau, Russia, September 17–22, 2018. CEUR Workshop Proceedings. CEUR-WS, 2018. Vol. 2260. P. 480–486. URL: https://ceur-ws.org/Vol-2260/57_480-486.pdf.

73. *Khaydarov S.M., Yamalutdinova G.S* Rekomendatel'naya sistema klassifikacii fiziko-matematicheskih dokumentov // Nauchnyj servis v seti Internet. 2018. № 20. S. 480–486.

74. *Biryal'cev E.V., Elizarov A.M., Zhil'cov N.G., Lipachev E.K., Nevzorova O.A., Solov'ev V.D.* Metody analiza semanticheskih dannyh matematicheskih elektronnyh kollekcij // Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy. 2014. № 4. S. 12–17.

75. *Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V.*

Methods for Analyzing Semantic Data of Electronic Collections in Mathematics // Automatic Documentation and Mathematical Linguistics. 2014. V. 48. No. 2. P. 81–85.

76. *Ahmetov D.Yu., Elizarov A.M., Lipachev E.K., Hajdarov Sh.M.* Programmnyj kompleks formirovaniya rekomendacij po podboru recenzentov dlya nauchnyh dokumentov v informacionnyh izdatel'skih sistemah // Svidetel'stvo o registracii programmy dlya EVM RU 2018611617, 02.02.2018.

77. *MacGregor J., Stranack K. and Willinsky J.* The Public Knowledge Project: Open Source Tools for Open Access to Scholarly Communication // S. Bartling and S. Friesike (Eds.) Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer, Cham, 2014. P. 165–175. https://doi.org/ 10.1007/978-3-319-00026-8_11.

78. Mathematics Subject Classification (MSC2010). https://mathscinet.ams.org/mathscinet/msc/pdfs/classifications2010.pdf.

79. MSC2020-Mathematics Subject Classification System. https://mathscinet.ams.org/msnhtml/msc2020.pdf.

80. MSC Conversion Table. https://mathscinet.ams.org/mathscinet/msc/conv.html?from=2010.

81. *Buswell S. et al.* (Eds.) The OpenMath Standard. Version: 2.0r2. The OpenMath Society, July 2019.
URL: https://openmath.org/standard/om20-2019-07-01/omstd20.html.

82. *Xie I, Matusiak K.* Discover Digital Libraries: Theory and Practice. Elsevier, 2016.

83. *Bouche T., Labbe O.* The New Numdam Platform // H. Geuvers et al. (Eds.). Proceedings of the 10th International Conference on Intelligent Computer Mathematics (CICM 2017), Edinburgh, UK, July 17–21, 2017. Lecture Notes in Computer Science. Springer, Cham, 2017. Vol. 10383. P. 70–82.
https://doi.org/10.1007/978-3-319-62075-6_6.
URL: https://zenodo.org/record/581405.

84. *Erxleben F., Günther M., Krötzsch M., Mendez J., Vrandečić D*. Introducing Wikidata to the Linked Data Web // P. Mika et al. (Eds.). Proceedings of the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19–23, 2014, Part I. Lecture Notes in Computer Science. Springer, Cham, 2014. Vol. 8796. P. 50–65. https://doi.org/10.1007/978-3-319-11964-9_4.

85. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase. Communications of the ACM. 2014. Vol. 57, Issue 10, October 2014. P. 78–85. https://doi.org/10.1145/2629489.

86. *Scharpf Ph., Schubotz M., Gipp B.* Mathematics in Wikidata // L.-A. Kaffee et al. (Eds.). Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), October 24, 2021. CEUR Workshop Proceedings. CEUR-WS, 2021. Vol. 2982.
URL: http://ceur-ws.org/Vol-2982/paper-1.pdf.

87. *Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachev E.K.* Popolnenie metadannyh dokumentov matematicheskih cifrovyh retro-kollekcij metodom semanticheskih setej // Nauchnyj servis v seti Internet: trudy XXIII Vserossijskoj nauchnoj konferencii. M.: IPM im. M.V. Keldysha, 2021. S. 22–33.
https://doi.org/10.20948/abrau-2021-22.
https://keldysh.ru/abrau/2021/theses/22.pdf.

88. *Harris S. et al. (Eds.).* SPARQL 1.1 Query Language. W3C Recommendation, 21 March 2013. URL: https://www.w3.org/TR/sparql11-query/.

89. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Extraction of Wikidata knowledge for the metadata formation for documents of digital mathematical collections // Russian Digital Libraries Journal. 2021. Vol. 24. No 6. P. 1023–1059.
https://doi.org/10.26907/1562-5419-2021-24-6-1023-1059.

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

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

92. *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.

93. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M.* Metadata Normalization Methods in the Digital Mathematical Library // M. Gorbunov-Posadov et al. (Eds.). Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019), Novorossiysk–Abrau, Russia, September 23–28, 2019. CEUR Workshop Proceedings. CEUR-WS, 2020. Vol. 2543. P. 136–148.
URL: https://ceur-ws.org/Vol-2543/rpaper13.pdf.

СВЕДЕНИЯ ОБ АВТОРАХ

ЕЛИЗАРОВ Александр Михайлович – доктор физико-математических наук, профессор, заслуженный деятель науки Республики Татарстан и Российской Федерации, профессор кафедры программной инженерии Института информационных технологий и интеллектуальных систем Казанского федерального университета.

Научные интересы: цифровые библиотеки, единое пространство научных математических знаний, интеллектуальный анализ данных, рекомендательные системы, облачные вычисления, технологии извлечения знаний.

Alexander ELIZAROV – Doctor of Physics and Mathematics, Professor, Honored Scientist of the Republic of Tatarstan and the Russian Federation, Professor of the Department of Software Engineering of the Institute of Information Technologies and Intelligent Systems of Kazan Federal University.

Research interests: digital libraries, common space of scientific mathematical knowledge, data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: amelizarov@gmail.com;

ORCID: 0000-0003-2546-6897

КИРИЛЛОВИЧ Александр Витальевич – кандидат технических наук, научный сотрудник Института информационных технологий и интеллектуальных систем Казанского федерального университета.

Научные интересы: онтологическое моделирование, Semantic Web, лингвистические открытые связанные данные, управление математическим знанием.

Alexander KIRILLOVICH – Ph.D. in Computer Science, worked as a researcher of the Institute of Information Technologies and Intelligent Systems of Kazan Federal University.

Research interests: ontology engineering, Semantic Web, Linguistic Linked Data, mathematical knowledge management.

e-mail: alik.kirillovich@gmail.com.

ORCID: 0000-0001-9680-449X

ЛИПАЧЁВ Евгений Константинович – кандидат физико-математических наук, доцент кафедры Интеллектуальных технологий поиска Института информационных технологий и интеллектуальных систем Казанского федерального университета.

Научные интересы: цифровые библиотеки, единое пространство научных математических знаний, интеллектуальный анализ данных, рекомендательные системы, облачные вычисления, технологии извлечения знаний.

Evgeny Konstantinovich LIPACHEV – Candidate of Physics and Mathematics, Associate Professor, Kazan Federal University.

Research interests: digital libraries, common space of scientific mathematical knowledge, data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: elipachev@gmail.com;

ORCID: 0000-0001-7789-2332

НЕВЗОРОВА Ольга Авенировна – кандидат технических наук, доцент кафедры информационных систем Института вычислительной математики и информационных технологий Казанского федерального университета.

Научные интересы: онтологическое моделирование, Semantic Web, лингвистические открытые связанные данные, управление математическим знанием.

Olga NEVZOROVA – Candidate of Technical Sciences, Associate Professor of the Information Systems Department of the Institute of Computational Mathematics and Information Technologies of the Kazan Federal University.

Research interests: ontology engineering, Semantic Web, Linguistic Linked Data, mathematical knowledge management.

email: onevzoro@gmail.com.

ORCID: 0000-0001-8116-9446.

Материал поступил в редакцию 21 ноября 2022 года

Переработанная версия – 6 марта 2023 года

УДК 81+004.048

СЕМАНТИЧЕСКИЙ РЕКОМЕНДАТЕЛЬНЫЙ СЕРВИС ПРИСВОЕНИЯ КОДА УДК МАТЕМАТИЧЕСКИМ СТАТЬЯМ

О. А. Невзорова[1] [0000-0001-8116-9446], Д. А. Альмухаметов[2] [0000-0002-4888-7937]

[1,2]*Казанский (Приволжский) федеральный университет, ул. Кремлевская, 35, г. Казань, 420008*

[1]onevzoro@gmail.com, [2]dnlanik@gmail.com

Аннотация

Классификация документов с присвоением кодов-классификаторов является традиционным способом систематизации и поиска документов по определенной тематике. Универсальная десятичная классификация (УДК) лежит в основе систематизации знаний, представленных в библиотеках, базах данных и других хранилищах информации. В России УДК является обязательным реквизитом всей книжной продукции и информации по естественным и техническим наукам. Выбор классификационных кодов связан с анализом структуры дерева классификатора и традиционно выполняется автором научной статьи.

В настоящей работе предложено решение задачи автоматизации подбора классификационного кода УДК для математической статьи на основе специального ресурса – онтологии OntoMath^{PRO} профессиональной математики, разработанной в Казанском федеральном университете. Подходом к решению задачи автоматизации является создание «кодовых карт» для каждого классифицирующего кода в дереве УДК в области математики. Под «кодовой картой» понимается взвешенный набор всех математических именованных сущностей, извлеченных с помощью онтологии OntoMath^{PRO} из коллекции статей с заданным кодом УДК. Создание «кодовых карт» основано на гипотезе о том, что выбор кода УДК обуславливается определённым набором классифицирующих признаков, которые можно представить классами из онтологии OntoMath^{PRO}. Предложенная гипотеза проверена и подтверждена: проверка гипотезы проведена на коллекции математических статей, опубликованных в журнале «Известия ВУЗов. Математика» в течение 1999–2009 гг.

***Ключевые слова**: Универсальная десятичная классификация, кодовая карта, онтология OntoMath^{PRO}, математическая статья*

ВВЕДЕНИЕ

В настоящее время рекомендательные системы используются в самых разных областях, выработаны основные подходы к их построению [1, 2]. Особый интерес представляют рекомендательные системы, ориентированные на издание и подготовку научных публикаций [3]. Такие системы формируют цифровую инфраструктуру электронных научных журналов, включающую программную платформу, реализующую основные рабочие процессы управления электронным журналом, и информационные системы, поддерживающие базовые и дополнительные сервисы с учетом, в частности, специфики предметной области этого журнала [4].

Классификация документов с присвоением кодов-классификаторов является традиционным способом систематизации и поиска знаний. Классификаторы – это тип метаданных в научных документах. Существуют различные национальные и международные универсальные системы классификации. В России широко используются такие классификационные системы, как Библиотечно-библиографическая классификация (ББК), Государственный рубрикатор научно-технической информации (ГРНТИ) и Универсальная десятичная классификация (УДК).

УДК (https://udcc.org) лежит в основе систематизации знаний, представленных в библиотеках, базах данных и других хранилищах информации. Эта классификация принята в качестве основной системы индексации научно-технической документации в большинстве стран мира. В России УДК является обязательным реквизитом для всей книжной продукции и информации по естественным и техническим наукам. В конце 2019 года данный классификатор содержал порядка 126 441 кодов. В настоящее время классификация переведена более чем на 50 языков.

Выбор классификационных кодов связан с анализом структуры дерева классификатора и занимает достаточно много времени. Ниже рассмотрена задача автоматизации подбора кода классификации УДК для математических статей из области УДК 51 «Математика» на основе специального ресурса – онтологии OntoMath^{PRO} профессиональной математики.

Смежные работы

Классификация научных текстов в соответствии с УДК основывается на ключевых словах, содержащихся в тексте [5]. Точно так же библиографические метаданные, такие как заголовок, описание и тематические теги, могут использоваться для дополнения библиографических записей публикации десятичной классификацией Дью (Dewey Decimal Classification, DDC) [6]. Распространение цифровых ресурсов и их интеграции в традиционную библиотечную среду создали потребность в автоматизированном инструменте для определения тематики публикации в соответствии со схемами библиотечной классификации.

Обзор методов, таких как контентно-ориентированная и совместная фильтрация, графические и гибридные методы, можно найти в работе Bai et al. [7]. Анализ использования сервисов рекомендаций для научных кругов представлен в исследовании Bell et al. [8]. В [9] дан исчерпывающий обзор современных рекомендательный систем на основе глубокого машинного обучения. Методы машинного обучения используются в различных научных рекомендательных системах [10, 11]. В [10] авторы исследуют возможность автоматического назначения первичной классификации с использованием схемы математической предметной классификации (Mathematics Subject Classification, MSC), рассматривая проблему назначения классифицирующего кода как задачу мультиклассовой классификации машинного обучения. В [11] обсуждается модель на основе машинного обучения, предназначенная для автоматической классификации старых оцифрованных текстов из словенской цифровой библиотеки. Классификационные коды УДК новых научных работ, назначенные специалистами людьми, использовались для построения классификационной модели УДК старых оцифрованных текстов. В этой модели использовались различные алгоритмы кластеризации. Авторы названной статьи утверждают, что наиболее эффективным классификатором был SVM с использованием TF-IDF. В отличие от описанных ранее работ, в нашей работе рассмотрена задача автоматизации подбора кода классификации УДК для математических статей на основе специального ресурса – онтологии OntoMath^{PRO} профессиональной математики [12].

Онтология OntoMath^{PRO}

Онтология OntoMath^{PRO} – прикладная онтология для автоматической обработки профессиональных математических статей на русском и английском языках, разработанная в Казанском федеральном университете. Эта онтология охватывает широкий спектр областей математики, таких как теория чисел, теория множеств, алгебра, анализ, геометрия, теория вычислений, дифференциальные уравнения, численный анализ, теория вероятностей и статистика. Каждый концепт онтологии имеет аннотацию, имя на русском и английском языках, включая синонимы. Терминологическими источниками, использованными при разработке OntoMath^{PRO}, служили классические учебники, интернет-ресурсы, такие как Кембриджский математический тезаурус, статьи из научных журналов, например, журнала «Известия высших учебных заведений. Математика».

В онтологии можно выделить две таксономии по отношению ISA – иерархия областей математики и иерархия объектов математического знания. Первая иерархия близка к Универсальной десятичной классификации. Верхний уровень второй таксономии содержит понятия трех типов: 1) основные математические понятия (например, «Множество» и «Оператор»); 2) понятия, относящиеся к конкретным областям математики и заданные в соответствующих иерархиях (например, «Элемент теории вероятностей» или «Элемент численного анализа»); 3) общие научные понятия (например, «Задача», «Метод», «Утверждение», «Формула» и пр.).

Онтология OntoMath^{PRO} разработана на языке OWL-DL/RDFS и содержит в настоящий момент времени 3 450 классов, 5 свойств объектов, 3 630 экземпляров подклассов свойств и 1 140 экземпляров других свойств. Постоянно происходит дальнейшее наполнение этой онтологии.

Описание подхода

Исследование проведено на коллекции статей из выпусков, опубликованных журналом «Известия высших учебных заведений. Математика» за 10 лет (с 1999 по 2009 годы). Коллекция содержит 1 356 математических статей в формате XML. Каждая статья имеет как минимум один код УДК. В рассмотренных выпусках

наибольшее количество статей пришлось на классифицирующий код УДК 517 «Анализ», всего в коллекции оказалось 883 статьи с данным кодом.

Предлагаемый нами подход к автоматическому назначению классифицирующего кода УДК математическим статьям основан на использовании онтологии OntoMath^{PRO}. Как отмечено выше, онтология содержит базовые понятия, такие как задача, система, теория, уравнение, формула и т. д. Ключевая идея предлагаемого подхода состоит в том, что выбор классифицирующего кода УДК базируется на определенных наборах классифицирующих признаков, которые использует автор статьи. Эти признаки представлены в онтологии базовыми математическими понятиями. Задачей исследования было выделение наиболее релевантных признаков среди онтологических понятий, определяющих выбор классифицирующего кода УДК.

Нами был проведен опрос экспертов-математиков с целью выяснения, какие признаки являются для них определяющими при выборе классифицирующего кода УДК для научной статьи. В результате был сделан вывод, что наиболее значимыми признаками являются метод, задача и уравнение, что составляет содержание принятой рабочей гипотезы.

Для проверки этой гипотезы был проведен ряд экспериментов на наиболее репрезентативной подколлекции с кодом УДК 517 («Анализ») из имеющейся коллекции математических статей. В экспериментах попарно сравнивались подколлекции с разными кодами УДК. Выбор кодов был основан на их положении в иерархии дерева УДК (разные поддеревья первого уровня в кодовом дереве с корневой вершиной под номером 517), родстве (потомки одного предка) и размере подколлекций.

В экспериментах использовалась подсистема семантической аннотации, которая обеспечивала функциональные возможности для аннотирования статей с точки зрения фиксированного набора предметных областей онтологии OntoMath^{PRO}. Из текста статьи извлекались все математические именованные сущности (Mathematical Named Entity, MNE), распознаваемые онтологией, и на основе словаря онтологии составлялся вектор документа.

Для процесса оценки релевантности классификационных признаков использовался модуль фильтрации математических именованных сущностей, который получал на вход два набора подколлекций статей с разными кодами УДК и список классифицирующих признаков. Результатом работы модуля являлся набор именованных математических сущностей, отобранных на основе выбранных классифицирующих признаков, для определенных кодов УДК. Модуль оценки сравнивал два полученных набора, определяя общие и специфичные признаки для каждого кода УДК. В результате модуль определял актуальность каждого классифицирующего признака для соответствующего кода УДК.

Обозначим S(f_i, c_j) – набор выделенных именованных сущностей для статей с кодом УДК c_j, отобранных по признаку f_i. Для оценки релевантности классифицирующего признака для определенного кода УДК использовалась следующая формула

$$REL_{c_j c_k}^{f_i} = \frac{S(f_i,c_j) \cap S(f_i,c_k)}{S(f_i,c_j) \cup S(f_i,c_k)}.$$

Оценка релевантности классифицирующего признака f_i представляет собой нечеткую лингвистическую переменную со значениями «слабый», «умеренный», «сильный». Были предложены следующие экспертные правила для выявления различий/сходства в паре подколлекций.

Если значение функции оценки $REL_{c_j c_k}^{f_i}$ находится в диапазоне [0..0.3], то можно говорить о сильном различии в паре подколлекций УДК по данному признаку.

Если значение функции оценки $REL_{c_j c_k}^{f_i}$ находится в диапазоне [0.3..0.7], то пара подколлекций УДК является умеренно различимой по данному признаку.

Если значение функции оценки $REL_{c_j c_k}^{f_i}$ находится в диапазоне [0.7..1], то пара подколлекций УДК слабо различима по данному признаку.

Результаты нескольких экспериментов представлены ниже. На диаграмме показано количество общих и специфичных терминов классифицирующих признаков для пары подколлекций с выбранными кодами УДК.

В первом эксперименте были рассмотрены подколлекции с кодами УДК одного уровня и сопоставимые по размерам: УДК 517.51 «Функции действительных

переменных. Действительные функции» (89 статей), УДК 517.54 «Конформное отображение и геометрические вопросы теорий комплексного переменного. Аналитические функции и их обобщение» (87 статей), УДК 517.97 «Вариационное исчисление и математическая теория оптимального управления» (75 статей).

Результаты эксперимента показаны на рис. 1, а интерпретация этих результатов в терминах введенной нечеткой лингвистической переменной представлена на таблице 1. Синим цветом на рис. 1 отображено число общих математических именованных сущностей для пары коллекций в сравнении, оранжевым – число уникальных сущностей для коллекции с подкодом 51, серым – для коллекции с подкодом 54, а желтым – для коллекции с подкодом 97.

Рассмотрим сравнение коллекций с подкодами 51 и 54, оранжевый и серый цвета. На графике видно достаточно представительное ядро методов у этих коллекций, что можно объяснить их родством в дереве УДК. Но при этом коллекция с подкодом 54 располагает большим числом уникальных задач и уравнений, и на основе данных «экспертных классов» мы можем различать эту пару УДК.

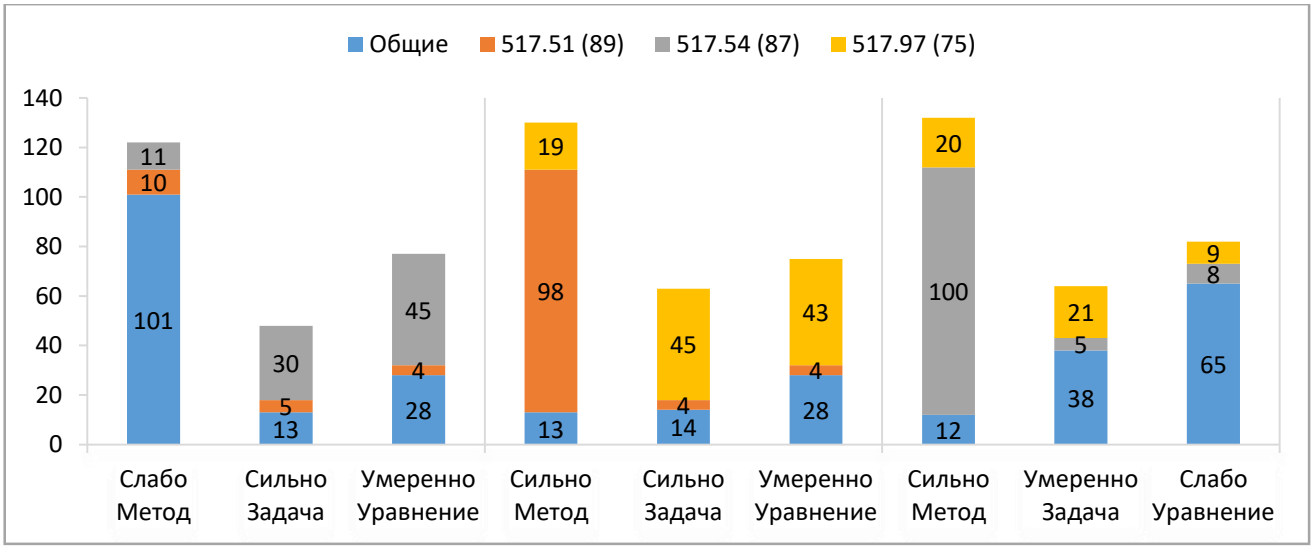


Рис. 1. Результаты эксперимента 1

	Метод	**Задача**	**Уравнение**
517.51 & 517.54	Слабо	Сильно	Умеренно
517.51 & 517.97	Сильно	Сильно	Умеренно
517.54 & 517.97	Сильно	Умеренно	Слабо

Таблица 1. Оценка релевантности классифицирующих признаков эксперимента 1

Рассмотрим коллекции с подкодами 51 и 97 (оранжевый и желтый цвета на рисунке 1). В коллекции с подкодом 97 не используется такое большое число методов, как в коллекции с подкодом 51. Но при этом в коллекции с подкодом 97, по сравнению с коллекцией с подкодом 51, преобладают задачи и уравнения. Данную пару мы можем различать по всем трем «экспертным классам».

Рассматривая пару коллекций с подкодами 54 и 97 (серый и желтый цвета на рисунке 1), мы видим такую же тенденцию по методам, как и в предыдущем сравнении. В столбце уравнений видно, что эти коллекции используют общий набор уравнений, и, следовательно, мы не можем различать данную пару по этому признаку. В задачах же преобладает коллекция с подкодом 97. Таким образом, для классификации этих коллекций можно использовать методы и задачи.

Второй эксперимент проводился между одноуровневыми подклассами одного класса УДК 517.9 «Дифференциальные, интегральные и другие функциональные уравнения. Вариационное исчисление и конечные разности», имеющими наибольшее количество статей среди подклассов (рис. 2). В эксперименте участвовали следующие коллекции: УДК 517.92 «Методы решения различных типов уравнений и систем уравнений» (192 статьи), УДК 517.95 «Дифференциальные уравнения с частными производными» (156 статей) и УДК 517.98 «Функциональный анализ и теория операторов» (133 статьи). Синим цветом на рисунке 2 отображено число общих математических именованных сущностей для пары коллекций в сравнении, оранжевым – число уникальных сущностей для коллекции с подкодом 92, серым – для коллекции с подкодом 95, а желтым – для коллекции с подкодом 98. Интерпретация результатов эксперимента согласно формуле оценки релевантности представлена в таблице 2.

Рассмотрим коллекции с подкодами 92 и 95 (оранжевый и серый цвета на рисунке 2). Пара коллекций использует общий набор уравнений и не может классифицироваться по этому признаку. В коллекции с подкодом 95 преобладают методы и задачи, и по этим «экспертным классам» мы можем различать данную пару.

Пару коллекций с подкодами 92 и 98 (оранжевый и желтый цвета на рисунке 2) мы можем уверенно различать только по методам, поскольку они используют общие наборы задач и уравнений.

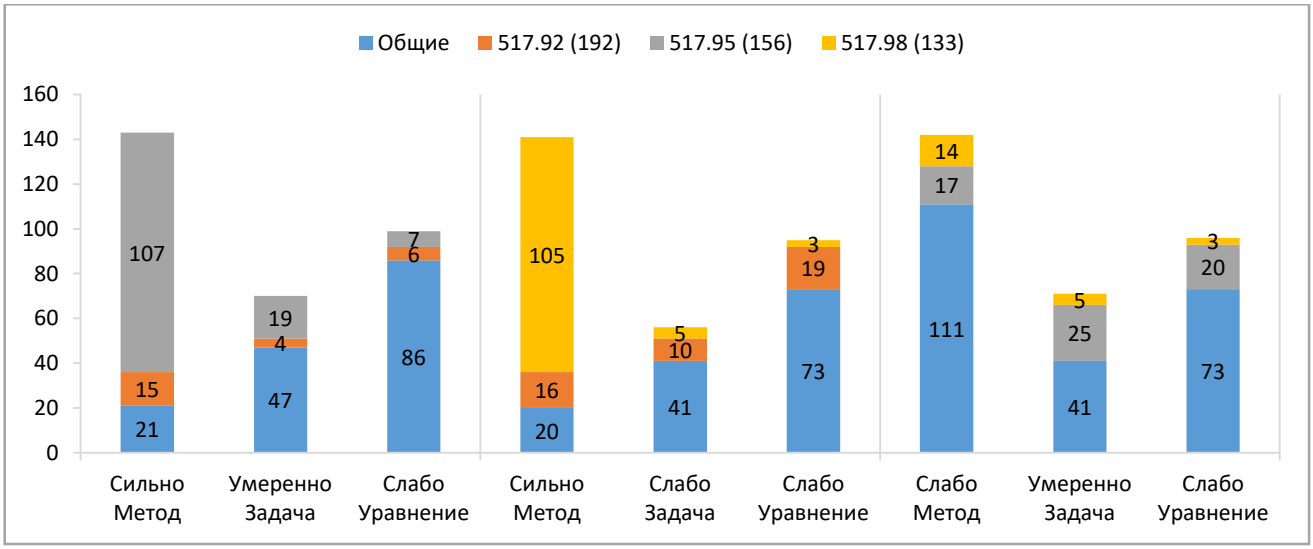


Рис. 2. Результаты эксперимента 2

	Метод	**Задача**	**Уравнение**
517.92 & 517.95	Сильно	Умеренно	Слабо
517.92 & 517.98	Сильно	Слабо	Слабо
517.95 & 517.98	Слабо	Умеренно	Слабо

Таблица 2. Оценка релевантности классифицирующих признаков эксперимента 2

Перейдем к паре коллекций с подкодами 95 и 98 (серый и желтый цвета на рисунке 2). Статьи с такими УДК используют общие наборы методов и уравнений и могут классифицироваться только по задачам.

Коллекции статей с кодами УДК одного предка обладают большим количеством общих представителей экспертных классов, что объясняется их родством в дереве УДК, тем не менее, мы все еще можем их различать.

В третьем эксперименте были рассмотрены классифицирующие коды УДК узкоспециализированной направленности: УДК 517.956 «Линейные и квазилинейные уравнения и системы» (57 статей), УДК 517.958 «Дифференциальные и интегральные уравнения математической физики» (59 статей), УДК 517.982 «Линейные пространства, снабженные топологией, порядком и другими структурами» (21 статья) и УДК 517.983 «Линейные операторы и операторные уравнения» (36 статей). Светло-синим цветом на рисунке 3 отображено число общих математических именованных сущностей для пары коллекций в сравнении, оранжевым – число уникальных сущностей для коллекции с подкодом 956, серым – для коллекции с подкодом 958, желтым – для коллекции с подкодом 982, а темно-синим – для коллекции с подкодом 983. Интерпретация результатов представлена в таблице 3.

По данным эксперимента видно, что все группы статей могут в той или иной степени классифицироваться по «экспертным классам». Малое количество извлеченных концептов «экспертных классов» в коллекциях с кодами 982 и 983 можно связать с недостаточным их представительством в нашей коллекции статей.

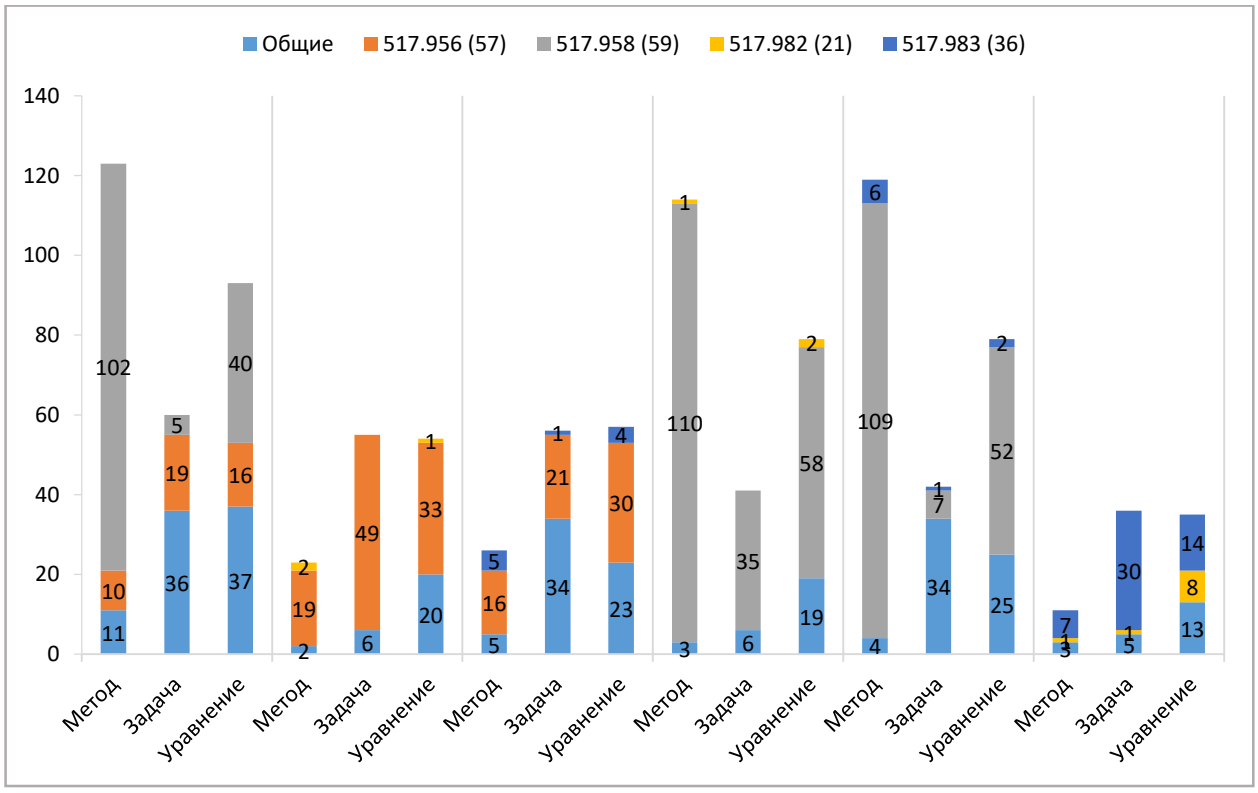


Рис. 3. Результаты эксперимента 3

	Метод	**Задача**	**Уравнение**
517.956 & 517.958	Сильно	Умеренно	Умеренно
517.956 & 517.982	Сильно	Сильно	Умеренно
517.956 & 517.983	Сильно	Умеренно	Умеренно
517.958 & 517.982	Сильно	Сильно	Сильно
517.958 & 517.983	Сильно	Слабо	Сильно
517.982 & 517.983	Умеренно	Сильно	Умеренно

Таблица 3. Оценка релевантности классифицирующих признаков эксперимента 3

Проведенное нами исследование подтверждает предложенную гипотезу о том, что группу математических кодов УДК можно классифицировать по таким признакам, как «метод», «задача» и «уравнение».

Основываясь на результатах проверки гипотезы, представляется перспективным создание «кодовых карт» для каждого кода УДК в области «Математика». Под кодовой картой мы подразумеваем взвешенный набор всех извлеченных именованных математических сущностей из подколлекции статей с определенным кодом УДК.

Кодовая карта

Кодовая карта строится на основе словаря онтологии OntoMath^{PRO}. На рисунке 4 представлена иерархия онтологии «Элемент математического знания», которая включает такие общие концепты, как *величина, геометрический объект, гипотеза, задача, метод, множество, неравенство, оператор, операция, отображение, оценка, преобразование, равенство, тензор, теорема, уравнение, утверждение, формула, характеристика* и др.

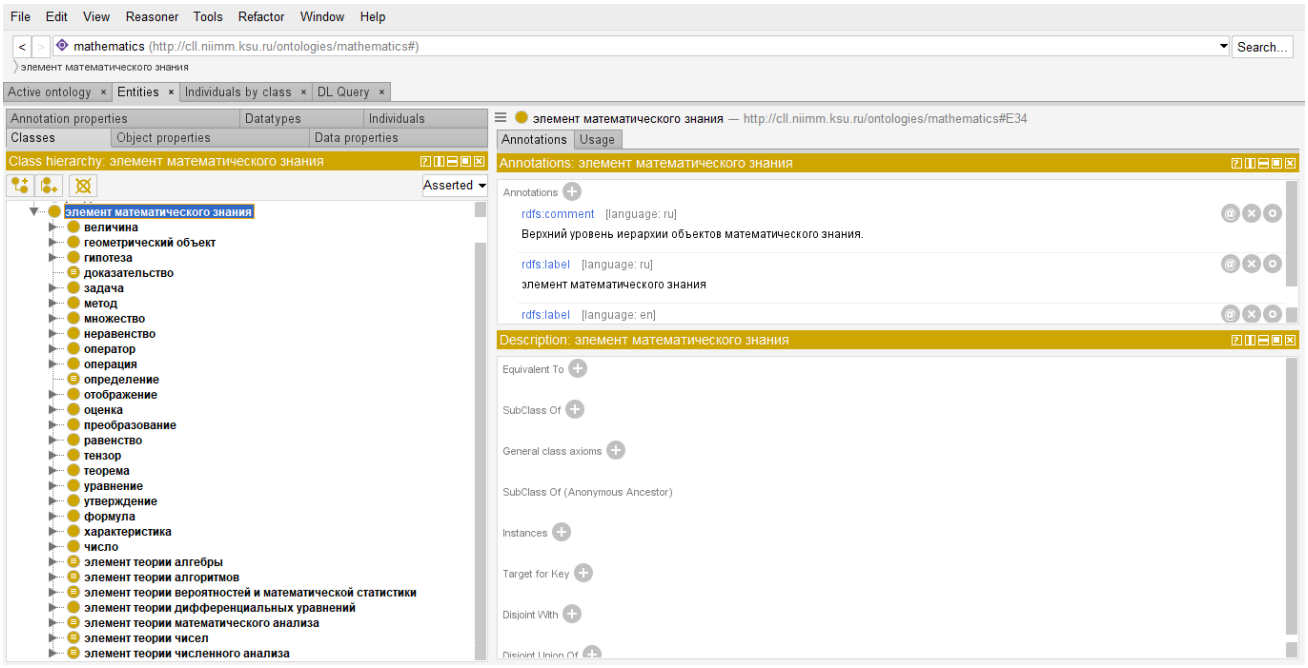
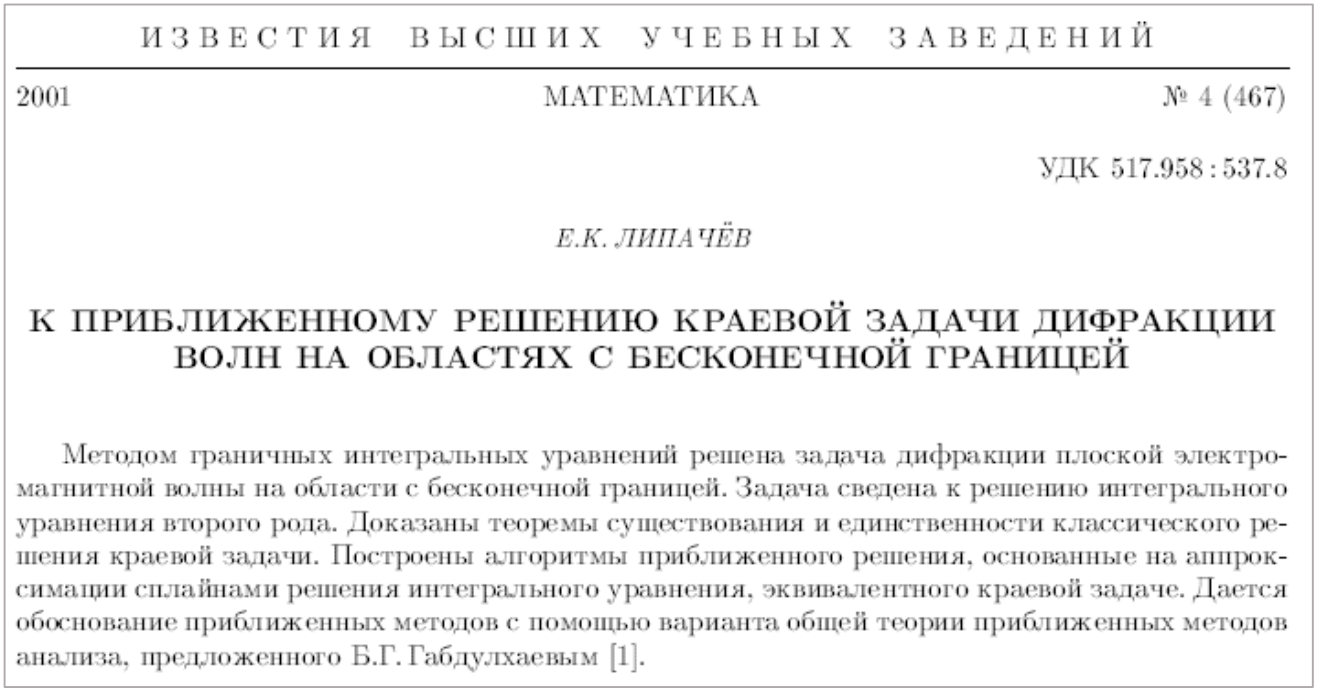


Рис. 4. Иерархия «Элементы математического знания» в онтологии OntoMath^{PRO}

Всего онтология OntoMath^{PRO} содержит сегодня 3 450 классов, 5 свойств объектов, 3 630 экземпляров подклассов свойств и 1 140 экземпляров других свойств. Например, класс *геометрический объект* содержит 333 подкласса, класс *задача* – 125 подклассов, а класс *метод* – 500 подклассов.

В рекомендательной системе предлагается использовать общий шаблон для формирования кодовых карт кодов УДК и карт статей. Шаблон содержит 2739 термов из 22 основных класса из иерархии элементов математического знания. Оценка близости статьи к определенному коду УДК происходит посредством нормировки карты статьи и сравнивания её с кодовой картой УДК.



ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

2001 МАТЕМАТИКА № 4 (467)

УДК 517.958 : 537.8

Е.К. ЛИПАЧЁВ

К ПРИБЛИЖЕННОМУ РЕШЕНИЮ КРАЕВОЙ ЗАДАЧИ ДИФРАКЦИИ ВОЛН НА ОБЛАСТЯХ С БЕСКОНЕЧНОЙ ГРАНИЦЕЙ

Методом граничных интегральных уравнений решена задача дифракции плоской электромагнитной волны на области с бесконечной границей. Задача сведена к решению интегрального уравнения второго рода. Доказаны теоремы существования и единственности классического решения краевой задачи. Построены алгоритмы приближенного решения, основанные на аппроксимации сплайнами решения интегрального уравнения, эквивалентного краевой задаче. Дается обоснование приближенных методов с помощью варианта общей теории приближенных методов анализа, предложенного Б.Г. Габдулхаевым [1].

Рис. 5. Фрагмент статьи из журнала «Известия высших учебных заведений. Математика».

Для классификации в качестве примера приведем статью Е.К. Липачёва «К приближенному решению краевой задачи дифракции волн на областях с бесконечной границей» [13]. Автор в своей статье указал два кода УДК: УДК 517.958 «Дифференциальные и интегральные уравнения математической физики» и УДК 537.8 «Электромагнетизм. Электромагнитное поле. Электродинамика. Теория Максвелла» (рис. 5).

Рекомендательная система также назначила статье код УДК 517.958. Выделим классификационные признаки, послужившие основанием для отнесения к конкретному классу УДК, и сравним две близкие кодовые карты в дереве классификатора.

Рассмотрим кодовые карты для кода УДК 517.956 «Линейные и квазилинейные уравнения и системы» и кода УДК 517.958, которые являются наследниками кода УДК 517.95 «Дифференциальные уравнения с частными производными».

Статистика по экспертным классам «методы», «задачи» и «уравнения», термы которых содержатся в кодовых картах, а также данные о пересечении списков термов из статьи и кодовых карт по этим классам приведена в таблице 4.

	Метод	**Задача**	**Уравнение**
517.956	59	51	37
517.958	75	51	29
Статья ∩ 517.956	5	1	5
Статья ∩ 517.958	9	3	5

Таблица 4. Статистика по классификационным термам из экспертных классов

В данном примере в пересечении списков термов из статьи и кодовой карты УДК 517.958 содержится 5 термов, которые входят в набор термов пересечения из статьи и кодовой карты УДК 517.956. Множество термов пересечения включает такие общие термы, как «вычислительная схема», «метод», «анализ», «спектральный метод» и «метод интегральных уравнений». Дополнительными классифицирующими термами для УДК 517.958 служат еще 4 терма: «метод граничных интегральных уравнений», «метод обобщенных потенциалов», «метод коллокаций» и «метод сплайн-коллокаций». Термы для класса уравнений совпадают для указанных кодов УДК, выделены термы «уравнение», «уравнение Фредгольма», «уравнение Фредгольма первого рода», «уравнение Фредгольма второго рода» и «уравнение Гельмгольца». По классу «задача» выделены общий терм «задача» и дополнительные термы по пересечению статьи и кодовой карты УДК 517.958 «задача численного решения интегральных уравнений» и «задача численного решения интегральных уравнений Фредгольма второго рода».

Реализация рекомендательной системы

Для реализации рекомендательной системы были выбраны высокоуровневый язык программирования общего назначения *Python* и свободный фреймворк для веб-приложений *Django*. В качестве СУБД использовалась *SQLite*. Для обработки загружаемых в систему научных статей в реальном времени применен менеджер задач с открытым исходным кодом *Celery*. В качестве брокера сообщений выбран *Redis*. В данный момент рекомендательная система работает с файлами в формате *PDF*. Для извлечения текста из статьи использован инструмент с открытым исходным кодом для оптического распознавания символов на основе нейронной сети *Tesseract OCR*.

На рис. 6 приведен интерфейс личного кабинета, в котором пользователь может вводить свои данные, а также увидеть статус обработки статьи и рекомендации по выбору классифицирующего кода УДК для загруженной в систему статьи. Система способна давать рекомендации по уточнению кода УДК (пример 1), правильно классифицировать код УДК статьи (примеры 2 и 3) или корректно определять общую тематическую направленность работы (пример 4).

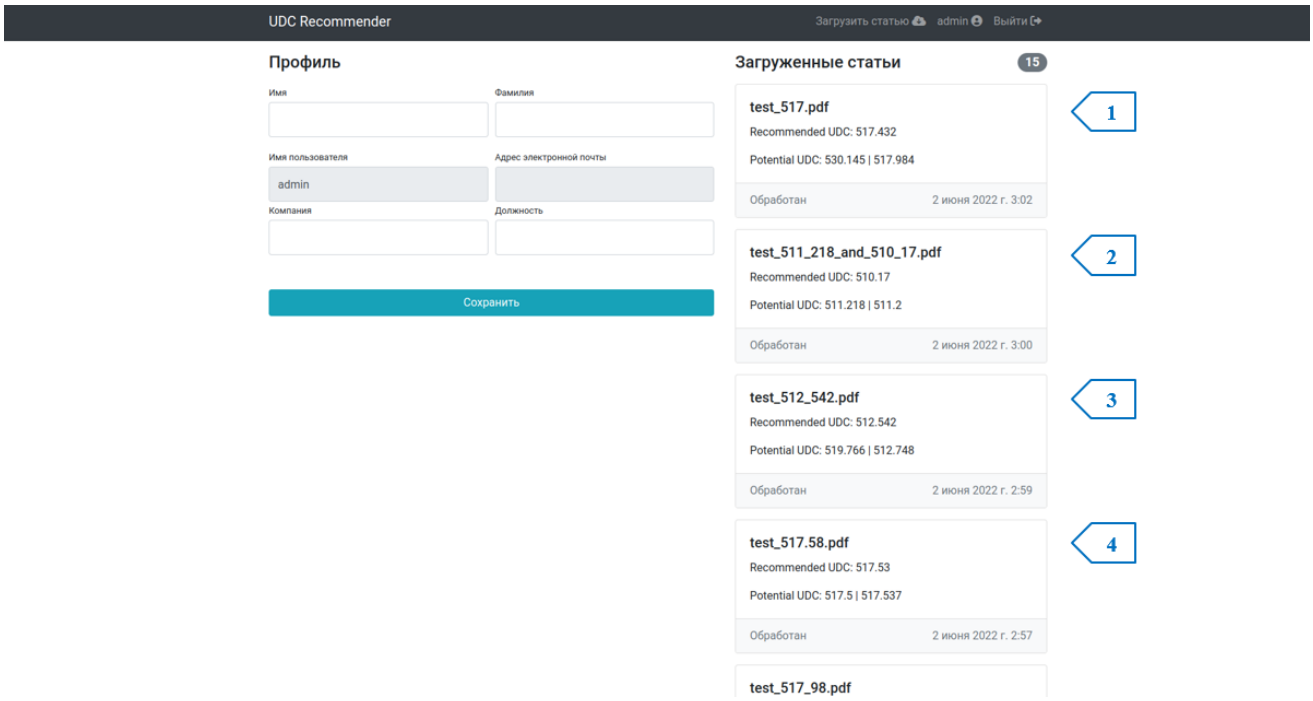


Рис. 6. Личный кабинет пользователя

Для разработки текущей версии рекомендательной системы использовалась коллекция статей журнала «Известия высших учебных заведений. Математика» за 50 лет (1968–2018 г.г.). Коллекция содержит более 6000 статей с назначенными классифицирующими кодами УДК. Статьям присвоены 622 различных кода УДК, из них 564 приходятся на раздел 51 «Математика», в котором имеется 1660 классификационных кодов. Процент успеха классификации варьируется от 30 до 80 процентов в зависимости от размера обучающей подколлекции и узкоспециализированной направленности кода УДК. Предполагается, что в дальнейшем к системе будут подключены внешние источники загрузки статей для увеличения размера обучающей коллекции и повышения качества классификации.

В таблице 5 приведены данные о качестве классификации для различных кодов УДК, находящихся на разных уровнях в иерархии дерева УДК. Здесь указаны код УДК, количество статей, содержащихся в подколлекциях с данным кодом УДК, и процент успешных классификаций наборов тестовых статей с данным кодом. Классификация считалась успешной, если хотя бы один из трех рекомендованных кодов совпадал с кодом статьи или уточнял его.

УДК	**Кол-во статей**	**Процент успеха**
510	48	84 %
511	67	87 %
512	472	74 %
514	443	64 %
515	54	76 %
517.51	540	47 %
517.54	372	58 %
517.97	150	42 %
517.98	470	53 %
517.512	156	52%
517.518	213	36%
517.544	212	47%
517.929	122	37%
517.956	183	64%
517.968	126	26%
517.983	121	47%

Таблица 5. Оценка качества классификации тестовых наборов

В подколлекциях с высоким уровнем в иерархии УДК – 510 «Фундаментальные и общие проблемы математики», 511 «Теория чисел», 512 «Алгебра», 514

«Геометрия» и 515 «Топология» – процент успешной классификации высок, поскольку названные области математики сильно отличаются по терминологии, и их легко отличить на основе словаря онтологии OntoMath^{PRO}.

При переходе на более низкий уровень иерархии УДК, в частности, на примере коллекции 517 «Анализ», процент успеха классификации снижается, поскольку тексты имеют более близкую направленность, и сложность классификации возрастает.

При переходе ниже в иерархии УДК, как ожидалось, процент успеха также снижается, несмотря на достаточно представительный размер подколлекций. Снижение не наблюдается у подколлекции 517.956 «Линейные и квазилинейные уравнения и системы», это, скорее всего, связано со спецификой тематики подколлекции. Самое большое снижение наблюдается у подколлекции 517.968 «Интегральные уравнения», поскольку термы из данной тематики широко применяются в смежных коллекциях.

В настоящее время проводятся дополнительные исследования для определения наиболее подходящей рекомендательной модели и выбора соответствующих весов для классифицирующих признаков.

Заключение

В статье представлены результаты разработки рекомендательной системы, ориентированной на автоматическое присвоение кодов УДК научным статьям в области УДК 51 «Математика». Решение задачи автоматизации подбора кода УДК для математической статьи основано на специальном ресурсе – онтологии OntoMath^{PRO} профессиональной математики. Подходом к решению задачи автоматизации является создание кодовых карт для каждого кода в дереве УДК в области математики. Под кодовой картой подразумевается взвешенный набор всех математических именованных сущностей, извлеченных с помощью онтологии OntoMath^{PRO} из коллекции статей с заданным кодом УДК. Создание кодовых карт основано на гипотезе о том, что выбор кода УДК обусловлен определённым набором классифицирующих признаков, в качестве которых могут выступать классы математических именованных сущностей, выбранных из онтологии.

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда, проект № 21-11-00105.

СПИСОК ЛИТЕРАТУРЫ

1. *Lu J., Wu D., Mao M., Wang W., Zhang G.* Recommender system application developments: A survey // Decision Support Systems. 2015. V. 74. P. 12–32.

2. *Ricci F.* Recommender Systems: Models and Techniques // Encyclopedia of Social Network Analysis and Mining. Springer: 2014, P. 1511–1522. https://doi.org/10.1007/978-1-4614-6170-8_88

3. *Elizarov A.M., Lipachev E.K.* Methods of processing large collections of scientific documents and the formation of digital mathematical library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.

4. *Elizarov A.M., Lipachev E.K.* Big Math methods in Lobachevskii-DML digital library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

5. *Romanov A.Y., Lomotin K.E., Kozlova E.S., Kolesnichenko A.L.* Research of neural networks application efficiency in automatic scientific articles classification according to UDC // 2016 International Siberian Conference on Control and Communications (SIBCON), Moscow, Russia, 12–14 May, 2016. IEEE: 2016, P. 7–11. https://doi.org/10.1109/SIBCON.2016.7491783

6. *Khoo M.J., Ahn J.W., Binding C., Jones H.J., Lin X., Massam D., Tudhope D.* Augmenting Dublin core digital library metadata with Dewey decimal classification // Journal of Documentation. 2015. V. 71. No. 5. P. 976–998.

7. *Bai X., Wang M., Lee I., Yang Z., Kong X., Xia F.* Scientific paper Recommendation: a survey // IEEE Access. IEEE. 2019. V. 7. P. 9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

8. *Beel J., Aizawa A., Breitinger C., Gipp B.* Mr. DLib: recommendations-as-a-service (RaaS) for academia // Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2017.

9. *Zhang S., Yao L., Sun A., Tay Y.* Deep Learning Based Recommender System: A Survey and New Perspectives // ACM Computing Surveys. 2019. V. 52(1). P. 1–38.

10. *Schubotz M.* et al. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020.

11. *Kragelj M., Kljajić Borštnar M.* Automatic classification of older electronic texts into the Universal Decimal Classification–UDC // Journal of Documentation. 2021. V. 77. No. 3.

12. *Nevzorova O.A., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* OntoMath^{PRO} Ontology: a Linked data hub for mathematics // 5th International Conference, KESW 2014, Kazan, Russia, September 29 – October 1, 2014. Springer: 2014. V. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4

13. *Липачёв Е.К.* К приближенному решению краевой задачи дифракции волн на областях с бесконечной границей // Изв. Вузов. Математика. 2001. № 4 (467). С. 69–72.

SEMANTIC RECOMMENDATION SERVICE FOR ASSIGNING UDC CODE TO MATHEMATICAL ARTICLES

O. A. Nevzorova[1] [0000-0001-8116-9446], D. A. Almukhametov[2] [0000-0002-4888-7937]

[1,2]*Kazan (Volga Region) Federal University, 35 Kremlyovskaya str., Kazan, 42008*
[1]onevzoro@gmail.com, [2]dnlanik@gmail.com

Abstract

Classification of documents with the assignment of classifier codes is a traditional way of systematizing and searching for documents on a specific topic. The Universal Decimal Classification (UDC) underlies the systematization of knowledge presented in libraries, databases and other information repositories. In Russia, UDC is an obligatory attribute of all book production and information on natural and technical sciences. The choice of classification codes is associated with the analysis of the structure of the classifier tree and is traditionally decided by the author of a scientific article. This article proposes a solution for automating the assigning the UDC classification code for a mathematical article based on a special resource – the OntoMath^{PRO} ontology for professional mathematics, developed at Kazan Federal University. An approach

to solving the problem is to create "code maps" for each classifying code in the UDC tree in the field of mathematics. Under the "code map" is meant a weighted set of all extracted, with the help of OntoMath^{PRO} ontology, mathematical named entities from the collection of articles with a given UDC code. The creation of "code maps" is based on the hypothesis that the choice of the UDC code is determined by a certain set of classifying features that can be represented by classes from the OntoMath^{PRO} ontology. The proposed hypothesis was tested and confirmed in the paper. The hypothesis was tested on a collection of mathematical articles An approach to solving the problem is to create "code maps" for each classifying code in the UDC tree in the field of mathematics. Under the "code map" is meant a weighted set of all extracted, with the help of OntoMath^{PRO} ontology, mathematical named entities from the collection of articles with a given UDC code. The creation of "code maps" is based on the hypothesis that the choice of the UDC code is determined by a certain set of classifying features that can be represented by classes from the OntoMath^{PRO} ontology. The proposed hypothesis was tested and confirmed in the paper. The hypothesis was tested on a collection of mathematical articles published during 1999-2009 in the "Izvestiya VUZov. Mathematics" journal.

Keywords*: the Universal Decimal Classification, code map, the OntoMath^{PRO} ontology, mathematical article*

REFERENCES

1. *Lu J., Wu D., Mao M., Wang W., Zhang G.* Recommender system application developments: A survey // Decision Support Systems. 2015. V. 74. P. 12–32.

2. *Ricci F.* Recommender Systems: Models and Techniques // Encyclopedia of Social Network Analysis and Mining. Springer: 2014, P. 1511–1522. https://doi.org/10.1007/978-1-4614-6170-8_88

3. *Elizarov A.M., Lipachev E.K.* Methods of processing large collections of scientific documents and the formation of digital mathematical library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.

4. *Elizarov A.M., Lipachev E.K.* Big Math methods in Lobachevskii-DML digital library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

5. *Romanov A.Y., Lomotin K.E., Kozlova E.S., Kolesnichenko A.L.* Research of neural networks application efficiency in automatic scientific articles classification according to UDC // 2016 International Siberian Conference on Control and Communications (SIBCON), Moscow, Russia, 12–14 May, 2016. IEEE: 2016, P. 7–11. https://doi.org/10.1109/SIBCON.2016.7491783

6. *Khoo M.J., Ahn J.W., Binding C., Jones H.J., Lin X., Massam D., Tudhope D.* Augmenting Dublin core digital library metadata with Dewey decimal classification // Journal of Documentation. 2015. V. 71. No. 5. P. 976–998.

7. *Bai X., Wang M., Lee I., Yang Z., Kong X., Xia F.* Scientific paper Recommendation: a survey // IEEE Access. IEEE. 2019. V. 7. P. 9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

8. *Beel J., Aizawa A., Breitinger C., Gipp B.* Mr. DLib: recommendations-as-a-service (RaaS) for academia // Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2017.

9. *Zhang S., Yao L., Sun A., Tay Y.* Deep Learning Based Recommender System: A Survey and New Perspectives // ACM Computing Surveys. 2019. V. 52(1). P. 1–38.

10. *Schubotz M.* et al. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020.

11. *Kragelj M., Kljajić Borštnar M.* Automatic classification of older electronic texts into the Universal Decimal Classification–UDC // Journal of Documentation. 2021. V. 77. No. 3.

12. *Nevzorova O.A., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* OntoMath^{PRO} Ontology: a Linked data hub for mathematics // 5th International Conference, KESW 2014, Kazan, Russia, September 29 – October 1, 2014. Springer: 2014. V. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4

13. *Lipachev E.K.* Approximation solution of the boundary value problem of wave diffraction on domain with infinite boundary // Izv. VUZ. Mathematics. 2001. No. 4 (467). P. 69–72.

СВЕДЕНИЯ ОБ АВТОРАХ

НЕВЗОРОВА Ольга Авенировна – доцент кафедры информационных систем Института вычислительной математики и информационных технологий Казанского федерального университета, к. т. н. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

Olga Avenirovna NEVZOROVA – Kazan Federal University, Institute of Computational Mathematics and Information Technologies, Associated Professor of the Department of Information System, PhD. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: onevzoro@gmail.com
ORCID: 0000-0001-8116-9446

АЛЬМУХАМЕТОВ Дамир Альбертович – инженер кафедры программной инженерии Института информационных технологий и интеллектуальных систем Казанского федерального университета. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

Damir Albertovich ALMUKHAMETOV – Engineer of the Department of Software Engineering of the Institute of Information Technology and Information Systems of Kazan Federal University. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: dnlanik@gmail.com
ORCID: 0000-0002-4888-7937

Материал поступил в редакцию 6 февраля 2023 года

УДК 004.415.25

ИНСТРУМЕНТЫ БАЛАНСИРОВАНИЯ ИГР

Г. Ф. Сахибгареева[1] [0000-0003-4673-3253], В. В. Кугуракова[2] [0000-0002-1552-4910], Э. С. Большаков[3] [0000-0002-2208-9515]

[1,2,3] *Институт информационных технологий и интеллектуальных систем Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

[1]gulnara.sahibgareeva42@gmail.com, [2]vlada.kugurakova@gmail.com, [3]edward.bolshakov117@gmail.com

Аннотация

Для раскрытия понятия игрового баланса и разработки подхода к автоматизации рутины при работе с игровой экономикой применены методы анализа данных и проведения экспериментов. По результатам анализа существующих определений выделены частный случай математического баланса и обобщенное дизайнерское определение игрового баланса. Благодаря анализу существующих подходов к балансированию и тестированию игр разработано видение собственного решения. На основе полученных выводов предложен подход к автоматизации балансирования в рамках генератора игрового прототипа. В качестве промежуточного итога представлены обновленная структура и порядок работы инструмента генерации игровых прототипов. Описаны перспективы дальнейшего развития исследований в данном направлении.

Ключевые слова*: игровой баланс, математический баланс, автоматическое балансирование игры, Machinations, генерация игровых прототипов.*

ВВЕДЕНИЕ

В научном и коммерческом направлениях разработки игр сформировалось разностороннее представление о том, что такое баланс, как над ним работать, как автоматизировать рутинные действия на этапе тестирования экономической системы игры. В данной работе рассмотрены разные определения игрового баланса и представлено различие между частным понятием математического баланса и обобщенным дизайнерским понятием игрового баланса.

Эффективным и единственным на данный момент прикладным инструментом для создания экономической системы игры вне игровых движков является платформа Machinations [1]. Чтобы доказать эффективность этого инструмента по сравнению с традиционными методами документирования, ниже представлены результаты ряда экспериментов. В итоге сформировано собственное видение способа автоматизации процесса балансирования игровой математической модели нетривиальным способом генерации на основе данных, полученных из текста на естественном языке. Выбранный функционал необходим для работы генератора игровых прототипов, представленного в ряде наших ранних работ [2–10].

Основное достижение настоящей работы заключается в том, что на основании проведенного анализа представлены результаты экспериментов, а также планы дальнейшего развития инструмента, призванного решить проблемы оптимизации ресурсов при разработке игровых и интерактивных проектов.

В первом разделе представлен обзор литературы, связанной с определением «игрового баланса». Во втором разделе рассмотрены разработанные инструменты балансировки игр. В третьем разделе описан онлайн-редактор динамического математического баланса компьютерных игр Machinations, а также представлены результаты двух экспериментов, доказывающих его эффективность. В четвертом разделе описано наше видение автоматизации процесса балансирования игры для работы инструмента генерации игрового прототипа и сформулированы перспективы развития. В заключении подведены итоги работы и сделан вывод: такая задача, как балансирование игрового прототипа, может быть автоматизирована, что существенно повысит качество артефактов этапа создания концепции.

ОПРЕДЕЛЕНИЯ ИГРОВОГО БАЛАНСА

Единого мнения о том, что такое игровой баланс, нет, как и практик работы с ним. Однако такой баланс влияет на успех игры прямым образом. От него зависит то, понравится ли игрокам процесс, останутся ли они с проектом надолго и посоветуют ли его друзьям [11].

В обзорной статье [12] 2019 года на основе десятка определений проведен семантический анализ термина «игровой баланс». Рассмотрим некоторые из них в хронологическом порядке, чтобы проследить развитие значения этого понятия.

Начнем с 2005 года. В книге [13] ведущего дизайнера и сценариста компаний Ubisoft и Microsoft Studios Ричарда Роуза III тема игрового баланса иногда поднимается в интервью с игровыми дизайнерами. Из контекста становится понятно, что речь идет о математическом балансе. Суть этого компонента игрового проекта заключается в том, чтобы соотнести значения параметров, которые участвуют в процессе игры, таким образом, чтобы игрокам было не только интересно, но и комфортно.

Кроме этого, на примере дизайна уровней автор приходит к выводу, что интерактив, повествование, продвижение по уровню и другие аспекты игры связаны и должны поддерживать и не перебивать друг друга, *«быть в балансе»* в широком смысле.

В 2010 году сооснователь Global Game Jam и доцент кафедры интерактивных игр и медиа Рочестерского технологического института Ян Шрайбер в блоге «Game balance concepts» утверждал, что почти в каждой игре есть параметры, поддающиеся количественной оценке [14]. В этой работе тема математического баланса игр подробно освещена на примерах. Однако автор добавляет, что числовые значения параметров, которые принимают участие в игре, имеют смысл только *в контексте*, т. е. зависят от текущего повествовательного события. Помимо прочего, автор советует учитывать доступность информации игрокам, их способность обрабатывать информацию, их ожидания и даже внешние факторы. Всё это влияет на восприятие, значит, и на игровой опыт.

Новак в своей книге [15] в 2012 году, как и другие авторы Роллингс и Адамс в их книге [16] в 2003 году, отмечают зависимость игрового баланса от мастерства игроков, но также разделяют **статический** и **динамический балансы**.

Статический баланс включает правила, числа, отношения и взаимодействия, которые возникают в игровом процессе. В свою очередь, динамический баланс иллюстрирует то, как игроки влияют на статический баланс своими действиями в режиме реального времени.

Интересно, что на Youtube-канале «Extra Credits» в видео 2012 года «Perfect imbalance – why unbalanced design creates balanced play» доказано, что легкий игровой дисбаланс мотивирует отдавать предпочтение новым стратегиям поведения, т. е. обучаться за счет ощущения легкого дискомфорта [17].

Основатель Ludeon Studios Тайнан Сильвестр в книге [18] в 2013 озвучил тезис о том, что невозможно создать игру, в которой игроки с разным уровнем мастерства могли бы иметь равный шанс на успех.

Дизайнеры должны ориентироваться на целевую аудиторию, чтобы создать то, с чем она справится. Таким образом, в основе игрового процесса должна быть честная игра. При этом важно, чтобы в игре были доступны разные стратегии, чтобы игроки могли принимать взвешенное решение в пользу каждой из них. Однако любой баланс бесполезен, если он разрушает повествование, плавность и темп игры, доступность и ясность правил.

Профессор игрового дизайна Университета Карнеги-Меллона Джесси Шелл в удостоенной многих наград книге [19] в 2015 году аккумулировал большой объем знаний об игровом дизайне. В его работе можно найти противоречащие друг другу явления вызова и успеха, мастерства и удачи. Под балансированием здесь он подразумевает ориентацию на целевую аудиторию. Формулировка общая и расплывчатая, но она ярко иллюстрирует, что игровой баланс в разных контекстах подразумевает соответствующие вызовы. Подобное часто случается в игровой индустрии, ведь в ней сложно найти два проекта, которые разрабатывались бы одинаково.

Ряд научных статей, которые основаны на мнениях практиков, посвящены теоретическому и практическому изучению и расширению понятия игрового баланса.

Так, в статье [11] 2006 года авторы приходят к выводу, что хороший игровой баланс увлекает игрока, дольше удерживая в игре. В этой работе приведены требования к динамическому балансу, который комфортен игрокам: игра адаптируется к начальному уровню навыков игроков, к развитию мастерства и остается правдоподобной.

В более поздней работе [20] 2018 года адаптивность игрового баланса авторы связывают с эмоциями и утверждают, что игра нравится игрокам, если она вызывает удовольствие, а не скуку или разочарование (казалось бы, очевидный вывод, однако существует немало примеров, которые работают не по так называемой дофаминовой петле – зависимости от удовольствия).

По итогу анализа упомянутых определений можно составить следующую картину: игра сбалансирована, если её сложность зависит от навыков целевой аудитории. Лучше всего, чтобы она адаптировалась под игроков в процессе игры.

Однако важно, чтобы игра бросала вызов, чтобы игроки испытывали чувство достижения. Нельзя рассматривать математический баланс в отрыве от остальных частей игры (интерактива, повествования, продвижения по уровню и др. [13]), всё связано и формирует контекст. Соответственно, все части игры должны находиться в балансе.

Игра должна приносить удовольствие, быть честной, справедливой, не вызывать излишние фрустрацию и скуку. Игра должна предоставлять разные стратегии, а не призывать использовать привычные тактики.

Справедливо упомянуть, что игровой баланс – это также инструмент дизайна, и если разработчики преднамеренно хотят создать состояние дискомфорта, то можно, соответственно, чрезмерно снизить или повысить сложность игры.

В силу перечисленных выше тезисов мы выделим следующие определения математического и игрового балансов.

Математический баланс выражается в параметрах и функциях, в математической модели игры, в экономической системе. По-другому, математический баланс часто называют игровой экономикой.

Игровой баланс подразумевает контекст, от которого зависит математический баланс. Контекстом могут выступать части повествования (сценария), уникальные игровые события, правила здравого смысла, объективное восприятие игровой ситуации дизайнером или игроком.

АВТОМАТИЗАЦИЯ ТЕСТИРОВАНИЯ И БАЛАНСИРОВАНИЯ

Связанные работы

Тестирование игр и, в частности, исполнения правил, заложенных в математический и игровой балансы, желательно проводить в большой аудитории. Однако тестировать игровые системы при помощи автотестов зачастую невозможно, так как каждая из игр представляет обычно уникальный продукт. И всё же способы автома-

тизации тестирования и балансирования игр уже существуют. Рассмотрим ряд работ, которые иллюстрируют работу подобных подходов.

Чтобы найти возможные игровые стратегии, применяют *коэволюционный метод*. Такой метод протестировали в игре «Захват флага» [21].

Для того чтобы оптимизировать тестирование, можно *формализовать* игровой процесс [22]. Однако такой подход локальный, его нельзя масштабировать на прочие игры.

Благодаря генеративному моделированию можно создавать архетипы игроков и далее тестировать игры с помощью *поиска по дереву методом Монте-Карло* синтетическими ИИ[1]-тестировщиками. Такой подход применен для игр Dungeon Crawl [23]. Кроме того, поиск по дереву методом Монте-Карло использован для балансирования стохастических игр [24], а также игр жанра tower defense [25].

Балансирование также возможно с помощью *эволюционных алгоритмов* ИИ. На примере аркадной игр Ms. Pac-Man и RTS StarCraft данный способ показал свою продуктивность [26]. Эволюционный алгоритм многокритериальной оптимизации применен для балансирования карточной игры Top Trumps [27].

Работа над математическим балансом возможна с помощью моделирования взаимосвязи между игровой динамикой и механикой. Оценка динамики игры может быть основана *на методах оператора Купмана* [28].

Предсказывать сложность недетерминированных игр-головоломок «три в ряд» можно с помощью *свёрточных нейронных сетей* [29].

Для работы с математическим балансом пошаговой стратегический игры ComPet был создан *генетический алгоритм* [30].

Для шутера с заданной продолжительностью игры для генерации сбалансированных уровней использованы *эволюционные вычисления* [31].

Интеллектуальные агенты могут быть созданы разными технологиями на основе поведения игроков. В одной из последних работ [32] авторы констатируют, что предшествующие подходы в области автоматизации тестирования игр решают проблему, в основном, эвристическими и обобщенными методами, оптимизируя данные о поведении игроков в виде нереалистичных архетипических

[1] ИИ – искусственный интеллект.

агентов – не удовлетворяя потребности в разнообразном тестировании. В противоположность этому авторы предлагают *подход глубокого моделирования поведения игрока*, Deep Player Behavior Modeling (DPBM), идея которого состоит в том, что у интеллектуальных агентов есть индивидуальность, которая имитирует поведение реальных игроков более достоверно. Обучение происходит на наборе данных игры Aion [33].

Интересно, что в большинстве работ их авторы приходят к выводу, что математический баланс зависит от контекста конкретной игры. Поэтому так интересно проанализировать работы в направлении интерактивного цифрового повествования, в котором контекст первичен. Например, балансировать события, которые происходят в игре, можно на основе биосигналов игрока [34]. Ожидания в таком случае удовлетворяются с помощью *модели эмоций*.

Вывод, который можно сделать на основе анализа существующих подходов, следующий: авторы перечисленных подходов [21–34] чаще склоняются к использованию ИИ и признают, что математический баланс зависит от контекста. Игровой баланс включает в себя и математическую модель игры, и реакции игроков, которые в свою очередь влияют на эту модель.

Machinations для создания математического баланса

На данный момент единственной платформой для разработки и прогнозирования игровых экономик и систем для премиума, free2play и play2earn игр является инструмент Machinations [1]. Он разработан группой игровых разработчиков с большим опытом работы и предназначен для создания и воспроизведения динамического математического баланса игры. Инструмент позволяет проектировать, балансировать и моделировать игровую экономику в виде динамических диаграмм, которые можно воспроизвести, чтобы посмотреть изменения системы в режиме реального времени.

Принцип работы Machinations необходимо подкрепить конкретным примером. Диаграмма [35] иллюстрирует поведение двух NPC из игры The Elder Scrolls V: Skyrim [36]. Первый – из самого маленького поселения Винтерхол, второй – из Виндхельм, самого большого и густонаселенного. В названной работе сопостав-

лены два способа работы: документирование требований к поведению NPC[2] и создание диаграмм в Machinations. Один из примеров диаграммы иллюстрирует поведение NPC, действия которого не зависят от времени суток.

На рис. 1 схематично изображено следующее: каждый день длится 1440 минут, именно столько ресурса поступает из доступного источника (истока) в накопитель (пул) в начале дня. Каждую минуту уничтожается одна единица ресурса времени. Ввиду того, что любое число больше нуля, единственный ресурс во втором пуле будет постоянно пребывать в одном состоянии, и поведение NPC будет постоянным в течение дня.

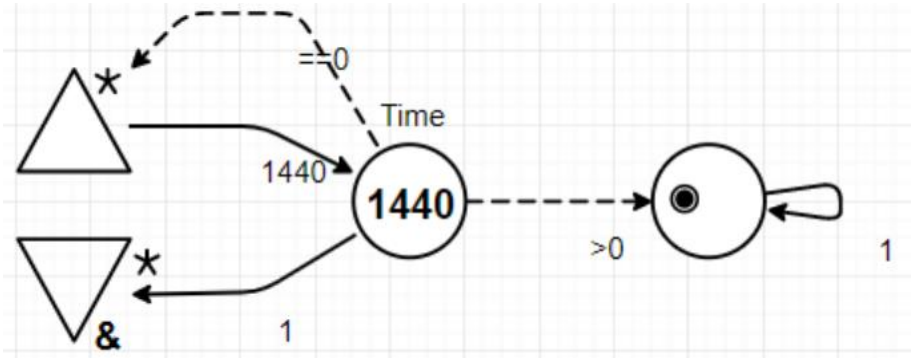


Рис. 1. Пример простой диаграммы баланса

Во фрагменте сложного поведения NPC показано, как ресурс изо дня в день движется по одним и тем же шагам сценария всегда в один и тот же период времени (рис. 2), что иллюстрирует одинаковое поведение NPC каждый день.

После оценки времени, затраченного на сборку диаграммы в редакторе Machinations и при ручном документировании игровых балансов, были сделаны выводы об эффективности использования диаграмм. Также диаграммы позволяют отказаться от объемных текстовых документов, предлагая максимальную репрезентативность.

Machinations для автоматизации математического баланса

Другой эксперимент показал, что Machination показывает хорошие результаты для автоматизации балансирования существующей игровой экономики [37].

[2] NPC (сокр. англ. Non-Player Controller) – неигровой персонаж в игре, который имеет свою логику поведения.

Для этого была создана диаграмма конкретной игровой механики – кровотечения, которое персонаж получает в результате ранения во время боя в настольной ролевой игре Dungeons and Dragons [38].

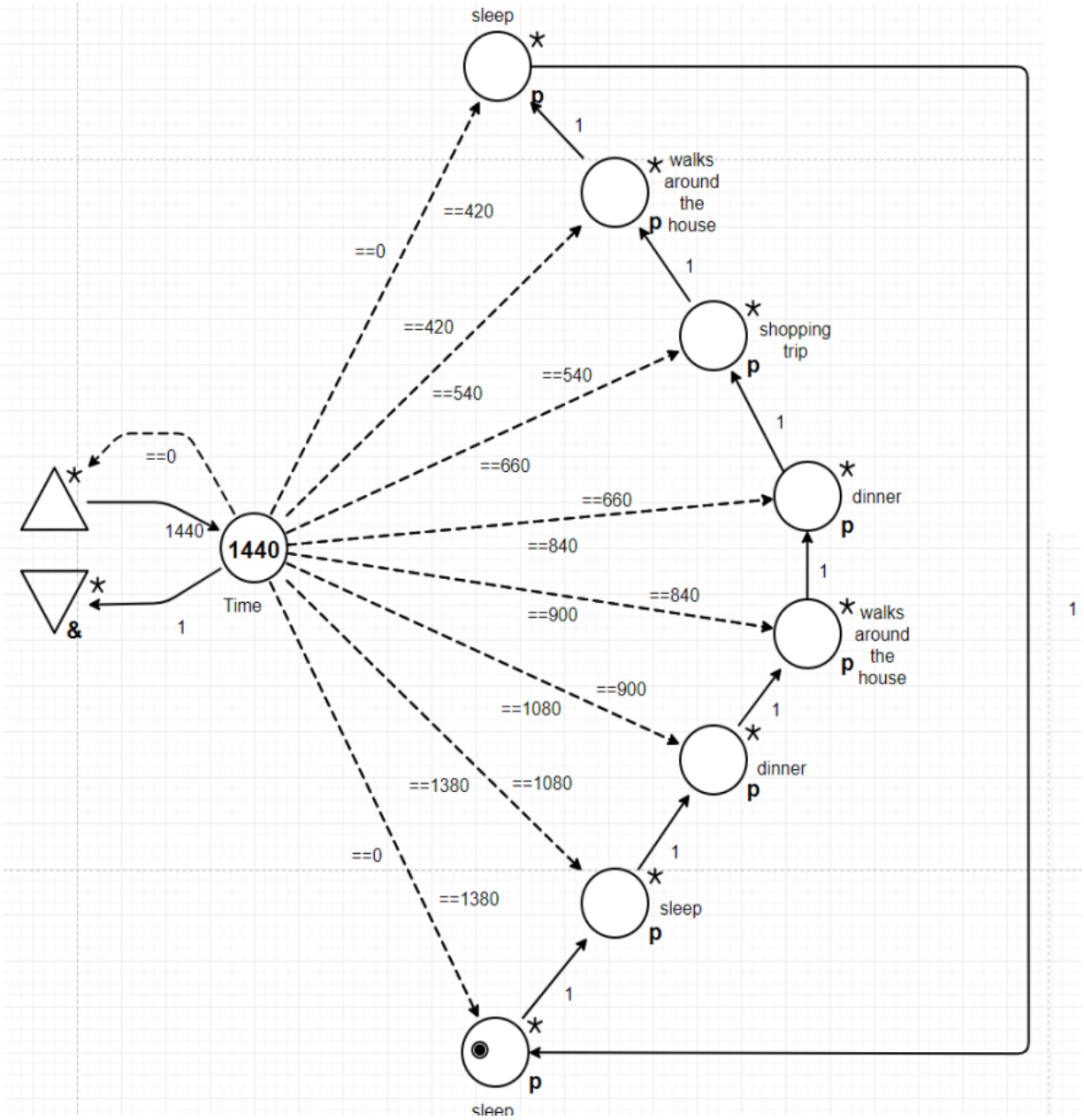


Рис. 2. Пример сложной диаграммы баланса

На рис. 3 схематично изображено следующее: урон от кровотечения, шанс ослабить или остановить кровотечение, урон от атак противника и шанс попадания атаки противника.

Сбалансированность диаграммы оценивается по количеству ходов боя. Построенная диаграмма была автоматически сбалансирована до желаемого состояния в результате многократной симуляции игрового процесса и корректировки.

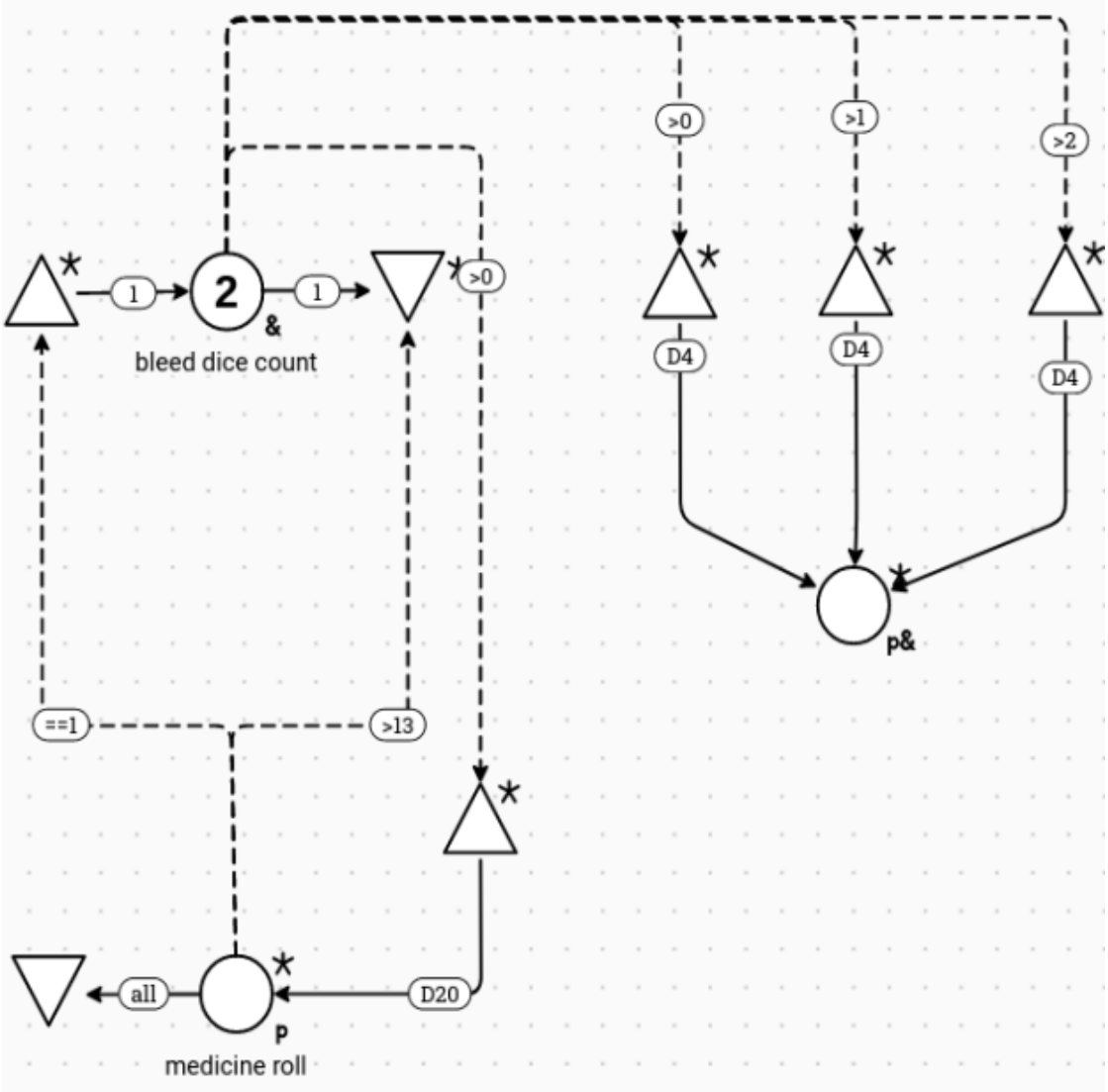


Рис. 3. Диаграмма баланса механики кровотечения

Пул «medicine roll» проверяет медицинские навыки персонажа с кровотечением. Если проверка покажет наличие таких навыков, то кровотечение уменьшится. В противном случае кровотечение усилится.

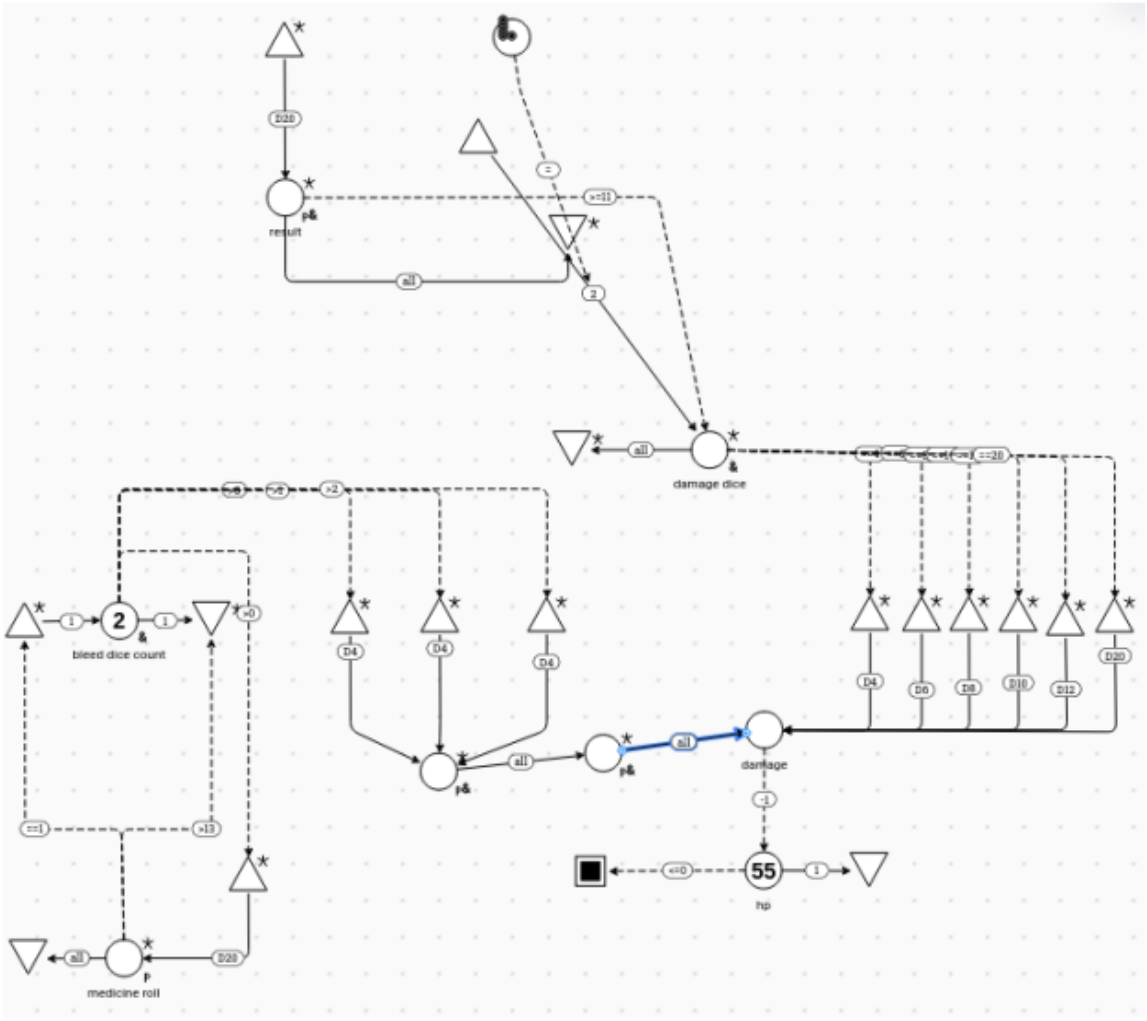


Рис. 4. Механика кровотечения в общей диаграмме

Пул «bleed dice count» отвечает за то, сколько конкретно крови потеряет игрок. Данный компонент отвечает за то, чтобы передавать информацию о том, сколько урона получает игрок от кровотечения.

В Machinations можно отправить необходимые параметры, которые будут корректировать диаграмму и составлять баланс игры (рис. 4). Чтобы подтвердить качество сгенерированных значений, алгоритм проверяет полученные данные и изменяет их с определенным шагом до тех пор, пока значения не удовлетворяют правилам сбалансированности. Так, в описанном эксперименте в результате имитаций 40 боевых ситуаций средняя длительность боев составила семь шагов.

Автоматическая генерация и проверка баланса составили чуть более трех часов. В то время, как ручная проверка может затянуться на дни и недели.

Эксперимент позволяет прийти к выводу, что автоматическое балансирование значительно сокращает время на разработку и повышает качество результата.

ИНТЕГРАЦИЯ ИНСТРУМЕНТА БАЛАНСИРОВАНИЯ В ГЕНЕРАТОР ИГРОВЫХ ПРОТОТИПОВ

Два эксперимента, описанных выше, показали, что динамические диаграммы математического баланса эффективны, сокращают время разработки и не требуют глубоких технических знаний и навыков.

В ходе разработки инструмента генерации игровых прототипов, описание которого представлено в ряде работ [2, 4, 6, 9], были получены подходы для автоматизации создания контента на основе данных, взятых из текстовой документации. Практический опыт работы над балансом игры показывает, что документирование математических функций, составляющих игровую экономику, происходит так же, как и предъявление требований для визуализации или кода – через текст. Более того, связь с контекстом формирует зависимость числовых значений. Например, из предложения «чем персонаж ближе к смерти, тем сильнее его удары» можно извлечь информацию о следующей зависимости: «чем меньше у персонажа игрока здоровья, тем больший урон наносят его удары по вражеским персонажам».

Чтобы рассмотреть более сложный пример, понадобится готовая диаграмма Machinations, признанная в открытом сообществе ресурса (комьюнити Machinations) как корректная [39]. Суть работы этой диаграммы в том, что из определенных ресурсов можно строить здания и другие игровые объекты. Всего в диаграмме пять рецептов. Более того, есть зависимость между ними: какие-то из них недоступны до тех пор, пока не будет использован предыдущий.

В случае, если данная диаграмма была бы одним из выходных файлов работы генератора игровых прототипов, входной текст на естественном языке, который описывает суть в художественной форме, мог бы быть следующим: «Игрок стоит у реки. Рядом с ним склад с досками. Игрок строит из досок мост и переходит по нему через реку. Игрок собирает камни и строит из них дом. Затем он пристраивает к дому террасу из досок и камней. После этого игрок создает арку для выхода из террасы из досок и камня. Из остатков досок игрок собирает скамейку».

На рисунке ресурсы и рецепты, которые упоминаются в тексте, подписаны соответствующим образом (рис. 5).

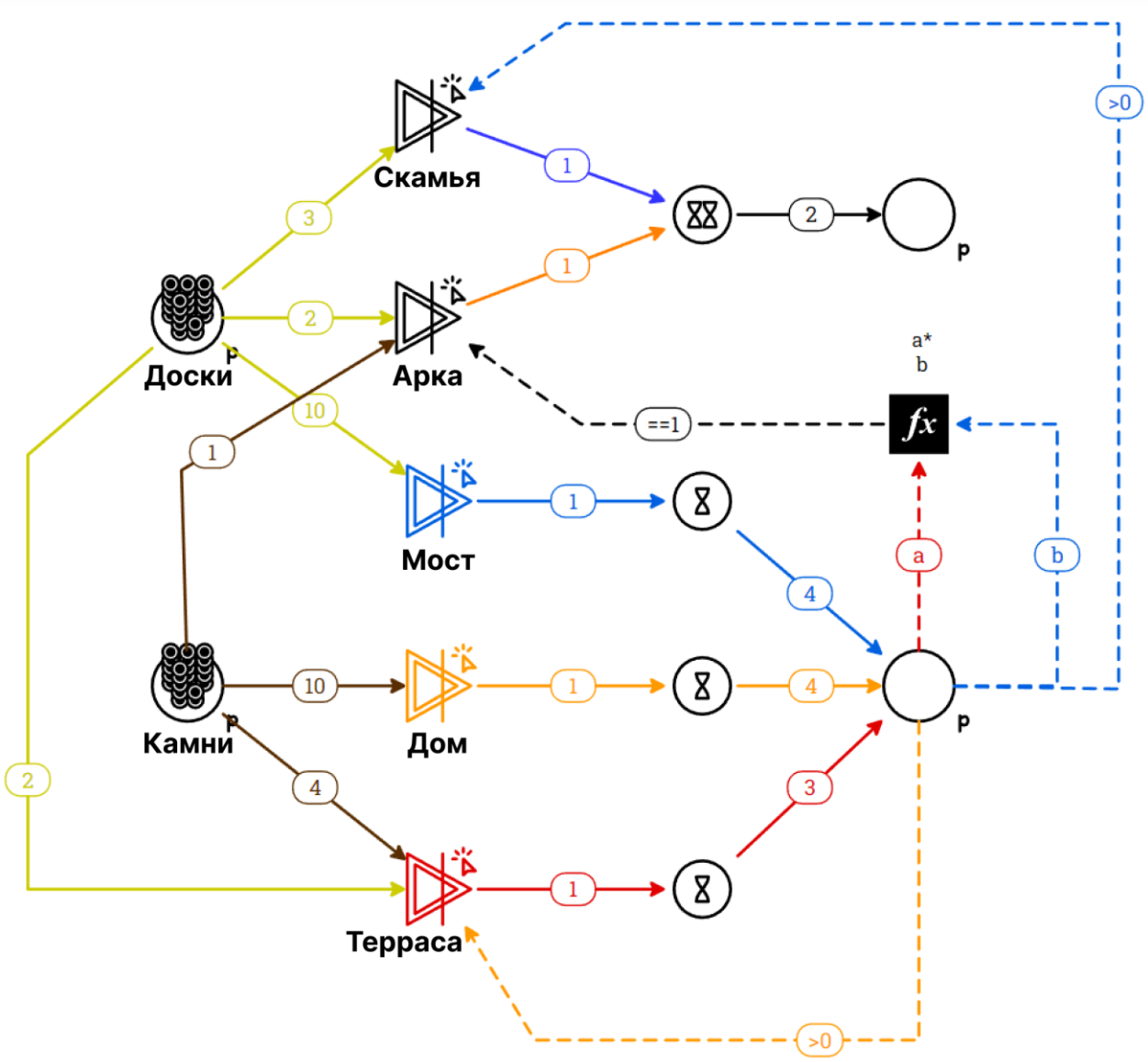


Рис. 5. Диаграмма ресурсов и рецептов

Проанализируем обновленную структуру инструмента генерации сценарного прототипа с точки зрения игрового баланса (рис. 6).

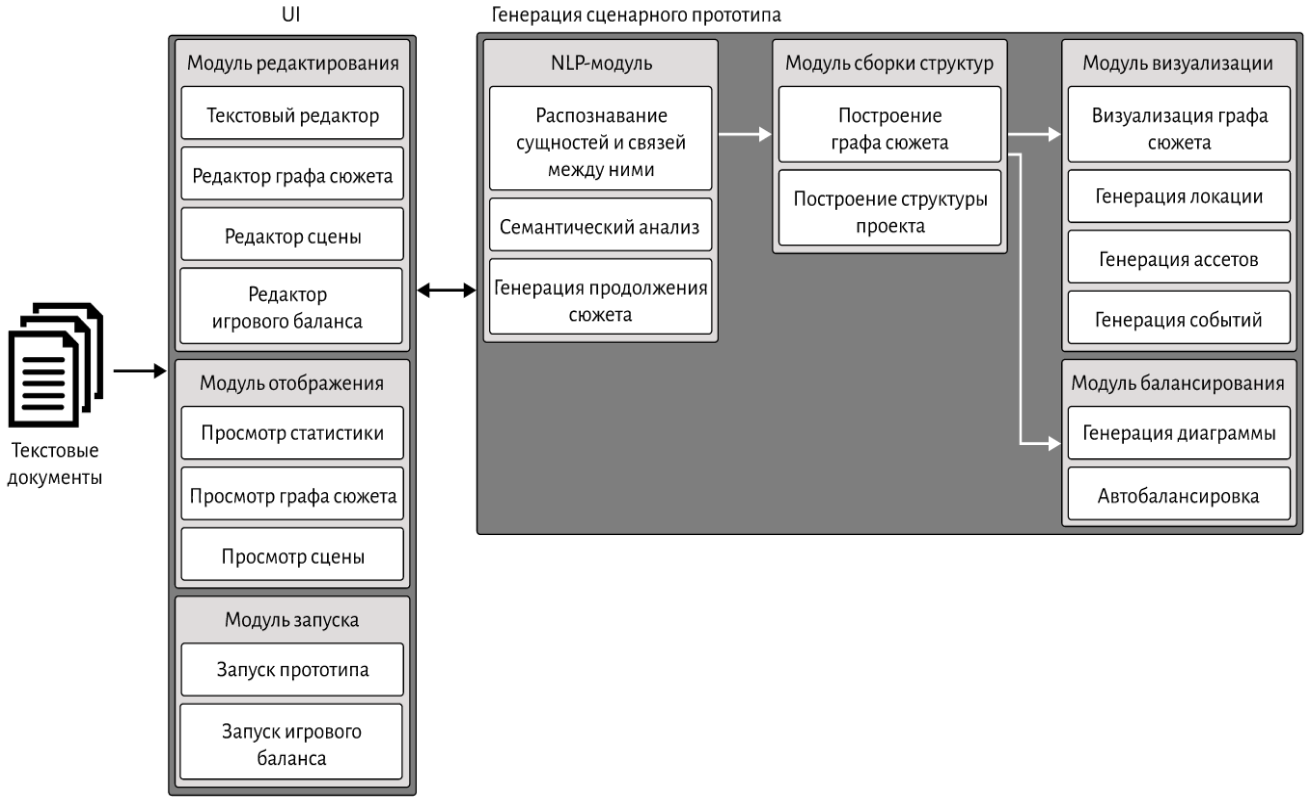


Рис. 6. Структура инструмента генерации сценарного прототипа

Инструмент принимает на вход набор текстовых документов. Первый этап, который проходит тест, это распознавание сущностей и связей между ними. В более ранних работах выделялись локации, персонажи и артефакты сценария, всевозможные предметы, без которых повествование невозможно [2, 40, 41]. Данная задача решается классическими методами NLP[3].

Однако возможно расширение анализируемых параметров. Например, благодаря использованию таких характерных эпитетов, как «мощная атака», «быстрый удар», «уворот» и т. п., можно формализовать параметры усилений и способностей, доступных игроку в текущем контексте событий.

Таким образом, предполагается, что в документации содержится вся информация, необходимая для анализа. Перевод этой информации на язык параметров и функций значительно облегчает программирование и реализацию игры.

[3] NLP – Natural Language Processing, обработка естественного языка.

На первых порах решение о присвоении той или иной схемы игрового баланса для текущего контекста может решаться присвоением наиболее подходящего шаблона из готового набора. В дальнейшем данный этап может быть полностью автоматизирован.

Параметры, извлеченные для игрового баланса можно, заполнить автоматически в случайном порядке, таких данных достаточно для тестирования. В дальнейшем, в результате исследования большого количества существующих систем возможно более осмысленное присвоение значений. В данном вопросе уместно применение алгоритмов машинного обучения.

И, наконец, полученную модель игрового баланса можно воплотить в виде логики и кода, использовав возможности визуального программирования и готовых библиотек.

При всей амбициозности идеи генерации игр из текста необходимо помнить, что честный взгляд на автоматическую генерацию цифрового контекста – это, в первую очередь, лояльность по отношению к артефактам генерации и неточностям. Задача автоматизации может не решать проблему производства качественного контента, но может ускорять этап прототипирования, что может позволить задействовать минимальное количество специалистов.

Необходимость в делегировании рутинной работы имеется в любых компаниях производства интерактивных проектов. Данный факт подтверждает опыт работы авторов статьи. Любой творческий проект затормаживает этап, когда существует неопределенность в том, какое решение для реализации задачи будет правильным, выигрышным.

Попытка манипулировать параметрами уже была предпринята в опубликованных работах [9, 40 – 42]. Успех в реализации малых задач говорит о том, что в дальнейшем локальные задачи можно будет объединить под эгидой объёмного многокомпонентного инструмента, способного заменить, а где-то и предложить новый способ решения задач разработки и прототипирования.

ЗАКЛЮЧЕНИЕ

Определение игрового баланса многогранно. Примечательно, что математический баланс не должен существовать вне контекста, должен зависеть от сценария и игрового дизайна. Зависимость числовых значений экономики игры и настроения повествования позволяет игрокам получать комфортный и интересный опыт.

На сегодняшний день существует множество подходов к работе над балансом игры. Большая часть из них использует технологии ИИ. Подходы балансирования направлены как непосредственно на процессы разработки игры, так и на процессы создания игрового повествования.

Ряд экспериментов показал, что эффективным способом балансирования является работа с инструментом Machinations. Его потенциал можно использовать для автоматизации тестирования и корректирования математического баланса (игровой экономики). Кроме того, его применение в конвейере генерации игровых прототипов делает его неотъемлемой частью оптимизации этапа проектирования игры.

На основе анализа актуальных источников получено представление о том, как интегрировать функцию автоматического балансирования в инструмент генерации игровых прототипов. Представлена обновленная структура работы инструмента.

Следующие четыре задачи на будущее формируют цель разработки функции автоматической генерации игрового баланса на основе текста на естественном языке:

1. создание адаптированных под разработку интерактивных проектов алгоритмов обработки игровой документации для извлечения параметров и составных математических функций из текста на естественном языке;
2. обработка извлеченных данных для генерации игровой экономики, зависящей от повествовательного контекста;
3. создание алгоритмов автоматического балансирования игровой модели на основе интеллектуального тестирования;

4. имплементация полученного игрового баланса в виде игровой логики в игровые проекты, которые получаются на выходе из инструмента генерации игровых прототипов.

Необходимо также отметить, что, будучи законченным, инструмент автоматической балансировки игр в разы сократит ресурсы на их разработку, в особенности для независимых разработчиков, а также повысит качество цифрового контента.

Благодарности

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»).

СПИСОК ЛИТЕРАТУРЫ

1. Machinations. URL: https://machinations.io/.

2. *Сахибгареева Г.Ф., Кугуракова В.В.* Концепт инструмента автоматического создания сценарного прототипа компьютерной игры // Электронные библиотеки. 2018. Т. 21. № 3–4. С. 235–249.

3. *Сахибгареева Г.Ф., Бедрин О.А., Кугуракова В.В.* Разработка компонента генерации визуализации сценарного прототипа видеоигр // Труды XXII Всероссийской научной конференции «Научный сервис в сети Интернет». ИПМ им. М.В. Келдыша. 2020. С. 581–603. https://doi.org/10.20948/abrau-2020-4.

4. *Сахибгареева Г.Ф., Бедрин О.А., Кугуракова В.В.* Раскадровка как одно из представлений сценарного прототипа компьютерных игр // Электронные библиотеки. 2021. Т. 24. № 2. С. 408–444.

5. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Visualization Component for the Scenario Prototype Generator as a Video Game Development Tool // Proceedings of the 22nd Conference on Scientific Services & Internet. CEUR Workshop Proceedings. 2020. V. 2784. P. 267–282.

6. *Кугуракова В.В., Сахибгареева Г.Ф., Нгуен А.З., Астафьев А.М.* Пространственная ориентация объектов на основе обработки текстов на естественном языке для генерации раскадровок // Электронные библиотеки. 2020. Т. 23. № 6. С. 1213–1238.

7. *Сахибгареева Г.Ф.* Применимость разветвленных структур для генерации сценарных прототипов видеоигр // 65-я Международная научная конференция Астраханского государственного технического университета. 2021. С. 596–600.

8. *Сахибгареева Г.Ф., Кугуракова В.В.* Прототипирование вариативности сюжета компьютерных игр // Труды XXIII Всероссийской научной конференции «Научный сервис в сети Интернет». ИПМ им. М.В. Келдыша. 2021. С. 347–360. https://doi.org/10.20948/abrau-2021-11.

9. *Сахибгареева Г. Ф., Кугуракова В. В.* Редактор интерактивной структуры для инструмента генерации сценарных прототипов // Электронные библиотеки. 2022. Т. 24. № 6. С. 1184–1202.

10. *Sahibgareeva G.F., Kugurakova V.V.* Branched Structure Component for a Video Game Scenario Prototype Generator // Proceedings of the 23nd Conference on Scientific Services & Internet. CEUR Workshop Proceedings, 2021. V. 3066. P. 101–111. https://doi.org/10.20948/abrau-2021-10-ceur.

11. *Andrade G., Ramalho G., Gomes A.S., Corruble V.* Dynamic game balancing: An evaluation of user satisfaction // Proceedings of the 2nd Conference on Artificial Intelligence and Interactive Digital Entertainment. AAAI Digital Library. 2006. V. 2. No. 1. P. 3–8.

12. *Becker A., Görlich D.* Game balancing — A semantical analysis // First International Workshop on Video Games, Gamification and Educational Innovation. CEUR Workshop Proceedings. 2019. V. 2486. P. 344–359.

13. *Rouse R.* Game design: Theory and practice // Plano: Jones & Bartlett Learning. 2005. 704 p.

14. Game balance concepts.
URL: http://gamebalanceconcepts.wordpress.com/.

15. *Novak J.* Game development essentials: an introduction // Cengage Learning. 2012. 510 p.

16. *Rollings A., Adams E.* Andrew Rollings and Ernest Adams on game design. Indianapolis: New Riders Publishing. 2003. 621 p.

17. Perfect imbalance — why unbalanced design creates balanced play.
URL: https://youtu.be/e31OSVZF77w.

18. *Sylvester T.* Designing games: A guide to engineering experiences. Sebastopol: O'Reilly Media. 2013. 413 p.

19. *Schell J.* The Art of Game Design: A Book of Lenses. Boca Raton: A K Peters/CRC Press. 2015. 600 p.

20. *Tijs T.J.V, Brokken D., IJsselsteijn W.A.* Dynamic game balancing by recognizing affect // Second International Conference «Fun and Games». Springer-Verlag GmbH. 2008. V. 5294. P. 88–93.

21. *Leigh R., Schonfeld J., Louis S.J.* Using coevolution to understand and validate game balance in continuous games // 10th annual conference on Genetic and Evolutionary Computation. ACM. 2008. P. 1563–1570.
https://doi.org/10.1145/1389095.1389394.

22. *Volz V., Rudolph G., Naujoks B.* Demonstrating the feasibility of automatic game balancing // Genetic and Evolutionary Computation Conference. Association for Computing Machinery. 2016. P. 269–276.

23. *Holmgard C., Green M., Liapis A., Togelius J.* Automated playtesting with procedural personas through MCTS with evolved heuristics // IEEE Transactions on Games. 2018. V. 11. No. 4. P. 352–362.

24. *Keehl O., Smith A.M.* Monster carlo 2: Integrating learning and tree search for machine playtesting // IEEE Conference on Games. 2019. P. 1–8.

25. *Beau P., Bakkes S.* Automated game balancing of asymmetric video games // IEEE Conference on Computational Intelligence and Games. IEEE. 2016. P. 333–340.

26. *Moroşan M., Poli R.* Automated Game Balancing in Ms PacMan and StarCraft Using Evolutionary Algorithms // 20th European Conference on the Applications of Evolutionary Computation. Springer, 2017. V. 10199. P. 377–392.

27. *Volz V., Rudolph G., Naujoks B.* Demonstrating the feasibility of automatic game balancing // Genetic and Evolutionary Computation Conference. ACM. 2016. P. 269–276.

28. *Avila A.M., Fonoberova M., Hespanha J.P., Mezic I., Clymer D., Goldstein J., Pravia M.A., Javorsek D.* Game Balancing using Koopman-based Learning // American Control Conference. IEEE. 2021. P. 710.

29. *Gudmundsson S.F., Eisen P., Poromaa E., Nodet A., Purmonen S., Kozakowski B., Meurling R., Cao L.* Human-like playtesting with deep learning // IEEE Conference on Computational Intelligence and Games. IEEE. 2018. P. 1–8.

30. *Morosan M., Poli P.* Lessons from Testing an Evolutionary Automated Game Balancer in Industry // Games, Entertainment, Media Conference. IEEE. 2018. P. 263–270.

31. *Karavolos D., Liapis A., Yannakakis G.N.* Using a Surrogate Model of Gameplay for Automated Level Design // IEEE Conference on Computational Intelligence and Games. IEEE. 2018. P. 1–8.

32. *Pfau J., Liapis A., Yannakakis G.N., Malaka R.* Dungeons & Replicants II: Automated Game Balancing Across Multiple Difficulty Dimensions via Deep Player Behavior Modeling // IEEE Transactions on Games. 2022. P. 1–11.

33. The Tower of Aion. URL: https://www.ncsoft.jp/aion/.

34. *Dworak W., Filgueiras E., Valente J.* Automatic Emotional Balancing in Game Design: Use of Emotional Response to Increase Player Immersion // 9th International Conference on Design, User Experience, and Usability. Springer. 2020. V. 12201. P. 426–438.

35. *Черечукина А.Н.* Содержание GDD как требований к разработке программного обеспечения // Казанский федеральный университет. 2019. 47 с. URL: https://kpfu.ru/student_diplom/10.160.178.20_236517_F_Cherechukina_1_.pdf.

36. The Elder Scrolls V: Skyrim.
URL: https://elderscrolls.bethesda.net/ru/skyrim.

37. *Галимзянов Г.Р.* Разработка инструмента автоматической корректировки внутриигровых параметров // Казанский федеральный университет. 2021. 35 с. URL: https://kpfu.ru/student_diplom/10.160.178.20_TXB9250VCS6S6OVSLZOCXQDP4J7WFCRV__J7FXN80EEZNIXS6Q_Galimzyanov.pdf.

38. Dungeons & Dragons. URL: https://dnd.wizards.com/.

39. Machination. URL: https://machinations.io/templates/book/figure-6-47-rts-building-mechanics-game-mechanics-advanced-game-design-book/.

40. *Доброквашина А.С., Газизова Э.А.* Автоматизация проектирования игрового прототипа на основании обработки формализованного игрового дизайн-

документа // Ученые записки Института социальных и гуманитарных знаний, 2019. Т. 17. № 1. С. 583–589.

41. *Вакатов С.А.* Разработка инструмента вариативности сюжета с запуском прототипа в виде текстовой игры // Казанский федеральный университет. 2021. 36 с. URL: https://kpfu.ru/student_diplom/10.160.178.20_TTKKD9XW59RG5L7TVLTB73YPTISE59Y16W5D1U435WOXWI10US_Vakatov.pdf.

42. *Вакатова Э.С.* Разработка функционала генерации продолжения сюжета для инструмента прототипирования сюжета в компьютерных играх // Казанский федеральный университет. 2021. 34 с. URL: https://kpfu.ru/student_diplom/10.160.178.20_PQK51KDGAPZ5Z82IKYY69MV84PCLTPERV0NNYJ33B7P5T7NJFP_F_Vakatova.pdf.

GAME BALANCE TOOLS

G. F. Sahibgareeva[1] [0000-0003-4673-3253], V. V. Kugurakova[2] [0000-0002-1552-4910], E. S. Bolshakov[3] [0000-0002-2208-9515]

[1, 2, 3]*Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[1]gulnara.sahibgareeva42@gmail.com, [2]vlada.kugurakova@gmail.com, [3]edward.bolshakov117@gmail.com

Abstract

To disclose the concept of game balance and to develop an approach to automate the routine when working with game economics, methods of data analysis and experimentation were applied. According to the results of the analysis of existing definitions, a special case of mathematical balance and a generalized design definition of game balance were singled out. By parsing the existing approaches for balancing and testing games, a vision of our own solution was developed. Based on the findings, an approach for automating balance within a game prototype generator has been proposed. As an intermediate result, an updated structure and operation procedure of the game prototype generation tool were presented. The prospects for further development in this direction are given.

***Keywords**: game balance, mathematical balance, automatic game balancing, Machinations, game prototypes generation*

REFERENCES

1. Machinations. URL: https://machinations.io/.

2. *Sahibgareeva G.F., Kugurakova V.V.* Koncept instrumenta avtomatich-eskogo sozdaniya scenarnogo prototipa komp'yuternoj igry // Russian Digital Library Journal. 2018. V. 21. No 3–4. P. 235–249.

3. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Razrabotka komponenta generacii vizualizacii scenarnogo prototipa videoigr // Trudy XXII Vserossijskoj nauch-noj konferencii «Nauchnyj servis v seti Internet». IPM im. M.V. Keldysha. 2020. S. 581–603. https://doi.org/10.20948/abrau-2020-4.

4. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Raskadrovka kak odno iz predstavlenij scenarnogo prototipa komp'yuternyh igr // Russian Digital Library Jour-nal. 2021. V. 24. No 2. P. 408–444.

5. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Visualization Component for the Scenario Prototype Generator as a Video Game Development Tool // Proceed-ings of the 22nd Conference on Scientific Services & Internet. CEUR Workshop Proceed-ings. 2020. V. 2784. P. 267–282.

6. *Kugurakova V.V., Sahibgareeva G.F., Nguen A.Z., Astaf'ev A.M.* Pros-transtvennaya orientaciya ob’ektov na osnove obrabotki tekstov na estestvennom yazyke dlya generacii raskadrovok // Russian Digital Library Journal. 2020. V. 23. No 6. P. 1213–1238.

7. *Sahibgareeva G.F.* Primenimost' razvetvlennyh struktur dlya generacii sce-narnyh prototipov videoigr // 65-ya Mezhdunarodnaya nauchnaya konferenciya Astra-hanskogo gosudarstvennogo tekhnicheskogo universiteta. 2021. S. 596–600.

8. *Sahibgareeva G.F., Kugurakova V.V.* Prototipirovanie variativnosti syu-zheta komp'yuternyh igr // Trudy XXIII Vserossijskoj nauchnoj konferencii «Nauchnyj servis v seti Internet». IPM im. M.V. Keldysha. 2021. S. 347–360. https://doi.org/10.20948/abrau-2021-11.

9. *Sahibgareeva G.F., Kugurakova V.V.* Redaktor interaktivnoj struktury dlya instrumenta generacii scenarnyh prototipov // Russian Digital Library Journal. 2022. V. 24. No 6. P. 1184–1202.

10. *Sahibgareeva G.F., Kugurakova V.V.* Branched Structure Component for a Video Game Scenario Prototype Generator // Proceedings of the 23nd Conference on Scientific Services & Internet. CEUR Workshop Proceedings, 2021. V. 3066. P. 101–111.

11. *Andrade G., Ramalho G., Gomes A.S., Corruble V.* Dynamic game balancing: An evaluation of user satisfaction // Proceedings of the 2nd Conference on Artificial Intelligence and Interactive Digital Entertainment. AAAI Digital Library. 2006. V. 2. No. 1. P. 3–8.

12. *Becker A., Görlich D.* Game balancing — A semantical analysis // First International Workshop on Video Games, Gamification and Educational Innovation. CEUR Workshop Proceedings. 2019. V. 2486. P. 344–359.

13. *Rouse R.* Game design: Theory and practice. Plano: Jones & Bartlett Learning. 2005. 704 p.

14. Game balance concepts.
URL: http://gamebalanceconcepts.wordpress.com/.

15. *Novak J.* Game development essentials: an introduction // Cengage Learning. 2012. 510 p.

16. *Rollings A., Adams E.* Andrew Rollings and Ernest Adams on game design. Indianapolis: New Riders Publishing. 2003. 621 p.

17. Perfect imbalance — why unbalanced design creates balanced play. URL: https://youtu.be/e31OSVZF77w.

18. *Sylvester T.* Designing games: A guide to engineering experiences. Sebastopol: O'Reilly Media. 2013. 413 p.

19. *Schell J.* The Art of Game Design: A Book of Lenses. Boca Raton: A K Peters/CRC Press. 2015. 600 p.

20. *Tijs T.J.V., Brokken D., IJsselsteijn W.A.* Dynamic game balancing by recognizing affect // Second International Conference «Fun and Games». Springer-Verlag GmbH. 2008. V. 5294. P. 88–93.

21. *Leigh R., Schonfeld J., Louis S.J.* Using coevolution to understand and validate game balance in continuous games // 10th annual conference on Genetic and Evolutionary Computation. ACM. 2008. P. 1563–1570.

22. *Volz V., Rudolph G., Naujoks B.* Demonstrating the feasibility of automatic game balancing // Genetic and Evolutionary Computation Conference. Association for Computing Machinery. 2016. P. 269–276.

23. *Holmgard C., Green M., Liapis A., Togelius J.* Automated playtesting with procedural personas through MCTS with evolved heuristics // IEEE Transactions on Games. 2018. V. 11. No. 4. P. 352–362.

24. *Keehl O., Smith A.M.* Monster carlo 2: Integrating learning and tree search for machine playtesting // IEEE Conference on Games. 2019. P. 1–8.

25. *Beau P., Bakkes S.* Automated game balancing of asymmetric video games // IEEE Conference on Computational Intelligence and Games. IEEE. 2016. P. 333–340.

26. *Moroşan M., Poli R.* Automated Game Balancing in Ms PacMan and StarCraft Using Evolutionary Algorithms // 20th European Conference on the Applications of Evolutionary Computation. Springer, 2017. V. 10199. P. 377–392.

27. *Volz V., Rudolph G., Naujoks B.* Demonstrating the feasibility of automatic game balancing // Genetic and Evolutionary Computation Conference. ACM. 2016. P. 269–276.

28. *Avila A.M., Fonoberova M., Hespanha J.P., Mezic I., Clymer D., Goldstein J., Pravia M.A., Javorsek D.* Game Balancing using Koopman-based Learning // American Control Conference. IEEE. 2021. P. 710.

29. *Gudmundsson S.F., Eisen P., Poromaa E., Nodet A., Purmonen S., Kozakowski B., Meurling R., Cao L.* Human-like playtesting with deep learning // IEEE Conference on Computational Intelligence and Games. IEEE. 2018.

30. *Morosan M., Poli P.* Lessons from Testing an Evolutionary Automated Game Balancer in Industry // Games, Entertainment, Media Conference. IEEE. 2018. P. 263–270.

31. *Karavolos D., Liapis A., Yannakakis G.N.* Using a Surrogate Model of Gameplay for Automated Level Design // IEEE Conference on Computational Intelligence and Games. IEEE. 2018. P. 1–8.

32. *Pfau J., Liapis A., Yannakakis G.N., Malaka R.* Dungeons & Replicants II: Automated Game Balancing Across Multiple Difficulty Dimensions via Deep Player Behavior Modeling // IEEE Transactions on Games. 2022. P. 1–8.

33. The Tower of Aion. URL: https://www.ncsoft.jp/aion/.

34. *Dworak W., Filgueiras E., Valente J.* Automatic Emotional Balancing in Game Design: Use of Emotional Response to Increase Player Immersion // 9th International Conference on Design, User Experience, and Usability. Springer. 2020. V. 12201. P. 426–438.

35. *Cherechukina A.N.* Soderzhanie GDD kak trebovaniĭ k razrabotke programmnogo obespecheniya // Kazanskij federal'nyj universitet. 2019. 47 s. URL: https://kpfu.ru/student_diplom/10.160.178.20_236517_F_Cherechukina_1_.pdf.

36. The Elder Scrolls V: Skyrim.
URL: https://elderscrolls.bethesda.net/ru/skyrim.

37. *Galimzyanov G.R.* Razrabotka instrumenta avtomaticheskoĭ korrektirovki vnutriigrovyh parametrov // Kazanskij federal'nyj universitet. 2021. 35 s. URL: https://kpfu.ru/student_diplom/10.160.178.20_TXB9250VCS6S6OVSL-ZOCXQDP4J7WFCRV__J7FXN80EEZNIXS6Q_Galimzyanov.pdf.

38. Dungeons & Dragons. URL: https://dnd.wizards.com/.

39. Machination. URL: https://machinations.io/templates/book/figure-6-47-rts-building-mechanics-game-mechanics-advanced-game-design-book/.

40. *Dobrokvashina A.S., Gazizova E.A.* Avtomatizaciya proektirovaniya igrovogo prototipa na osnovanii obrabotki formalizovannogo igrovogo dizajn-dokumenta // Uchenye zapiski Instituta social'nyh i gumanitarnyh znanij, 2019. T. 17. № 1. S. 583–589.

41. *Vakatov S.A.* Razrabotka instrumenta variativnosti syuzheta s zapuskom prototipa v vide tekstovoj igry // Kazanskij federal'nyj universitet. 2021. 36 s. URL: https://kpfu.ru/student_diplom/10.160.178.20_TTKKD9XW59RG5L7TVLTB73YPTISE59Y16W5D1U435WOXWI10US_Vakatov.pdf.

42. *Vakatova E.S.* Razrabotka funkcionala generacii prodolzheniya syuzheta dlya instrumenta prototipirovaniya syuzheta v komp'yuternyh igrah // Kazanskij federal'nyj universitet. 2021. 34 s. URL: https://kpfu.ru/student_diplom/10.160.178.20_PQK51KDGAPZ5Z82IKYY69MV84PCLTPERV0NNYJ33B7P5T7NJFP_F_Vakatova.pdf.

СВЕДЕНИЯ ОБ АВТОРАХ

САХИБГАРЕЕВА Гульнара Фаритовна – старший преподаватель кафедры программной инженерии Института ИТИС КФУ. Сфера научных интересов – игровая сценаристика, нарративный дизайн, изучение вопроса эффективности создания игрового прототипа и возможности автоматизации данного процесса.

Gulnara Faritovna SAHIBGAREEVA – assistant of the Department of Software Engineering of the Institute ITIS KFU. Research interests - game scripting, narrative design, studying the issue of the effectiveness of creating a scenario prototype and the possibility of automating this process.

email: gulnara.sahibgareeva42@gmail.com

ORCID: 0000-0003-4673-3253

КУГУРАКОВА Влада Владимировна – к. т. н., доцент кафедры программной инженерии Института ИТИС КФУ, руководитель НИЛ разработки интеллектуальных инструментов для компьютерных игр. Сфера научных интересов – иммерсивность виртуальных сред, различные аспекты проектирования игр, AR/VR.

Vlada Vladimirovna KUGURAKOVA, PhD., Docent of the Institute ITIS KFU, Head of laboratory of intelligent tools design for computer games development. Research interests include immersiveness of virtual environments, problems of generating realistic visualization, various aspects of game design, AR/VR.

email: vlada.kugurakova@gmail.com

ORCID: 0000-0002-1552-4910

БОЛЬШАКОВ Эдуард Сергеевич – лаборант-исследователь НИЛ разработки интеллектуальных инструментов для компьютерных игр. Сфера научных интересов – игровая сценаристика, игровой дизайн.

Eduard Sergeevich BOLSHAKOV – laboratory researcher at the laboratory of intelligent tools design for computer games development. Research interests – game scripting, game design.

email: edward.bolshakov117@gmail.com

ORCID: 0000-0002-2208-9515

Материал поступил в редакцию 27 декабря 2022 года

УДК 004.91 + 004.774

МЕТОДЫ И ИНСТРУМЕНТЫ, ИСПОЛЬЗУЕМЫЕ ПРИ ПОДГОТОВКЕ ПУБЛИКАЦИЙ НАУЧНЫХ СТАТЕЙ В ФОРМАТЕ HTML

Р. Ю. Скорнякова[0000-0001-7372-3574]

Институт прикладной математики им. М.В. Келдыша Российской академии наук, г. Москва

rimmaskorn@gmail.com

Аннотация

Наряду с традиционной формой электронного представления полных текстов научных статей – форматом PDF – в последние годы все большее распространение получает формат HTML, обладающий для онлайн-публикаций рядом преимуществ за счет имеющихся в нем средств для лучшей структуризации материала, вставки мультимедийного контента и реализации разного рода интерактивных и динамических возможностей. В связи с этим становится весьма актуальной задача получения HTML-версии научной статьи из исходного формата материала, присланного автором. В настоящей работе рассмотрены различные подходы к подготовке HTML-версий полных текстов научных статей, применяемые в издательствах, и описаны используемые при этом программные инструменты. Основное внимание уделено инструментам, применяемым для исходных материалов в формате Word. Изложены также основы стандарта JATS XML, широко применяемого при подготовке онлайн-публикаций журнальных статей.

Ключевые слова*: HTML-версия научной статьи, XML-версия научной статьи, стандарт обмена научными статьями, JATS, преобразование форматов научных статей*

ВВЕДЕНИЕ

В настоящее время подавляющее большинство научных журналов имеет онлайн-версии и предоставляет полные тексты статей для открытого доступа или на коммерческой основе. Основным форматом представления полных текстов является PDF, однако в последние годы наметилась тенденция к публикации полных

текстов научных статей в формате HTML. В работе [1] проанализированы преимущества и недостатки этих форматов и сделаны выводы о дальнейших тенденциях в их использовании. Главное преимущество формата HTML – возможность предоставления дополнительного функционала, который сложно или невозможно реализовать в PDF, например, связи библиографических ссылок с библиографическими базами данных, встроенные мультимедиа-материалы [2], динамическое подгружение информации с других сайтов [3], в частности, даты последней редакции живой публикации в библиографической ссылке [4, 5].

В работе [6] изложены результаты опроса читателей о предпочтениях в выборе формата публикации. Большинство опрошенных считает удобным использовать HTML-версию для предварительного просмотра статьи и определения, насколько статья отвечает их интересам, а PDF-версию – для более внимательного чтения. Однако при наличии в HTML-версии динамики и интерактивных возможностей предпочтения могут быть отданы этому формату. Вывод, сделанный в работах [1, 6], состоит в том, что в ближайшем будущем HTML-формат полностью PDF-формат не заменит, и электронные журналы будут публиковать статьи в обоих форматах.

В связи с этим встает вопрос, как организовать процесс получения двух синхронизованных между собой версий из материала, присланного автором. Если получение PDF-версии из любого источника не представляет труда – все программы редактирования и верстки текстов дают возможность экспорта в PDF, то создание HTML-версии научной статьи не является столь простой задачей. Дело в том, что при преобразовании исходного формата в формат HTML целью является не воспроизведение внешнего вида текста (для этого годится формат PDF, HTML в таком случае и не нужен), а получение таким образом структурированного файла, чтобы можно было:

- при помощи общего стилевого оформления создать удобный, единый для всех статей онлайн-журнала дизайн, адаптируемый под размер устройства, используемого читателем;
- организовать удобную навигацию по тексту;
- реализовать динамические и интерактивные возможности;

- реализовать масштабируемое представление математических формул, доступное для машинной обработки и поиска.

Встроенные конвертеры, имеющиеся в редакторах, обычно используемых для набора текста статьи, такого качества HTML не дают.

История научных публикаций в HTML-формате насчитывает более 20 лет, однако единого подхода к созданию HTML-версий полных текстов статей за это время не выработано. Технологические цепочки получения HTML-версий в разных издательствах могут быть различными. Подход во-многом зависит от кадровых и финансовых возможностей издательства. Мы рассмотрим наиболее популярные из этих подходов и опишем программные инструменты, используемые при таких подходах. Основное внимание будет уделено инструментам, применяемым для исходных текстов в формате Word.

XML-ПРЕДСТАВЛЕНИЕ СТАТЬИ. СТАНДАРТ JATS

Один из наиболее распространенных подходов к формированию HTML-версии научной статьи состоит в предварительном создании XML-версии в соответствии со стандартом, принятым в данном журнале или издательстве. Для получения HTML используется XSLT или какой-либо иной способ автоматического преобразования. Часто XML-версия статьи используется и для автоматического преобразования в PDF, что позволяет получать синхронизированные версии статьи из одного источника.

Используемые XML-схемы отражают структуру научной статьи. В них, как правило, предусматриваются отдельные элементы для заголовка, метаданных, аннотации, библиографического списка, формул, рисунков, таблиц, затекстовых ссылок и т. п. Библиографическая ссылка может быть структурирована более детально с выделением отдельных элементов для авторов, названий работ, названий журналов и т. д.

Преимущество такого подхода состоит в том, что все статьи могут быть автоматически представлены в едином дизайне, и этот дизайн при желании нетрудно изменить. Кроме того, HTML-элементы и атрибуты, полученные при автоматическом преобразовании из XML-элементов, могут быть использованы для организации удобной навигации и реализации интерактивных и динамических

возможностей. Например, можно реализовать появление всплывающей подсказки, содержащей полный или частичный текст библиографической ссылки, при наведении курсора мыши на место ссылки внутри статьи.

Еще одно преимущество такого подхода – в том, что XML-формат отделяет структуру статьи от ее представления и тем самым упрощает хранение и обмен информацией, поиск данных, доступ к ним и управление ими. Современные СУБД предоставляют возможности для хранения данных в XML-формате и быстрого поиска в них. Реляционные СУБД, такие как Oracle, Microsoft SQL Server, расширили свои типы данных типом XML, имеются и специализированные XML-СУБД, для которых XML является основным форматом хранения. Одну из таких СУБД – MarkLogic Server – использует, например, для хранения статей издательство Nature Publishing Group[1], выпускающее большое число журналов, в т. ч. журнал Nature[2].

Обоснованию использования формата XML в издательских процессах посвящены работы [7–9]. Подробно об использовании XML-разметки при издании цифрового контента говорится в главе 3 сборника [10]. Преимуществам использования формата XML для научных публикаций посвящена работа [11]. В ней предлагается, в частности, в соответствии с концепцией Семантической паутины, использовать XML-представление научной статьи для формального описания научного знания, содержащегося в ней, путем добавления в XML-разметку элементов и атрибутов, отражающих понятия из конкретных научных областей, с использованием определенных словарей и онтологий.

Широкое использование формата XML для обмена журнальными статьями и хранения их в электронных библиотеках потребовало выработки для этой цели единого стандарта. За основу был взят разработанный в Национальной медицинской библиотеке США (NLM) стандарт NLM DTD, выпущенный в 2003 году и ставший де факто стандартом для хранения и обмена открытыми научными публикациями. Стандарт NLM DTD был доработан совместно с другими организациями и опубликован в 2006 году под названием JATS (Journal Article Tag Suite) как официальный стандарт Национальной организации по стандартизации информации

[1] https://publons.com/publisher/7/nature-publishing-group
[2] https://www.nature.com/

США (NISO)[3]. Текущая официальная версия JATS – 1.3, название NISO стандарта – ANSI/NISO Z39.96-2021 [12].

В публикациях [13–15] изложены основы стандарта JATS и рассказана история его создания. Изначально предполагалось, что издательства и веб-порталы будут использовать собственные наборы XML-элементов и преобразовывать документы к единому стандарту при обмене XML-документами друг с другом, при сохранении их в едином хранилище и/или при использовании общих программных инструментов и ресурсов. Разработчики стандарта проанализировали DTD XML-документов более 40 издательств и сотен журналов для выделения их общей структуры, общих метаданных, определения разметки библиографических ссылок и названий элементов. В результате анализа выяснилось, что DTD документов из различных источников на 80% совпадают. Разработанная модель целиком включила эту общую часть, а также отдельные структуры из несовпадающих 20%.

Поскольку в основу стандарта легли реально используемые в издательствах XML-схемы, он оказался удобен не только для обмена статьями, но и для подготовки статей к публикации в журнале. В настоящее время, по словам авторов работ [13, 14], с этой целью он используется большинством средних и мелких издательств Северной Америки и Европы. Значительная часть крупных издательств, в которых накоплено большое число XML-документов и налажен основанный на собственных схемах процесс подготовки публикаций, в основном продолжает использовать свои старые XML-схемы, однако некоторые из этих издательств запустили процесс перехода на стандарт JATS. Например, такое крупное издательство как Nature Publishing Group/Palgrave Macmillan, выпускающее порядка 180 журналов, перешло на формат JATS при выпуске новых журналов и планирует переход на этот формат в процессе подготовки выпусков старых журналов. В работе [16] изложены мотивы, побудившие это издательство начать переход с собственных XML-схем на стандарт JATS XML, и описан процесс перехода. Стоит отметить, что анализ собственных DTD и сравнение их с JATS, произведенные в издательстве, показали, что структурных элементов JATS достаточно для отображения информации, содержащейся в имевшихся XML-документах, расширение JATS не потребовалось.

[3]https://www.niso.org/

Стандарт JATS достаточно гибок: обязательных элементов в нем не очень много, можно использовать только необходимое подмножество. Различные издательства и порталы часто используют собственные спецификации JATS, составленные из элементов и атрибутов, входящих в основное описание стандарта. В качестве примеров разновидностей спецификаций JATS XML можно привести JATS, удовлетворяющий требованиям онлайн-архива медицинских статей PubMed Central[4], или SCJATS[5] – спецификацию, используемую популярной платформой для научных публикаций Silverchair.

Стандарт допускает расширения, в том числе элементами, отражающими семантику предметной области. Например, TaxPub[6] является расширением JATS, относящимся к области таксономии.

JATS де факто стал международным стандартом. Он используется более чем в 25 странах, в том числе в России, например, размещенными на издательской платформе ARPHA[7] российскими журналами «Nuclear Energy and Technology»[8] (издание МИФИ) и «Population and Economics»[9] (издание экономического факультета МГУ). На JATS XML основан язык представления метаданных цифровой математической библиотеки Lobachevskii-DML [17]. С целью расширения использования стандарта JATS ведутся работы по улучшению поддержки в нем многоязычия [18].

ОСНОВНЫЕ ЭЛЕМЕНТЫ JATS XML

Стандарт JATS включает в себя три набора элементов и атрибутов:

- Journal Archiving and Interchange Tag Set [19] – набор для хранения содержания журнала и обмена им;
- Journal Publishing Tag Set [20] – набор для подготовки публикации журнальных статей;
- Article Authoring Tag Set [21] – набор для первоначального ввода содержимого журнальных статей.

[4] https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/style.html
[5] http://specifications.silverchair.com/xsd/1/9/XMLSpecJournProc.html
[6] https://github.com/plazi/TaxPub
[7] https://arphahub.com/
[8] https://nucet.pensoft.net/
[9] https://populationandeconomics.pensoft.net/

Первый набор содержит элементы и атрибуты, описывающие содержание и метаданные журнальных статей. Он позволяет описывать как полное содержание статьи, так и метаданные в отдельности. Целью этого набора тегов является предоставление стандартизированного формата, в котором можно хранить информацию о ранее опубликованных журнальных статьях и в который контент из различных источников может быть переведен с минимальными потерями. Этот набор включает наибольшее число элементов, при этом правила использования для него являются наименее жесткими.

Второй набор тегов предназначен для XML-разметки статьи в издательстве для последующего преобразования в выходной формат. Как правило, эта разметка делается из исходного материала, предоставленного автором в каком-либо ином формате, например, в формате Word. В этом наборе меньше элементов, но среди них больше предписанных, и в большей степени регулируется порядок элементов.

Цель последнего набора тегов – предоставить пользователям стандартизованный формат для создания новых статей с помощью программных инструментов, управляемых моделями. В нем меньше всего элементов, но правила для этого набора являются самими жесткими.

Подробнее о назначении каждого набора и принципах их использования можно прочитать на сайте Национального центра биотехнологической информации США (NCBI[10]), посвященном JATS [22].

Корневым элементом во всех наборах является элемент <article> – статья. Он включает в себя элементы-контейнеры

- <front> – для заголовка и метаданных;
- <body> – для текста статьи;
- <back> – для дополнительной информации, включающей благодарности, библиографию, приложения, глоссарий и т. п.

Графики, таблицы, видео, относящиеся к статье, могут содержаться как в ее теле, так и отдельно. Для элементов, расположенных отдельно, предусмотрен контейнер <floats group>. В публикацию могут быть включены также отзывы на статью (<response>) и дополнительные материалы, оформленные как подстатьи

[10] https://www.ncbi.nlm.nih.gov/

(<sub-article>). Последние могут не иметь непосредственного отношения к статье, а относиться к журналу в целом.

Библиография <ref-list>, входящая как необязательный элемент в контейнер <back>, содержит отдельные библиографические ссылки в элементах <ref>. При этом библиографическая ссылка может быть оформлена как <element-citation> или <mixed-citation>. Первый вариант предназначен для оформления ссылки при помощи отдельных составляющих элементов без пунктуации и пробелов. Во втором случае ссылка представляется так, как она должна выглядеть в итоговом документе, при этом внутри нее могут быть выделены отдельные структурные элементы. Предписанных элементов в обоих случаях нет, но для распознавания ссылок компьютерными сервисами, такими, как, например, Crossref[11], желательно использовать определенный набор.

Тело статьи может включать отдельные разделы <sec>. Абзацам, так же, как и в HTML, соответствует элемент <p>. Внутри абзаца для смыслового выделения отдельных фрагментов предусмотрены элементы <bold>, <italic> и т. п.

Таблица вместе с заголовком и описанием помещается в контейнер <table-wrap>, а собственно таблица – в элемент <table>.

Предусмотрены также элементы для

- списков – <list>;
- рисунков – <fig>;
- указателей на внешние файлы, содержащие медиа-объекты – <media>;
- групп формул – <disp-formula-group>;
- фрагментов программного кода – <code>;
- ссылок внутри документа – <xref>.

Это далеко не полный список. Список всех элементов и атрибутов с описанием их назначения для каждого из трех наборов JATS доступен на сайте NCBI [19–21].

[11] https://www.crossref.org/

На сайте специально созданной для этой цели рабочей группы NISO – JATS4R (JATS For Reuse) [23] имеются практические рекомендации по использованию JATS с примерами. В качестве примеров форматирования в соответствии со стандартом JATS можно использовать также имеющиеся в открытом доступе научные статьи: некоторые ресурсы, например, CODATA Data Science Journal[12], PLOS – Public Library of Science[13], PeerJ – the Journal of Life and Environmental Sciences[14], предоставляют читателю возможность скачивать статьи в формате JATS XML, а хранилище SpringerLink предоставляет возможность получать статьи в формате JATS XML через Springer Open Access API[15].

ТЕХНОЛОГИИ ПОЛУЧЕНИЯ ВЫХОДНЫХ ФОРМАТОВ СТАТЬИ

Большинство западных издательств так или иначе использует формат JATS XML в своих рабочих процессах. По этапам, на которых применяется JATS XML, эти процессы упрощенно делятся на три основные группы, получившие условные названия XML-First (Рис. 1), XML-Middle (Рис. 2, Рис. 3, Рис. 4) и XML-Last (Рис. 5, Рис. 6).

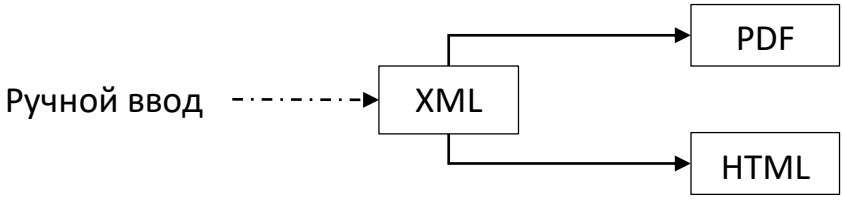


Рис. 1. Схема процесса XML-First

В процессах вида XML-First предполагается ручной ввод содержимого статьи в формате XML с использованием специализированных редакторов. Наиболее подходящей для этого вида рабочих процессов является XML-схема Article Authoring Tag Set. Процесс вида XML-Middle основан на использовании программ-конвертеров, преобразующих исходный формат статьи, присланной автором (обычно Word или LaTeX), в формат JATS XML, как правило, соответствующий схеме Journal Publishing Tag Set. Иногда процессы вида XML-Middle, при которых

[12] https://datascience.codata.org/
[13] https://plos.org/
[14] https://peerj.com/
[15] https://support.springer.com/en/support/solutions/articles/6000195668-springerlink-api-details

оба выходных формата PDF и HTML получаются из XML, называют процессами XML-First, имея при этом ввиду, что XML является основным форматом для хранения и обмена и создается прежде выходных форматов. Процесс вида XML-Last предполагает получение JATS XML в соответствии со схемой Journal Archiving and Interchange Tag Set на конечном этапе, после формирования основных выходных форматов.

Поскольку формат XML достаточно просто конвертируется в форматы HTML и PDF, издательствам было бы удобно, если бы авторы присылали статьи, уже набранные в формате JATS XML. Однако по словам эксперта в области редакционных и издательских технологий Билла Касдорфа, в прошлом президента Общества научных издательств (Society for Scholarly Publishing[16], SSP), а ныне руководителя собственного консалтингового агентства, успешные применения такой стратегии ему не известны, хотя попытки делались раньше и продолжаются до сих пор [24]. Авторам существенно проще набирать тексты статей в редакторах, традиционно предназначенных для этого. К тому же конкретные спецификации JATS, например, требования к наличию тех или иных элементов, в разных журналах могут быть разными, а авторы часто пишут статьи до принятия решения, в какой именно журнал статья будет направлена. Поэтому процессы XML-First, как правило, реализуются через привлечение дополнительного персонала или обращение к сторонним компаниям.

Процессы вида XML-Middle, основанные на использовании программ-конвертеров, как разработанных внутри самих издательств, так и имеющихся на

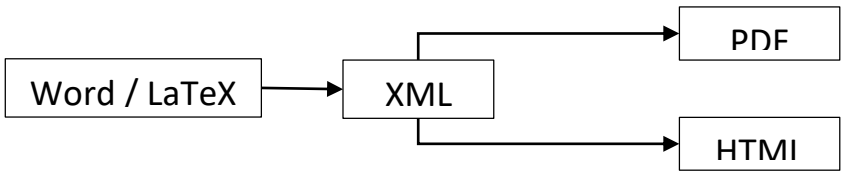


Рис. 2. Схема процесса XML-Middle (для PDF и HTML)

[16] https://www.sspnet.org/

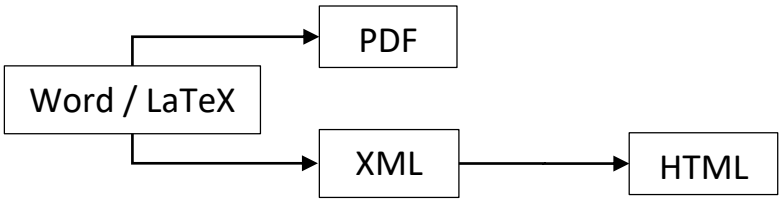


Рис. 3. Схема процесса XML-Middle (для HTML)

рынке программного обеспечения, требуют меньших временных и финансовых затрат [24, 25] и являются наиболее популярными. Существующие конвертеры из Word в JATS XML можно разделить на два класса: первые основываются на предположении, что исходный файл соответствует определенному шаблону, в котором для выделения семантики используется разметка стилями; вторые задействуют искусственный интеллект для анализа «сырого» файла. В первом случае результат получается более качественный, но необходима предварительная работа персонала издательства по приведению исходных файлов в соответствие с нужным шаблоном. Во втором случае необходима работа по доведению преобразованных файлов до полного соответствия стандарту JATS. Иногда издательства требуют, чтобы присылаемые тексты изначально были оформлены в соответствии с шаблоном, нужным для преобразования в JATS XML, однако далеко не все авторы достаточно хорошо владеют всеми возможностями редактора, и требуется работа профессионала по устранению ошибок, например, часто встречающейся в MS Word ошибки использования локального форматирования вместо предопределенного стиля.

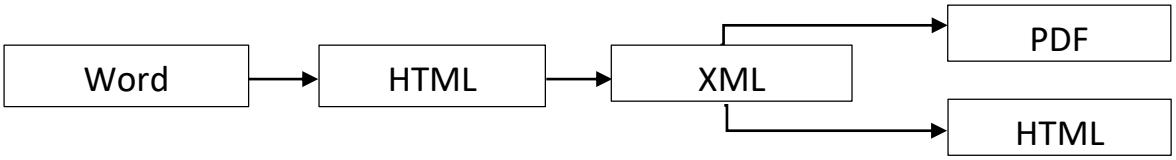


Рис. 4. Схема процесса XML-Middle с использованием промежуточного HTML

Несмотря на наличие различных программ-конвертеров и накопленный опыт работы с ними, преобразование исходных форматов в формат JATS XML остается довольно трудоемким и/или финансово затратным. Для формата Word основная сложность для преобразования состоит в отсутствии в формате необходимой семантики, отражающей структуру научной статьи, а для формата LaTex – сложность и вариабельность самого формата. Поэтому наряду с непосредственным преобразованием исходных форматов в JATS XML рассматриваются и другие

подходы. Например, авторы работы [26] предлагают использовать HTML как промежуточный формат при преобразовании формата Word в JATS XML (Рис. 4). Разработанный ими инструмент преобразует документ Word в HTML с сохранением внешнего вида. Отсутствующую семантику предлагается вносить не в документ Word, а в документ HTML, что, по мнению авторов, существенно проще при наличии специализированных инструментов.

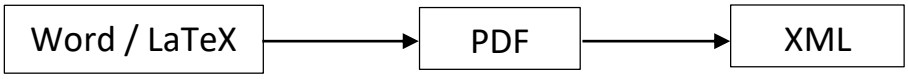


Рис. 5. Схема процесса XML-Last (без HTML)

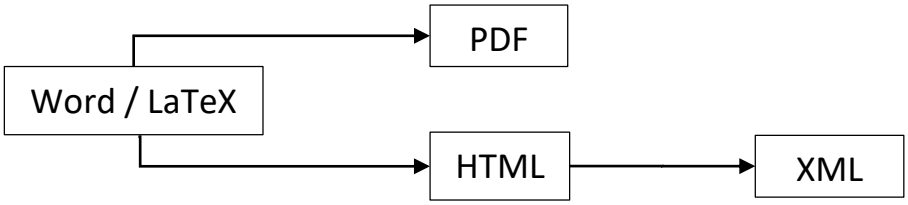


Рис. 6. Схема процесса XML-Last

Технология XML-Last используется в тех случаях, когда в издательстве налажены рабочие процессы получения выходных форматов, и издатели не имеют возможности или не хотят их изменять, а формат JATS XML используется только для хранения и/или обмена информацией. Часто в таких случаях выходной XML содержит только метаданные.

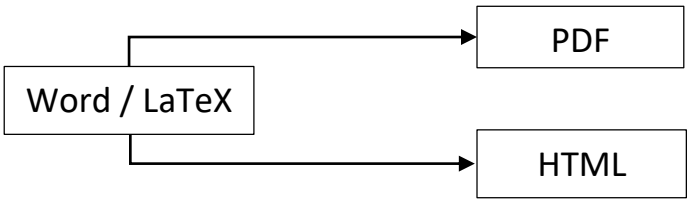


Рис. 7. Схема процесса без использования XML

Хотя использование формата JATS XML для журнальных публикаций дает много преимуществ, остаются еще издательства, не включающие получение этого формата в свои рабочие процессы по причине отсутствия достаточных финансовых и кадровых ресурсов. В этом случае производится непосредственная конвертация исходного формата рукописи в PDF и HTML без предварительной или последующей конвертации в JATS XML (Рис. 7).

Существенное увеличение доли онлайн-публикаций по сравнению с печатными, развитие средств и рост популярности семантической разметки веб-страниц, рост числа пользователей, знакомых с языком HTML, и наличие множества доступных инструментов для работы с этим форматом, и в то же время «недружественность» XML-формата по отношению к читателю и сложность получения

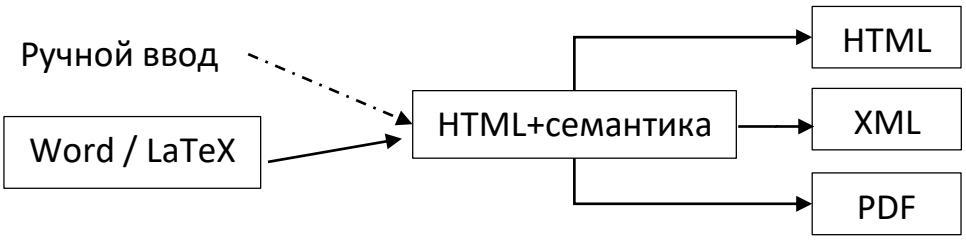


Рис. 8. Подход HTML-First

XML-версии статьи привели к появлению в последнее время интереса к подходу, условно называемому HTML-First (Рис. 8), при котором основным форматом для хранения научных статей и преобразования их в другие форматы является HTML. Одно из крупнейших академических издательств Wiley анонсировало в 2018 году начало процесса перевода своих производственных процессов с технологии XML-first, в основе которой лежал проприетарный формат WileyML, на технологию HTML-First [66]. Дополнительным стимулом к развитию подхода HTML-First послужило появление инициативы Linked Research[17], призывающей ученых самим публиковать результаты своих исследований и в рамках которой идет разработка инфраструктуры для таких публикаций.

Главное преимущество подхода HTML-First в том, что, в отличие от XML, HTML легко визуализируется при помощи браузеров, но для того чтобы HTML-версия научной статьи могла служить исходным форматом для хранения и преобразование в другие форматы, в HTML-код должна быть добавлена семантика научной статьи. С этой целью консорциум WWW (W3C) разрабатывает стандарт Scholarly HTML[18], в основе которого лежит тип ScholarlyArticle[19] широко используемого стандарта семантической разметки веб-страниц Schema.org[20]. Семантику

[17] https://linkedresearch.org/
[18]https://w3c.github.io/scholarly-html/
[19] https://schema.org/ScholarlyArticle
[20] https://schema.org/

научной статьи в HTML в стандарте Scholarly HTML предлагается вводить с помощью синтаксиса RDFa или JSON-LD. Авторы статьи [67] предлагают свой вариант стандарта HTML для научных статей – Research Articles in Simplified HTML (или RASH), формальная грамматика которого описана на языке RELAX NG, и который, помимо добавления семантики научной статьи, ограничивает использование языка HTML 32-мя элементами. Существуют и другие варианты внесения семантики научной статьи в HTML, например, PubCSS[21] – набор HTML-шаблонов и CSS для представления научных публикаций как в HTML, так и в PDF, Dokieli[22] – где статьи представляются в формате HTML+RDFa. Общепринятого стандарта описания структуры научной статьи, такого как JATS XML, для HTML пока нет.

Текст научной статьи в формате HTML при подходе HTML-First может либо вводиться вручную, либо получаться с помощью конвертеров из традиционно используемых форматов Word или LaTeX.

ПРОГРАММНЫЕ ИНСТРУМЕНТЫ

В зависимости от того, какой из вариантов рабочего процесса выбран, используются различные типы программных инструментов. Для процессов вида XML-First используются специализированные XML-редакторы, поддерживающие стандарт JATS, и инструменты, преобразующие JATS XML в HTML и PDF. Технология XML-Middle требует наличия программ-конвертеров из исходного формата (как правило, Word или Tex) в JATS XML. На начальном этапе, перед конвертацией документа Word, могут использоваться дополнительные инструменты, позволяющие при помощи специального форматирования внести в документ необходимую семантику. После преобразования в XML для исправления ошибок и доведения результата преобразования до полного соответствия стандарту JATS могут понадобиться XML-редакторы. Облачные XML-редакторы используются также для совместного редактирования статьи авторами и редакторами издательства. В процессах вида XML-Last используются конвертеры из HTML или PDF в XML. На конечном этапе также могут быть использованы XML-редакторы. При любом из вариантов рабочего процесса, если процесс включает получение на каком-либо

[21] https://github.com/thomaspark/pubcss/
[22] https://dokie.li/docs

из этапов документа в формате JATS XML, для проверки соответствия этого документа стандарту необходим JATS XML-валидатор. Он может быть встроен в XML-редактор или установлен отдельно. При подходе HTML-First могут использоваться HTML-редакторы, конвертеры из традиционных форматов (Word, Tex) в специализированный HTML-формат и конвертеры из специализированного HTML-формата в PDF и XML. Авторы, знакомые с HTML, могут вводить тексты статей непосредственно в HTML-формате с использованием определенных шаблонов.

Существуют программные продукты, совмещающие в себе несколько из вышеописанных функций, а также целые издательские платформы, включающие в себя как часть специализированные редакторы и конвертеры.

XML-РЕДАКТОРЫ С ПОДДЕРЖКОЙ JATS

Одним из наиболее популярных XML-редакторов является коммерческий редактор Oxygen XML Editor [27], устанавливаемый как отдельное приложение или как плагин к среде разработки Eclipse[23]. Он имеет встроенную поддержку широко используемых стандартов семантической XML-разметки документов DITA[24], DocBook[25], TEI[26], XHTML. В последнюю версию редактора была добавлена и поддержка JATS.

При создании научного контента часто применяют коммерческий WYSIWYG онлайн-редактор Fonto [28], предназначенный в первую очередь для пользователей, не знакомых с XML. Редактор может работать с различными XML-схемами, в том числе с JATS. Fonto не имеет своего хранилища данных и, в отличие от Oxygen, не может использоваться автономно, а только в интеграции с другими системами: системами управления цифровыми активами (DAM), системами управления содержимым (CMS), репозиториями; к примеру, в интеграции с платформой MarkLogic Data Hub[27].

[23] https://www.eclipse.org/eclipseide/
[24] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=dita
[25] http://docs.oasis-open.org/docbook/docbook/v5.1/os/docbook-v5.1-os.html
[26] https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
[27] https://www.fontoxml.com/partners/marklogic/

Редактор Fonto лежит в основе облачного редактора LiXuid Manuscript [29] компании Aries Systems, интегрированного с популярными издательскими платформами Editorial Manager и ProduXion Manager, разработанными этой компанией для оптимизации процессов редактирования и публикации отредактированных статей. LiXuid Manuscript имеет Word-подобный интерфейс с автоматической разбивкой на страницы, организованной при помощи Adobe InDesign[28].

Встроенную поддержку стандартов семантической разметки документов, включая JATS, имеет также коммерческий WYSIWYG XML-редактор Xeditor [30], реализованный как веб-приложение и имеющий средства для интеграции с другими системами: CMS, DAM, базами данных, редакционными системами. Он может работать как локально, так и через облако, пользовательский интерфейс осуществляется через любой современный браузер.

Перечисленные редакторы могут быть интегрированы с широко используемым редактором математических формул MathType[29].

Ведутся также работы по созданию свободно распространяемых специализированных JATS XML-редакторов с открытым исходным кодом. В 2018 году на основе JavaScript-библиотеки для редактирования веб-контента Substance[30] консорциумом Substance, включающим сообщества Public Knowledge Project (PKP)[31], Collaborative Knowledge Foundation (CoKo)[32], онлайн-библиотеки SciELO[33], Érudit[34] и журнал eLife[35], был создан WYSIWYG JATS XML редактор Texture [31, 32], который может устанавливаться как отдельно, так и как плагин к свободно распространяемой редакционно-издательской системе Open Journal Systems (OJS)[36]. Имеющаяся версия редактора предназначена в первую очередь издателям для использования на этапе доведения «до ума» результата автоматического преобразования в JATS XML исходного варианта рукописи, как в плане соответствия стандарту, так

[28] https://www.adobe.com/ru/products/indesign.html
[29] https://docs.wiris.com/mathtype
[30] https://github.com/substance/substance
[31] https://pkp.sfu.ca/
[32] https://coko.foundation/
[33] https://scielo.org/
[34] https://apropos.erudit.org/
[35] https://elifesciences.org/
[36] https://pkp.sfu.ca/ojs/

и в плане содержания статьи. Использование редактора упрощает эти процессы благодаря тому, что доведением до соответствия стандарту может заниматься сотрудник издательства, не знакомый с XML, и возможности редактора позволяют авторам и сотрудникам издательства работать над текстом статьи совместно, аналогично совместной работе с документом в Google Docs[37]. В дальнейшем разработчики планировали расширить пользовательский интерфейс, чтобы редактором могли пользоваться и авторы в процессе написания статьи, однако работа над редактором была прекращена в 2019 году, оставшись незавершенной. Не все элементы JATS были реализованы, хотя значительная их часть, включая таблицы, рисунки, цитирование, формулы, в редакторе присутствуют. Формулы поддерживаются в формате Tex.

На смену редактору Texture должен прийти редактор Libero [33]. На данный момент – это тоже незавершенная работа. В основе редактора лежит ProseMirror[38] – набор инструментов с открытым исходным кодом для создания редакторов форматированного текста в интернете. Изначально редактор создавался командой разработчиков журнала eLife как часть свободно распространяемой издательской платформы Libero Publisher[39], но в 2021 году работы по разработке платформы были прекращены, а редактор передан для дальнейшего развития сообществу Coko Foundation.

ИНСТРУМЕНТЫ ДЛЯ ВАЛИДАЦИИ И ВИЗУАЛИЗАЦИИ JATS XML

В Сети можно найти как JATS-валидаторы общего назначения, проверяющие XML-файл на соответствие JATS DTD, так и специализированные, производящие проверку на соответствие версии стандарта, используемой конкретным издательством или порталом. Последние помимо проверки на соответствие JATS DTD могут включать проверку на соответствие дополнительным требованиям.

К валидаторам общего назначения относится XML-валидатор [34], представленный на сайте архива находящихся в свободном доступе статей по биомедицинской тематике PubMed, поддерживаемого Национальным центром биотехнологической информации США (NCBI), основным разработчиком JATS.

[37] https://www.google.ru/intl/ru/docs/about/
[38] https://prosemirror.net/
[39] https://github.com/libero/publisher

Рабочая группа JATS4R (JATS for Reuse) Национальной организации по стандартизации информации США (NISO), выдающая рекомендации по использованию JATS, предоставляет свой валидатор [35], осуществляющий проверку на соответствие JATS DTD и рекомендациям этой группы. Исходный код отдельных его компонент (пользовательского интерфейса[40], веб-службы[41], Schematron-правил[42] и используемых DTD[43]) выложен на GitHub. Его можно кастомизировать и использовать для проверки на соответствие требованиям конкретного журнала.

Примерами специализированных валидаторов могут служить валидатор PMC Style Checker [36] упомянутого выше онлайн-архива PubMed или ScienceCentral Style Checker [37] Научного центра республики Корея.

Наиболее распространенный подход к визуализации XML-документов– использование XSL-преобразований. В открытом доступе имеются XSL-файлы для преобразования JATS XML в HTML и PDF, разработанные в Национальном центре биотехнологической информации США (NCBI) – JATS Preview Stylesheets [38]. Они предназначены для предварительного просмотра статей, представленных в формате JATS XML, а также для использования в качестве отправной точки для дальнейшей адаптации под требования конкретного издательства [39, 40]. Журнал PeerJ публикует на портале GitHub XSL-преобразования [41], используемые в этом журнале, и php-код, их выполняющий.

Другой подход – преобразование формата XML в формат JSON и динамическая прорисовка при помощи JavaScript-кода. Такой подход используется в издательстве Nature Publishing Group/Palgrave Macmillan, где тексты статей в формате JATS XML хранятся в СУБД MarkLogic и извлекаются по запросу [16]. Этот же подход используется в разработанном в журнале открытого доступа eLife средстве просмотра JATS XML-файлов Lens [42]. Окно просмотра делится на две части (Рис. 9): слева отображается основной текст статьи, а справа – дополнительный контент. При нажатии на ссылку, указывающую на рисунок или библиографическую ссылку, в дополнительной панели автоматически отображается нужный контент.

[40] https://github.com/JATS4R/jats-validator-ui
[41] https://github.com/JATS4R/jats-validator
[42] https://github.com/JATS4R/jats-schematrons
[43] https://github.com/JATS4R/jats-dtds

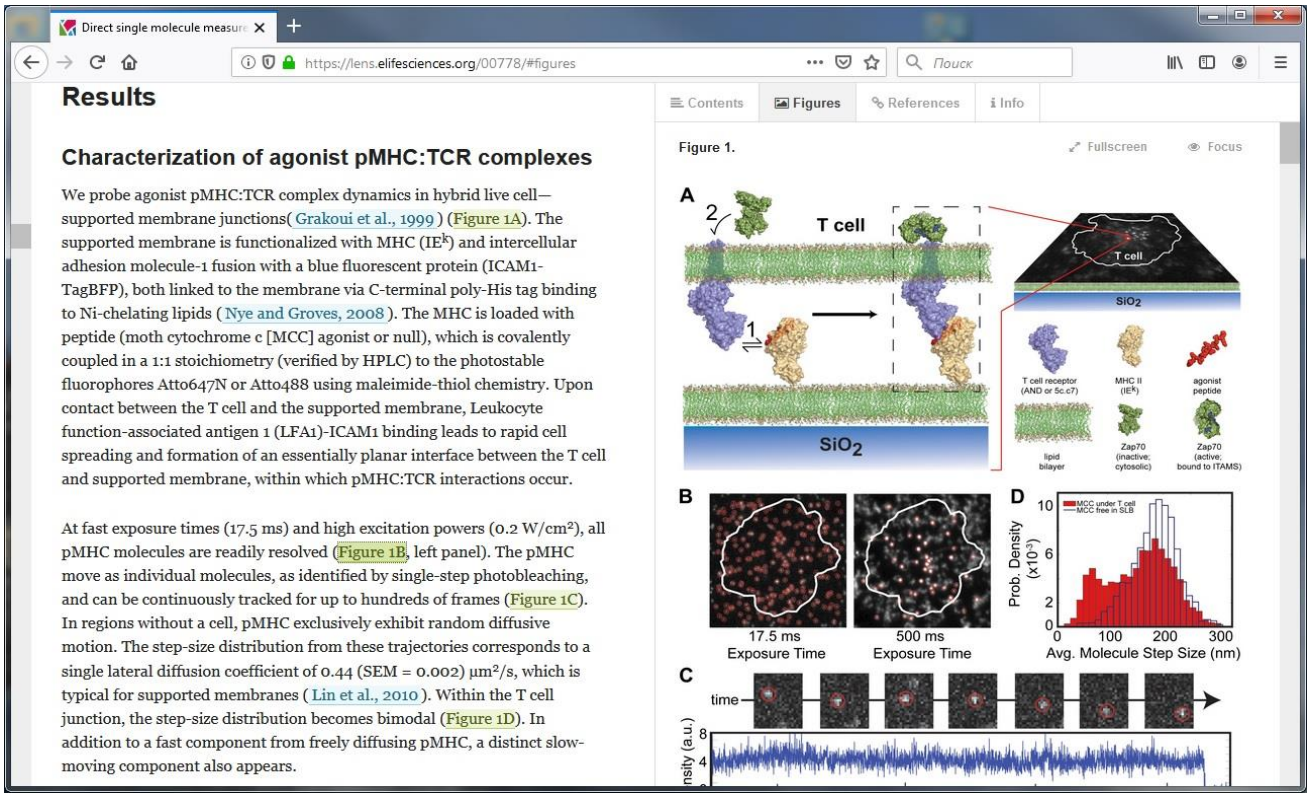


Рис. 9. Вид статьи во вьюере Lens

Вьюер eLife Lens представляет собой JavaScript-библиотеку с открытым исходным кодом [43], допускающим модификацию и расширения.

ИНСТРУМЕНТЫ ДЛЯ ПРЕОБРАЗОВАНИЯ ДОКУМЕНТОВ WORD В ФОРМАТ JATS XML

В издательской среде признано, что среди инструментов, предоставляющих возможность конвертировать документы Word в формат JATS XML, наиболее качественные результаты дают решения компании Inera: eXtyles JATS и eXtyles Custom [44]. eXtyles JATS – это готовое решение для получения XML, удовлетворяющего требованиям портала PubMed Central и агенства Crossref, eXtyles Custom – решение, настраиваемое под требования к JATS XML конкретного издателя. Устанавливаемые как плагины к Word, эти продукты позволяют автоматизировать трудоемкие аспекты процесса производства XML-документов — вычищение, форматирование, редактирование и собственно преобразование в XML.

Преобразование документа Word в формат JATS XML в eXtyles основано на использовании предопределенной палитры пользовательских стилей, включающей как стили абзацев, так и символьные стили. Стилям, как правило, соответствуют элементы JATS XML.

Применение стилей абзацев в плагине eXtyles происходит через отдельный диалог, в котором стили разделены на несколько групп. В eXtyles JATS используются следующие группы:

- **Front** — стили абзацев вступительной части (Рис. 10);
- **Trans** — стили абзацев, содержащих переводной текст (Рис. 10);
- **Body** — стили абзацев основной части статьи (Рис. 11);
- **List** — стили для списков (Рис. 11);
- **Object** — стили для объектов, таких как таблицы, рисунки, текстовые поля (Рис. 11);
- **Back** — стили для абзацев заключительной части (Рис. 10)

Для ускорения процесса разметки после применения выбранного стиля курсор автоматически переходит на следующий абзац.

Рис. 10. Стили абзацев в eXtyles JATS (Front, Trans, Back)

Рис. 11. Стили абзацев в eXtyles JATS (Body, List, Object)

Символьные стили (они доступны через основное меню стилей Word) используются для выделения элементов внутри абзацев, например, для выделения отдельных элементов библиографической ссылки: имен авторов, названий статей, года выхода, интервала страниц и т. д. (Рис. 12). Как правило, расстановка стилей для элементов библиографической ссылки не производится вручную. В eXtyles имеется функция обработки библиографических ссылок. Она автоматически определяет тип библиографической ссылки (журнал, книга и т. д.) и реструктурирует ссылки в соответствии со стилем оформления списка литературы, используемым данным издательством. Ручное применение стилей библиографии необходимо только для исправления ошибок.

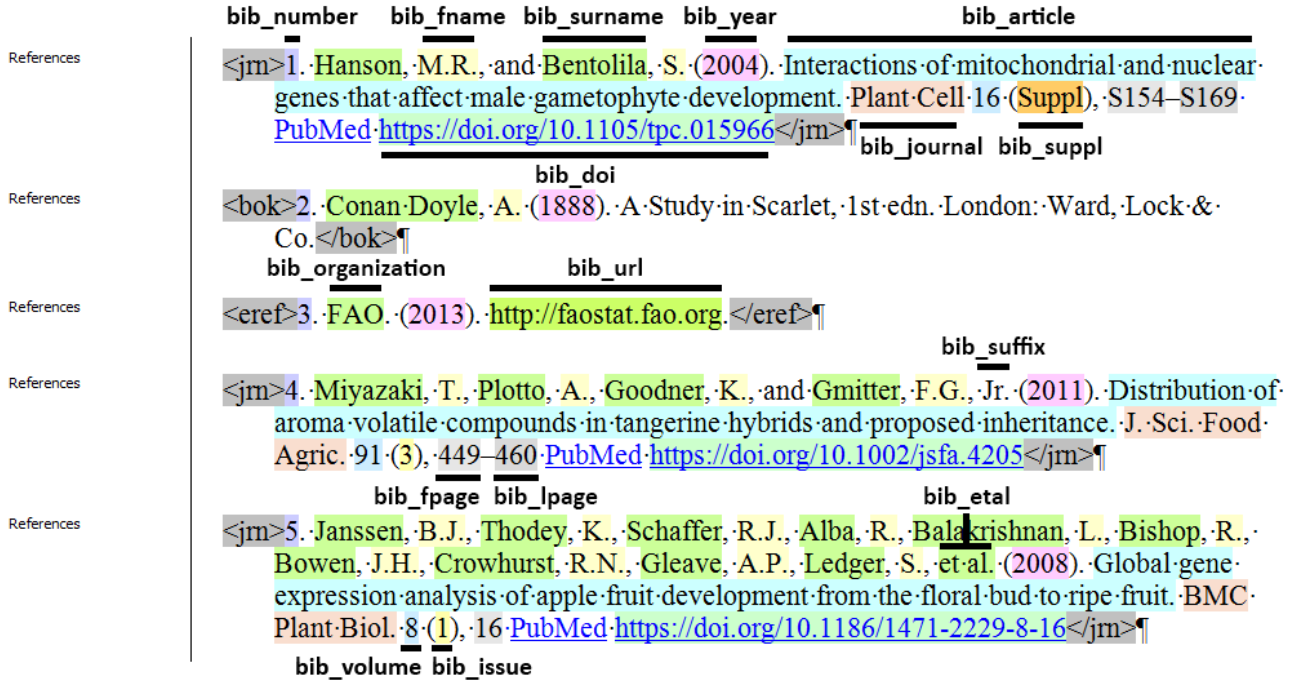


Рис. 12. Использование символьных стилей eXtyles JATS для элементов библиографической ссылки

Помимо разметки стилями, eXtyles предоставляет возможность проверки библиографических ссылок на соответствие стандартам (ISO, EN и др.) и базам данных PubMed и CrossRef. Проверка производится путем обращения к веб-службам соответствующих порталов.

Перед расстановкой стилей обычно делается предварительное автоматическое форматирование документа, которое опционально может включать в себя вычищение документа от нежелательных символов, применение основного стиля

ко всем обычным абзацам, распознавание библиографических списков, применение к каждой библиографической ссылке определенного стиля и др. (Рис. 13).

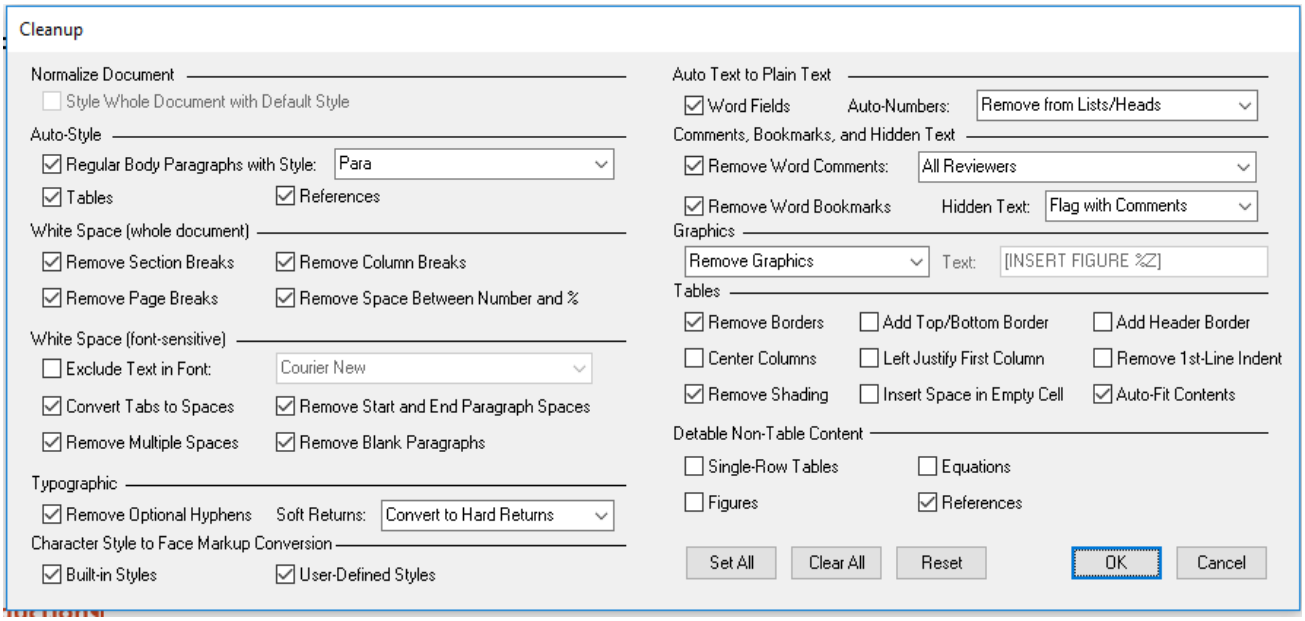


Рис. 13. Опции предварительного форматирования документа

При преобразовании документа Word в JATS XML eXtyles использует также контекст и предположение о наличии определенных ограничений, например, таблицы в документе не должны использоваться для форматирования. В последних версиях eXtyles в предварительное форматирование документа включена опция «детаблитизации» подобных фрагментов текста.

Конвертер eXtyles позволяет экспортировать математические формулы, созданные в том числе при помощи редактора формул MathType, в один из следующих форматов, допустимых в JATS XML: MathML, изображение или их комбинацию. Хотя формулы, созданные с помощью Microsoft Equation 3.0 или Microsoft Equation Builder, тоже конвертируются, рекомендуется сначала преобразовать их в формат MathType.

Основываясь на двадцатилетнем опыте развития и эксплуатации eXtyles, компания Inera в 2019 году выпустила новый продукт eXtyles Arc [45], который, используя технологии искусственного интеллекта, позволяет получать JATS XML из документа Word без предварительной ручной разметки документа стилями. Продукт включает два решения: eXtyles Arc Metadata Extraction и eXtyles Arc Full-Text Extraction. Первое предназначено для извлечения метаданных, второе – для

преобразования в JATS XML полного текста статьи. Хотя eXtyles Arc не требует детальной разметки стилями, определенные ограничения на документ Word все же накладываются. Например, документ не должен содержать фигуры и графические объекты SmartArt; все изображения должны предоставляться в виде отдельных файлов; нельзя использовать таблицы для форматирования; документ не должен содержать вложенные таблицы; нельзя использовать встроенные таблицы Excel и еще ряд других ограничений.

Компания Inera тесно сотрудничает с компанией Typefi Systems[44], предоставляющей решения для генерации различных форматов публикаций из одного источника (Рис. 14). Разработанная компанией издательская платформа Typefi [46] позволяет производить рендеринг сложных макетов с использованием динамических шаблонов и дизайнерских методов, задействуя для этого Adobe InDesign, широко используемое программное обеспечение для профессиональной верстки страниц. Использование eXtyles для преобразования формата Word

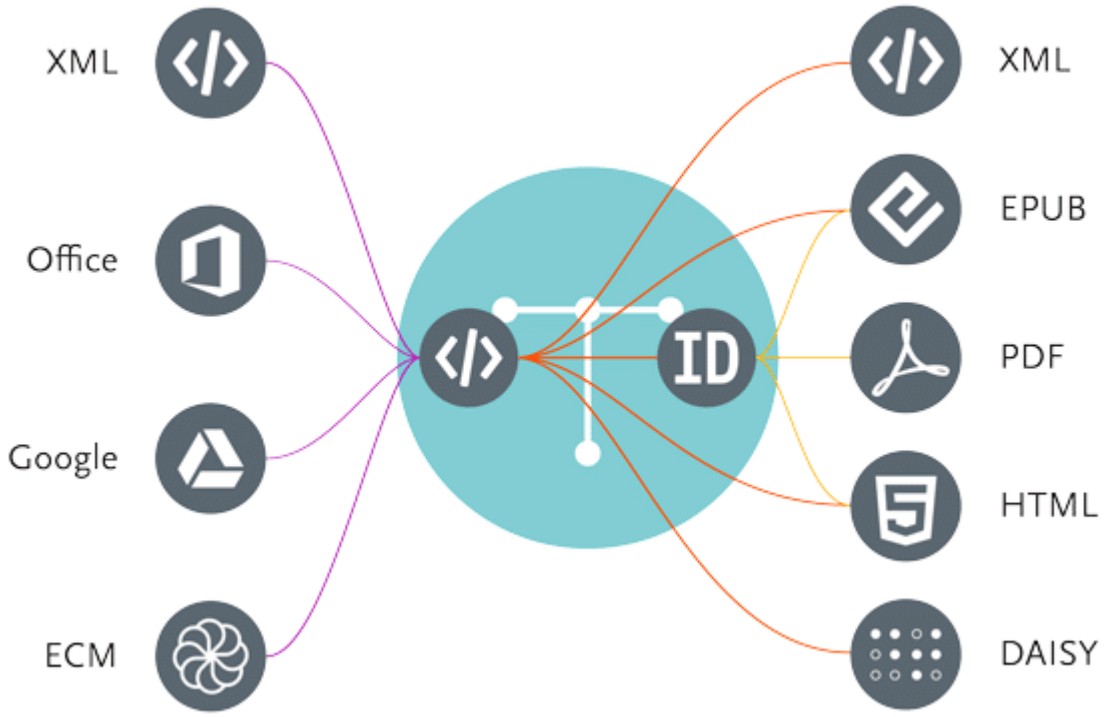


Рис. 14. Преобразования форматов, осуществляемые платформой Typefi
Источник рисунка https://www.typefi.com/wp-content/uploads/lotus-diagram.png

[44]https://www.typefi.com/

в формат JATS XML и Typefi для преобразования JATS XML в выходные форматы [25, 47] позволяет получать качественные результаты за довольно короткое время, однако продукты эти очень дорогие (порядка десятков тысяч долларов в год [48]), и их покупку могут позволить себе только крупные издательства с большим бюджетом.

Среди менее дорогих инструментов (порядка тысяч долларов в год[45]) хорошие отзывы [24] получил выпущенный в 2017 году компанией Ictect программный продукт Intelligent Content for Journals [49], работа которого основана на технологиях искусственного интеллекта. Тестирование, проведенное крупными издателями, входящими в ассоциацию STM[46], показало, что этот инструмент создает правильные и детальные структуры JATS из более чем половины исходных рукописей, а более 90% рукописей могут быть усовершенствованы для получения правильного результата менее чем за десять минут обычными редакторами контента, не знакомыми с XML.

Инструмент представляет собой сервис, с помощью которого можно загрузить документ Word на специализированный сервер Intelligent Content Server и автоматически получить от него на выбор два документа: документ Word с добавленной в него разметкой тегами JATS и документ JATS XML (Рис. 15). Сервис предлагается в облачном и локальном вариантах, в первом случае сервер управляется компанией Ictect, во втором – клиенты запускают сервер самостоятельно. При

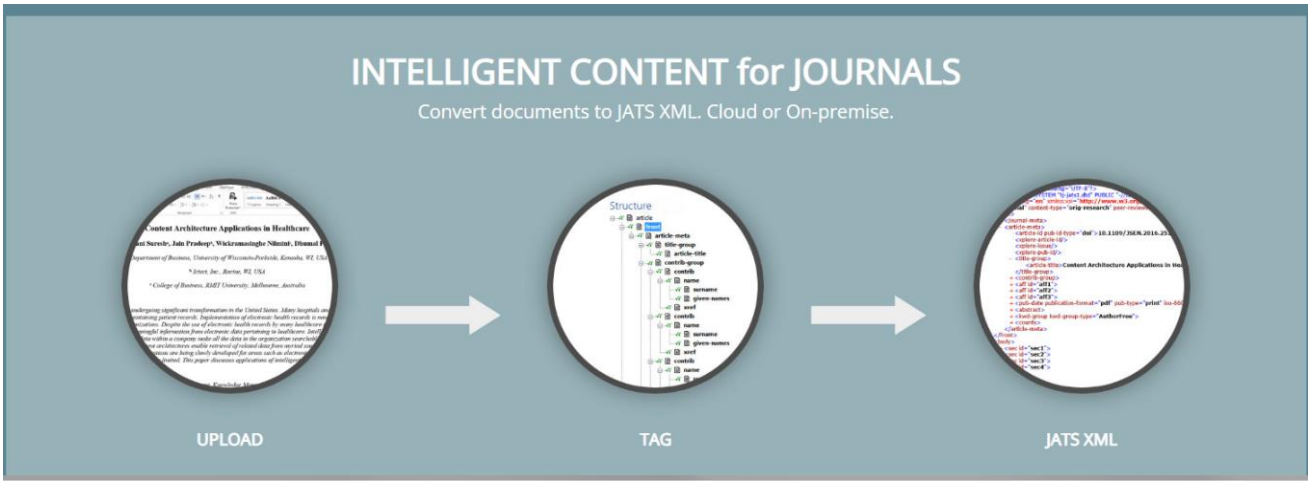


Рис. 15. Преобразование документа Word при помощи Ictect

Источник рисунка https://www.ictect.com/images/new_website_images/splash-jats-2b-tag.jpg

[45] https://typeset.io/resources/top-4-ms-word-docx-to-jats-xml-converters/

[46] https://www.stm-assoc.org/

установке, в обоих случаях, сервис настраивается на используемую клиентом версию JATS.

Добавленную в документ Word JATS-разметку можно увидеть и отредактировать с помощью разработанного компанией Ictect плагина Intelligent Content Tools (icTools). При наличии плагина окно документа Word делится на две панели: в левой панели показывается сам документ Word, в правой – иерархическая структура, соответствующая JATS-разметке (Рис. 16). Панели синхронизированы между собой: когда пользователь помещает курсор на конечный элемент в правой панели, в левой – соответствующая часть текста выделяется цветом; аналогично, если дважды щелкнуть мышью в каком-нибудь месте левой панели, соответствующий элемент в правой панели выделяется жирным шрифтом.

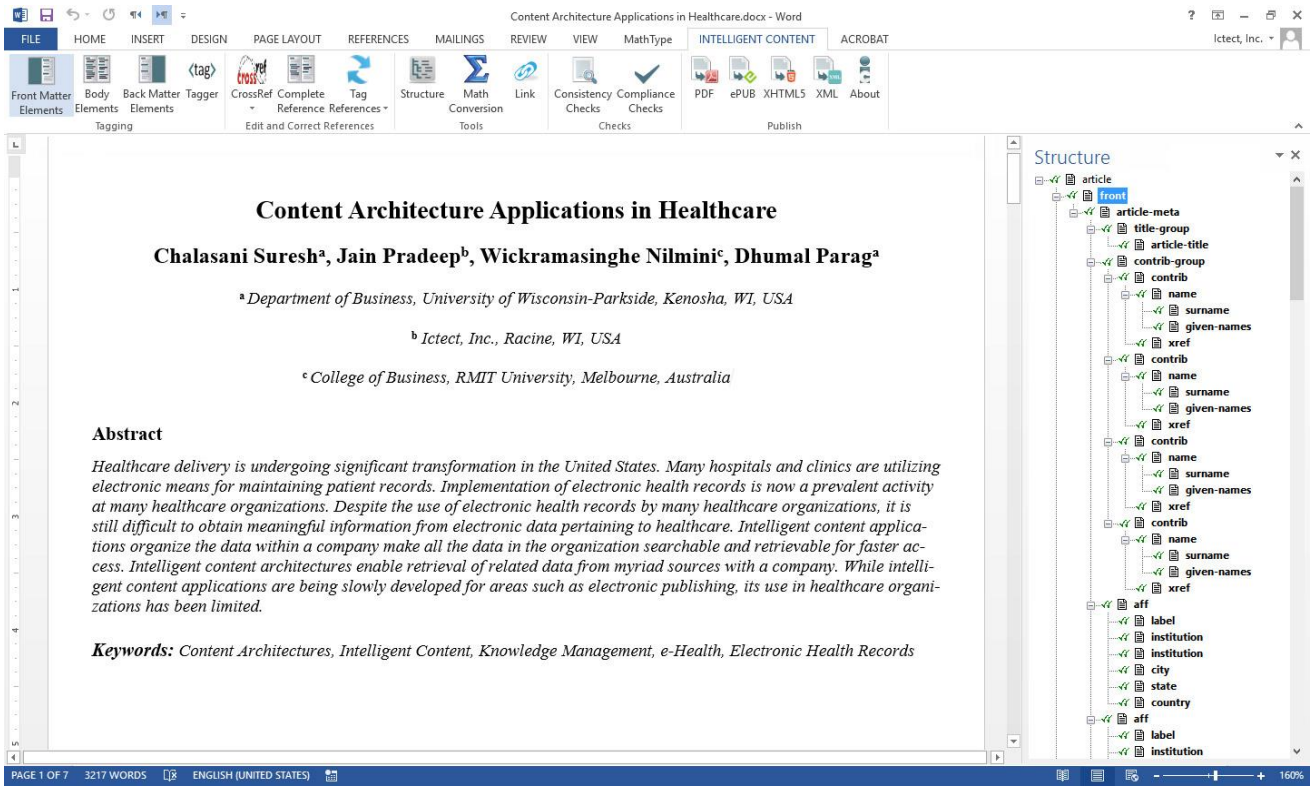


Рис. 16. JATS структура документа Word в плагине icTools

Если Intelligent Content Server не смог распознать какую-то часть контента, он помечает ее желтым цветом, а в структуре ставит ей в соответствие элемент unknown. После редактирования помеченного текста пользователь может вызвать распознавание подходящей части контента. Например, если в библиографической ссылке есть слово Vol., а конкретный номер пропущен, вместо элемента

volume в структуре создается элемент unknown. После вставки номера и вызова функции «Reference for periodical» из меню «Back matter elements» нужный элемент появляется в структуре.

Такой подход с использованием двух панелей имеет определенные преимущества перед разметкой стилями, используемой в плагине Inera eXtyles, поскольку пользователь видит одновременно исходный текст и конечный результат, и изменения в тексте практически сразу отражаются в конечном результате.

Конвертер Ictect распознает библиографические ссылки, основываясь на правилах оформления ссылок, принятых в данном журнале. Если ссылка оформлена правильно, то распознавание ее отдельных элементов происходит полностью автоматически, если есть ошибки, то ссылка размечается частично и исправляется в плагине icTools с помощью обращения к REST API CrossRef[47]. Согласно информации, размещенной на сайте компании, Ictect поддерживает преобразование формул в форматы MathML и LaTeX, допустимые в JATS, и рисунки, содержащиеся как внутри документа, так и во внешних файлах. С помощью плагина icTools можно проверять документ на соответствие руководству по стилю, принятому в данном журнале, и экспортировать документ в форматы HTML, PDF и ePUB. Имеется также ряд других возможностей, упрощающих и ускоряющих совместную работу авторов и редакторов.

Работа Intelligent Content for Journals основана на анализе содержимого на английском языке, и на данный момент его пользователями являются только американские издательства.

Стоит отметить, что упомянутая выше платформа Typefi [46] также предоставляет средства для конвертации научных статей из формата Word в формат JATS XML. При помощи входящего в состав платформы модуля Typefi Writer, плагина к Word, производится разметка документа, затем размеченный документ конвертируется во внутренний формат платформы – Content XML, и из него уже осуществляется конвертация в другие форматы, в том числе в JATS. TypefiWriter конвертирует формулы, созданные редактором MathType, в формат MathType

[47] https://www.crossref.org/documentation/retrieve-metadata/rest-api/

EPS[48], который может быть преобразован в формат MathML в модуле Typefi Designer при помощи дополнительного подключаемого стороннего модуля movemen MathTools[49]. Конвертация в JATS с помощью Typefi обходится дешевле, чем с помощью Inera eXtyles, однако и результаты получаются менее качественными. Typefi чаще используется для конвертации уже готового XML в выходные форматы.

Имеются и дешевые предложения, но их качество оставляет желать лучшего. К примеру, конвертер [50] предлагает компания SciSpace (предыдущее название Typeset.io), основным продуктом которой является облачный редактор для научных статей. О конвертере и редакторе довольно много отрицательных отзывов. Наша попытка конвертировать в JATS XML исходный текст статьи в формате docx, воспользовавшись демонстрационной версией, окончилась неудачей. Конвертер не смог правильно выделить разделы, поставив, в частности, разделитель в середине абзаца. Местонахождение формул было определено более или менее правильно, но сами формулы конвертированы не были. На посланный в техподдержку вопрос, может ли их продукт конвертировать формулы, набранные в редакторе MathType, ответа не последовало, вместо этого пришло несколько писем с рекламой возможностей продуктов компании. Распознавание отдельных элементов библиографических ссылок тоже оказалось неудовлетворительным.

Ряд свободно распространяемых инструментов с открытым исходным кодом также декларируют, что включают возможность конвертации документов Word в формат JATS XML, однако это либо незаконченные разработки, либо конвертация осуществляется с потерями, либо инструмент представляет собой основу, которую надо дополнять пользовательским кодом.

Одним из таких инструментов является Pandoc [51] – написанный на языке Haskell универсальный конвертер разметок документов, включающий библиотеку и инструмент командной строки. Pandoc позволяет конвертировать различные форматы, в частности, формат DOCX в JATS, а также DOCX в HTML и HTML в JATS. Инструмент состоит из множества отдельных конвертеров считывания, пре-

[48] https://www.adobe.com/creativecloud/file-types/image/vector/eps-file.html
[49] http://movemen.com/

образующих исходный формат во внутреннее промежуточное представление документа в виде абстрактного синтаксического дерева (AST), и множества отдельных конвертеров записи, преобразующих это представление в целевой формат. Внутреннее представление Pandoc обладает более слабыми выразительными возможностями, чем многие из преобразуемых им форматов, поэтому при конвертации возможны потери. Pandoc позволяет настраивать конвертацию путем добавления программ-фильтров на языке Haskell или Python, преобразующих AST, а также создания пользовательских конвертеров из AST в целевой формат при помощи скриптов на языке lua.

Среди свободно распространяемых универсальных конвертеров можно отметить также Transpect [52] – фреймворк, созданный немецкой компанией Le-Tex для преобразования различных форматов, базирующиеся на XML. При конвертации в качестве промежуточного формата используется специально введенный разработчиками формат Hub XML[50], представляющий собой видоизмененный формат DocBook, в котором не обязательно наличие разделов <section> и добавлены атрибуты стилей в формате CSSa[51], т. е. CSS, представленного в виде XML-атрибутов. Работа инструмента основана на XSL-преобразованиях с использованием языка XProc. Управляющий код написан на языке Java и использует XML Calabash[52] – интерпретатор языка XProc.

Конвертация документа Word в файл формата JATS XML производится в несколько этапов:

- вначале при помощи модуля docx2hub[53] файл формата DOCX преобразуется в файл формата Hub XML, содержащий всю информацию о форматировании исходного файла,
- затем этот файл преобразуется в файл того же формата, но более пригодный для конвертации в JATS,
- и уже затем происходит конвертация непосредственно в JATS XML.

[50] https://github.com/le-tex/Hub
[51] https://github.com/le-tex/CSSa
[52] https://xmlcalabash.com/
[53] https://github.com/transpect/docx2hub

Семантика вносится на втором, промежуточном этапе – происходит идентификация элементов списков на основании отступов; по соответствию имен разделов регулярным выражениям устанавливается их иерархия; таблицы и рисунки объединяются с их заголовками и т. п. Для этого используются специальные XSL-преобразования, которые могут быть изменены пользователем для настройки на требования конкретного издательства. Третий этап также использует настроечную информацию, содержащуюся в XSL-файлах и XML-файлах специального формата, по которой определяется, в частности, как должны интерпретироваться названия стилей абзацев и символов. При наличии в исходном документе формул в формате MathType нужен еще один шаг для их преобразования в формат MathML. Он осуществляется с помощью модуля mathtype-extension[54].

Усилия по созданию конвертеров документов Word в JATS XML предпринимаются и сообществом Public Knowledge Project (PKP), главным образом с целью интеграции их в разработанную PKP свободно распространяемую платформу для автоматизации редакционно-издательских процессов Open Journal Systems (OJS), широко используемую как зарубежными, так и отечественными издательствами [53, 54]. Была попытка использовать для конвертации упомянутый выше инструмент Pandoc, однако результат тестирования оказался неудовлетворительным: Pandoc плохо распознавал структуру документа и иерархию его частей. В настоящий момент для использования с OJS предлагаются два конвертера: meTypeset [55, 56] и docxToJats [57]. Оба конвертера выполняют функцию конвертации лишь частично и не могут сделать процесс преобразования документа Word в XML полностью автоматическим. Предполагается,

Рис. 17. Схема рабочего процесса в OJS с использованием конвертера Word в JATS XML

Источник рисунка https://i0.wp.com/ojs-services.com/wp-content/uploads/2021/12/xml_publishing_in_ojs_-_project_summary_user_guide6.png?resize=1024%2C133&ssl=1

[54] https://github.com/transpect/mathtype-extension/

что результат преобразования будет дорабатываться вручную с помощью описанного выше XML-редактора Texture (Рис. 17).

Написанный на языке Python инструмент meTypeset использует эвристический подход и не предполагает наличие в документе Word специальных пользовательских стилей. Вначале он при помощи XSL-преобразований, разработанных для проекта OxGarage[55], делает преобразование документа формата DOCX в формат TEI, а затем, анализируя встроенные стили для заголовков, использование жирных шрифтов, курсива, подчеркиваний и изменения размеров шрифтов, определяет структуру документа. Для того чтобы определить, какой из выделенных разделов может быть кандидатом на раздел библиографических ссылок, он использует список фраз-синонимов для библиографии на разных языках. Выделение отдельных ссылок происходит с использованием возможно имеющихся в исходном документе Word тегов XML, вставленных плагинами Zotero[56] или Mendeley Cite[57], а также путем нахождения в конце документа очень коротких абзацев, имеющих одинаковую структуру отступа. Согласно документации, имеющейся на GitHub, инструмент поддерживает изображения, таблицы, списки, сноски. Формулы поддерживаются, но только в формате OMML (Office Math Markup Language) – собственном формате Word.

Инструмент docxToJats представляет собой PHP-библиотеку для конвертации документов формата DOCX в формат JATS XML. Библиотека используется как подмодуль в плагине для OJS «DOCX to JATS XML Converter Plugin»[58]. Для определения структуры документа используются встроенные стили заголовков. Инструмент поддерживает списки, таблицы, изображения в формате JPEG и PNG. Конвертация формул и сносок пока не реализована, планируется включить их в следующую версию.

Разработка meTypeset практически прекращена, docxToJats продолжает дорабатываться и на данный момент рассматривается как более предпочтительный для OJS.

[55] https://wiki.tei-c.org/index.php/OxGarage
[56] https://www.zotero.org/
[57] https://www.mendeley.com/reference-management/mendeley-cite
[58] https://github.com/Vitaliy-1/docxConverter

В работе [58] описана попытка встроить JATS XML в рабочий процесс основанной на OJS службы публикаций библиотеки Арктического университета Норвегии, предпринятая в 2020-м году и окончившаяся неудачей. Были опробованы оба конвертера. Для использования meTypeset потребовалось предварительно отформатировать документы определенным образом, результаты получились смешанные. Результаты конвертации при помощи docxToJats оказались сырыми и потребовали существенной ручной доработки выходного XML-файла. Тем не менее, авторы решили, что docxToJats им подойдет больше, поскольку продолжает активно разрабатываться. Редактирование выходного XML осуществлялось при помощи редактора Texture, в котором реализовано лишь ограниченное подмножество элементов JATS. Были еще трудности, связанные с несовместимостью плагина JATS Parser с используемой версией OJS. В итоге уложиться в сроки, отведенные для подготовки публикаций, не удалось. Авторы пришли к выводу, что, хотя на данном этапе встроить JATS XML в рабочий процесс не удалось, оставлять эти попытки не стоит, планируя в дальнейшем использовать либо улучшенные версии опробованных инструментов, либо другие, совместимые с OJS, инструменты, которые, возможно, удастся отыскать.

ИНСТРУМЕНТЫ ДЛЯ ПРЕОБРАЗОВАНИЯ ДОКУМЕНТОВ WORD В ФОРМАТ HTML

Существует множество автоматических конвертеров документов Word в формат HTML, как коммерческих, так и бесплатных, однако это инструменты общего назначения, не учитывающие специфику научных статей. В полученном HTML не будут выделены аннотация, авторы, библиографический список, отдельные элементы библиографического списка и т. п. Результат преобразования надо будет существенно дорабатывать вручную. В особенности это касается статей, содержащих математические формулы. Бесплатные конвертеры (большой, но далеко не полный список таких конвертеров можно найти, например, в обзоре [59]) формулы либо пропускают, либо преобразуют в картинки, что исключает возможность их машинной обработки и существенно ограничивает возможности визуального представления для лучшего восприятия человеком.

Возможность конвертации в формат HTML имеется в самом редакторе Word. Преобразовать документ в формат HTML можно, вызвав диалог «Сохранить

как» и выбрав один из вариантов формата: «Веб-страница в одном файле», «Веб-страница» или «Веб-страница с фильтром». Первые два варианта – «тяжелые», они содержат много лишней информации, которая нужна, чтобы страница в браузере отображалась в точности так же, как и в самом приложении Word, и могла быть преобразована обратно в исходный формат. Последний вариант – более «легкий», фильтрованная веб-страница содержит только основную информацию о форматировании, файлы получаются существенно меньшего размера, такой вариант более пригоден для размещения в Сети. Однако существенным недостатком этого преобразования является то, что часть таблиц и формулы конвертируются в изображения (используются форматы gif, png, jpg) с низким разрешением: формулы, особенно сложные, получаются плохо читаемыми.

Плагин MathType также предоставляет возможность преобразования документа Word в HTML через вызов функции «Publish to MathPage». При этом есть две опции для формул – переводить их в изображения или в формат MathML. К сожалению, эта функция работает ненадежно. Из трех документов с формулами нам удалось успешно конвертировать в HTML только один. В процессе преобразования двух других возникала нераспознанная ошибка (Рис. 18), HTML-страница при этом создавалась, но часть формул была конвертирована не в MathML, а в изображения с низким разрешением. Переустановка MathType избавиться от проблемы не помогла. Жалобы на возникновение подобной ошибки нам встречались и в Сети.

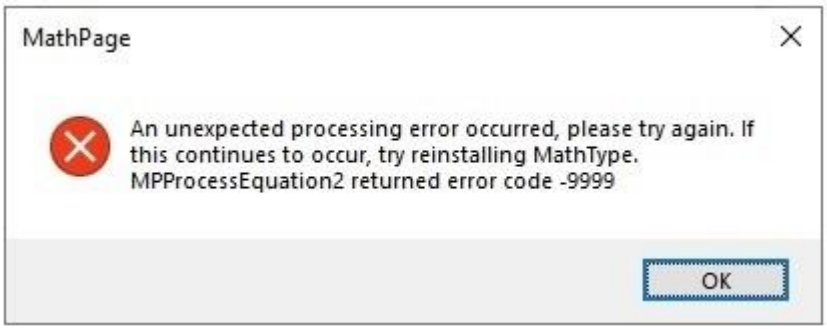


Рис. 18. Ошибка при попытке создания MathPage

Неплохого качества HTML получается при помощи условно бесплатного инструмента XMLmind Word To XML [60] французской компании XMLmind Software[59],

[59] https://www.xmlmind.com/

написанного на языке Java. XMLmind предоставляет облачный сервис, позволяющий бесплатно конвертировать в HTML (а также популярные XML-схемы для разметки документов DITA, DocBook и др.) ограниченное число документов в день. Формулы преобразуются в формат svg, дающий достаточно качественное изображение. Коммерческий вариант продукта позволяет управлять результатом преобразования программным путем при помощи скриптов на языке XED, основанном на языке XPath.

Довольно популярен коммерческий инструмент Doc Converter Pro [61], позволяющий конвертировать документы Word в различные форматы, в том числе HTML, PDF, EPUB. Пользователь может настраивать преобразование через создание собственных шаблонов. Настройка включает множество деталей, в частности, выходной формат для изображений: PNG, BMP, WMF, GPG или GIF; в каком виде должны быть представлены формулы: в виде изображений, MathML или текста, а также управление результирующим HTML и CSS при помощи регулярных выражений поиска и замены. Doc Converter Pro представлен в трех видах: как онлайн-сервис, как десктоп-приложение для Windows и как Rest API.

Существуют также различные конвертеры из Word в HTML с открытым исходным кодом, написанные на разных языках программирования и создаваемые с различной целью.

Упомянутая выше некоммерческая организация Collaborative Knowledge Foundation разработала конвертер с открытым исходным кодом XSweet [62], представляющий собой серию XSL-преобразований. Одна из целей создания этого инструмента – использовать результат конвертации в HTML для дальнейшего преобразования в JATS XML [26] при помощи Pandoc [51] и скриптов на языке lua. Работа над инструментом еще не завершена. Конвертер поддерживает списки, в том числе вложенные, таблицы, сноски, гиперссылки, однако конвертация изображений и формул на данный момент не реализована.

Конвертер документов Word в формат HTML входит в пакет Open XML PowerTools [63], написанный на языке C# и использующий библиотеку Open XML SDK, разработанную Microsoft для работы с документами Microsoft Office. Конвертер стремится в точности повторить внешний вид документа Word, что не соответствует, на наш взгляд, целям публикации научной статьи в HTML-формате.

Среди конвертеров документов Word в HTML-формат с открытым исходным кодом, написанном на языке Java, пользуется популярностью конвертер, входящий в пакет Opensagres XDocReport [64]. Конвертер допускает множество настроек, в том числе выбор лежащей в основе библиотеки работы с документами Word. Можно выбрать, например, Apache POI. На наш взгляд, для конвертации научных статей этот инструмент не очень удобен, в частности, из-за сложности использования.

Достаточно простой и ясный HTML-код получается при конвертации документов Word с использованием инструмента с открытым исходным кодом Mammoth [65], разработанного английским программистом Майклом Уильямсоном. Конвертер использует только семантическую информацию и игнорирует второстепенные детали. Например, абзацы со стилем Heading1 преобразуются в h1-элементы, при этом конвертер не пытается точно скопировать стиль заголовка (шрифт, размер текста, цвет и т. д.). Для пользовательских стилей имеется возможность сопоставить эти стили с соответствующим HTML-кодом, например, стилю BibliographyHeading сопоставить h1.bibliography. Инструмент поддерживает списки, таблицы, изображения, сноски, ссылки, форматирование текста (жирный шрифт, курсив, подчеркивание, зачеркивание, надстрочный и подстрочный индексы), однако форматирование таблиц не поддерживается, не поддерживаются и формулы. Mammoth используется как плагин в популярной системе управления содержимым сайта WordPress. Исходный код представлен на языках программирования Java, C#, Python, JavaScript. Код C# получен автоматическим преобразованием из кода Java. На наш взгляд, Mammoth является подходящим инструментом с открытым исходным кодом, чтобы использовать его в качестве основы для написания конвертера исходного текста научной статьи в формате Word в формат HTML.

ИНСТРУМЕНТЫ ДЛЯ ИСПОЛЬЗОВАНИЯ В РАМКАХ ПОДХОДА HTML-FIRST

Для авторов, знакомых с языком HTML, программист из Филадельфии Томас Парк создал библиотеку таблиц стилей CSS и шаблонов HTML – PubCSS [68],

поддерживающую на данный момент форматы научных статей ACM[60] и IEEE[61]. По мнению Парка, язык HTML проще для авторов, чем LaTeX, и, хотя сложнее, чем Word, но имеет больше возможностей для структурирования контента; HTML можно рассматривать как компромисс между Word и LaTeX.

В рамках проекта SOLID[62] – инициативы Тима Бернерса-Ли по редецентрализации Сети, цель которого – предоставить пользователям полный контроль своих данных, включая контроль доступа и место хранения, был разработан инструмент с открытым исходным кодом dokieli [69], предоставляющий средства для создания и аннотирования статей непосредственно в браузере. Подробнее о dokieli и возможностях его использования в децентрализованных авторских и издательских системах можно прочесть в работе [70].

Авторы работы [67] разработали набор инструментов с открытым исходным кодом для работы с научными статьями в предложенном им формате – RASH Framework [71]. Платформа включает в себя файлы CSS и скрипты на языке Javascript для визуализации RASH-документов, валидаторы на соответствие HTML-документа стандарту RASH, веб-редактор на языке JavaScript для создания научных статей в формате RASH [72], набор XSL-преобразований для преобразования документов Word, составленных в соответствии с специальными рекомендациями в формат RASH, а также для преобразования RASH в LaTeX.

ЗАКЛЮЧЕНИЕ

В настоящее время основной подход к подготовке публикации научных статей в формате HTML состоит в предварительном создании XML-версий статей со схемой, отражающей структуру научной статьи. Такой подход позволяет не только получать качественный HTML, легко воспринимаемый человеком, но и делает статью доступной для машинной обработки.

В США разработан стандарт XML-представления научной статьи, получивший название Journal Article Tag Suite или, сокращенно, JATS. Он стал универсальным

[60] https://www.acm.org/publications/authors/reference-formatting#:~:text=ACM%20IN%2DTEXT%20CITATION%20STYLE&text=Sequential%20parenthetical%20citations%20are%20enclosed,%5B1999%5D...%22
[61] https://www.ieee.org/conferences/publishing/templates.html
[62] https://solidproject.org/

стандартом при обмене информацией о научных статьях и широко используется при подготовке публикаций.

Для создания XML-версии статьи используются два основных подхода:

- непосредственный ввод содержимого статьи в XML-формат специально обученным персоналом;
- конвертация в XML-формат материала, присланного автором в одном из традиционно используемых форматов (Word, LaTeX).

Первый подход чаще всего реализуется через аутсорсинг.

Для исходного материала в формате Word наилучшие результаты при конвертации в JATS XML дает коммерческий продукт Inera eXtyles [44], часто используемый в комбинации с Typefi [46] для получения выходных HTML и PDF-форматов статьи из одного источника. Для внесения семантики, отсутствующей в документе Word и относящейся к структуре научной статьи, Inera eXtyles использует множество специальных пользовательских стилей, разметку исходного материала которыми должны осуществлять сотрудники издательства.

В последнее время наметилась тенденция в использовании для создания конвертеров из Word в JATS XML технологий искусственного интеллекта. Продукты Ictect [48] и Inera eXtyles Arc [45], основанные на этих технологиях, не требуют предварительной разметки документа специальными стилями. Они дают менее качественные результаты, чем традиционно используемый Inera eXtyles, но время, необходимое для доработки выходного XML до соответствия стандарту, получается небольшим. По всей видимости, дальнейший прогресс в совершенствовании конвертеров статей из формата Word в JATS XML будет связан именно с этим подходом.

Упомянутые продукты являются дорогими, и далеко не все издательства могут их себе позволить. Разработка бесплатных инструментов с открытым исходным кодом ведется, но пока их качество не достигло такого уровня, чтобы свести к минимуму использование ручного труда.

В последние годы наметился интерес к подходу, условно называемому HTML-First, при котором основным форматом для хранения научных статей и преобразования их в другие форматы является HTML с добавленной в него семантической разметкой, отражающей структуру научной статьи. Общепринятого стандарта этой разметки пока нет, он находится в стадии разработки. На наш взгляд, подход может

иметь большие перспективы, в особенности, если получит развитие инициатива Linked Research.

СПИСОК ЛИТЕРАТУРЫ

1. *Чебуков Д.Е.* Об HTML версии полного текста научной статьи // Труды XX Всероссийской научной конференции «Научный сервис в сети Интернет», г. Новороссийск, 17–22 сентября 2018 г. М.: ИПМ им. М.В. Келдыша, 2018. С. 487–498. URL: https://keldysh.ru/abrau/2018/theses/16.pdf, https://doi.org/10.20948/abrau-2018-16.

2. Анимация и видео в научной публикации / *М.М.Горбунов-Посадов [и др.]* // Препринты ИПМ им. М.В. Келдыша. 2014. № 104. 32 с. URL: https://library.keldysh.ru/preprint.asp?id=2014-104.

3. *Китаев Е.Л., Скорнякова Р.Ю.* Скрейпинг «на лету» внешних веб-ресурсов, управляемый разметкой HTML-страницы // Препринты ИПМ им. М.В. Келдыша. 2019. № 20. 31 с. https://doi.org/10.20948/prepr-2019-20, URL: https://library.keldysh.ru/preprint.asp?id=2019-20.

4. *Горбунов-Посадов М.М.* Живая публикация // Открытые системы. 2011. № 4. С. 48–49. URL: https://keldysh.ru/gorbunov/live.htm.

5. *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. URL: https://library.keldysh.ru/preprint.asp?id=2017-82, https://doi.org/10.20948/prepr-2017-82.

6. *Aalbersberg I.J.* PDF versus HTML – which do researchers prefer? // Elsevier connect. 9 Jul 2013. URL: https://www.elsevier.com/connect/pdf-versus-html-which-do-researchers-prefer.

7. *Kasdorf W.E.* The XML revolution // Learned Publishing. 2001. Vol. 14, No. 3. P. 223–231. https://doi.org/10.1087/095315101750240485.

8. *Young D., Madans P.* XML: Why Bother? // Publishing Research Quarterly. 2009. No. 25. P. 147–153. https://doi.org/10.1007/s12109-009-9120-4.

9. *Rech D.A.* Instituting an XML-First Workflow // Publishing Research Quarterly. 2012. No. 28. P. 192–196. https://doi.org/10.1007/s12109-012-9278-z.

10. *Kasdorf W.E.* The Columbia Guide to Digital Publishing. NYC: Columbia University Press, 2003. 816 p.

11. *Murray-Rust P., Rzepa H.S.* Scientific publications in XML – towards a global knowledge base // Data Science. 2002. No. 1. P. 84–98.
https://doi.org/10.2481/dsj.1.84.

12. Standardized Markup for Journal Articles: Journal Article Tag Suite (JATS) // NISO website, 07.07.2021. URL: https://www.niso.org/standards-committees/jats.

13. *Lapeyre D.A.* Introduction to JATS (Journal Article Tag Suite) // XML.com. 12.10.2018. URL: https://www.xml.com/articles/2018/10/12/introduction-jats/.

14. *Usdin B.T., Lapeyre D.A.* JATS/BITS/NISO STS // Proceedings of the Symposium on Markup Vocabulary Ecosystems. Balisage Series on Markup Technologies, vol. 22 (2018), Washington, DC, USA, 30.07.2018.
https://doi.org/10.4242/BalisageVol22.Usdin01.

15. *Beck J.* NISO Z39.96 The Journal Article Tag Suite (JATS): What Happened to the NLM DTDs? // The Journal of Electronic Publishing. 2011. Vol. 14, issue 1.
https://doi.org/10.3998/3336451.0014.106.

16. *Donohoe P., Sherman J., Mistry A.* The Long Road to JATS // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 21–22, 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK279831/.

17. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Базовые сервисы фабрики метаданных цифровой математической библиотеки Lobachevskii-DML // Электронные библиотеки. 2020. Т. 23. № 3. С. 336–381.
https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

18. *Lizzi V.* Improving JATS for multilingual articles // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, May 3–4, 2022.
URL: https://www.ncbi.nlm.nih.gov/books/NBK579699/.

19. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3 // National Center for Biotechnology Information.
URL: https://jats.nlm.nih.gov/archiving/tag-library/1.3/chapter/set-intro.html.

20. Journal Publishing Tag Library NISO JATS Version 1.3 // National Center for Biotechnology Information.

URL: https://jats.nlm.nih.gov/publishing/tag-library/1.3/chapter/journal-tag-set-intro.html.

21. Article Authoring Tag Library NISO JATS Version 1.3 // National Center for Biotechnology Information.
URL: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.3/chapter/set-intro.html.

22. Journal Article Tag Suite // National Center for Biotechnology Information URL: https://jats.nlm.nih.gov/.

23. JATS4R (JATS for Reuse). Официальный сайт. URL: https://jats4r.org/.

24. *Kasdorf W.E.* Getting from Word to JATS XML // The Association of Learned and Professional Society Publishers blog. 18.10.2018.
URL: https://blog.alpsp.org/2018/10/getting-from-word-to-jats-xml.html.

25. *Adam L.R., Perera C.* eXtyles, Typefi, and the NLM Journal Publishing DTD // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, November 1–2, 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK47080/.

26. *Piez W.* HTML First?: Testing an alternative approach to producing JATS from arbitrary (unconstrained or "wild") .docx (WordML) format // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 25–26, 2017.
URL: https://www.ncbi.nlm.nih.gov/books/NBK425546/.

27. Oxygen XML Editor. URL: https://www.oxygenxml.com/.

28. Fonto Editor. URL: https://www.fontoxml.com/.

29. LiXuid Manuscript.
URL: https://www.ariessys.com/blog/introduction-lixuid-manuscript-xml/.

30. XEditor. URL: https://www.xpublisher.com/products/xeditor.

31. Texture JATS XML editor. URL: https://github.com/substance/texture.

32. *Garnett A., Aufreiter M., Buchtala O., Alperin J. P.* Introducing Texture: An Open Source WYSIWYG Javascript Editor for JATS // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 25–26, 2017.
URL: https://www.ncbi.nlm.nih.gov/books/NBK425544/.

33. Libero Editor. URL: https://gitlab.coko.foundation/libero/editor.

34. PMC XML Validator // National Center for Biotechnology Information.
URL: https://www.ncbi.nlm.nih.gov/pmc/tools/xmlchecker/.

35. JATS4R Validator // JATS4R, NISO Working Group.

URL: https://validator.jats4r.org/.

36. PMC Style Checker // National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/pmc/tools/stylechecker/.

37. ScienceCentral Style Checker // The Korean Federation of Science and Technology Societies. URL: https://www.e-sciencecentral.org/tools/stylechecker/.

38. JATS Preview Stylesheets // GitHub.com. URL: https://github.com/ncbi/JATSPreviewStylesheets.

39. *Piez W.* Fitting the Journal Publishing 3.0 Preview Stylesheets to Your Needs: Capabilities and Customizations // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, November 1–2, 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK47104/.

40. *Graham T.* Formatting JATS: as easy as 1-2-3 // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 1–2, 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK189779/.

41. PeerJ/jats-conversion: *Conversion and validation for JATS XML* // GitHub.com URL: https://github.com/PeerJ/jats-conversion.

42. Seeing through the eLife Lens: A new way to view research // Inside eLife, Jun 6, 2013. URL: https://elifesciences.org/inside-elife/0414db99/seeing-through-the-elife-lens-a-new-way-to-view-research.

43. Lens // GitHub.com URL: https://github.com/elifesciences/lens.

44. Inera JATS Solutions. URL: https://www.inera.com/jats-solutions/.

45. Inera eXtyles Arc. URL: https://www.inera.com/extyles-arc/.

46. Typefi: Aut*omated publishing for print, online, and mobile.* URL: https://www.typefi.com/products-services/.

47. Q&A: End-to-end automation with eXtyles Arc and Typefi. URL: https://www.typefi.com/qa-end-to-end-automation-with-extyles-arc-and-typefi/.

48. *Eve M.P.* The Means of (Re-)Production: Expertise, Open Tools, Standards and Communication // Publications. 2014. No. 2. P. 38–43. https://doi.org/10.3390/publications2010038.

49. Ictect Intelligent Content for Journals. URL: https://www.ictect.com/JATS-XML.

50. SciSpace JATS XML Converter.
URL: https://typeset.io/for-publishers/jats-xml/.

51. Pandoc. URL: https://pandoc.org/.

52. Transpect. An Open Source framework for converting and checking data. URL: https://transpect.github.io/.

53. *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://rdl-journal.ru/article/view/377/468.

54. *Галявиева М.С., Елизаров А.М., Липачёв Е.К.* Цифровая инфраструктура электронного научного журнала: автоматизация редакционно-издательских процессов и система сервисов // Электронные библиотеки. 2016. Т. 19, № 5. С. 408–465. URL: https://rdl-journal.ru/article/view/404/489.

55. meTypeset. URL: https://github.com/withanage/meTypeset.

56. *Garnett A., Alperin J.P., Willinsky J.* The Public Knowledge Project XML Publishing Service and meTypeset: Don't call it "Yet Another Word-to-JATS Conversion Kit" // Journal Article Tag Suite Conference (JATS-Con) Proceedings 2015 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2015.
URL: https://www.ncbi.nlm.nih.gov/books/NBK279666/.

57. DocxToJats. URL: https://github.com/Vitaliy-1/docxToJats.

58. *Ekanger A., Odu O.* How we tried to JATS XML // Ravnetrykk. 2020. No. 39. P. 156–162. https://doi.org/10.7557/15.5517.

59. 13 Best Free Word to HTML Converter Software for Windows.
URL: https://listoffreeware.com/free-word-to-html-converter-software-windows/.

60. XMLmind Word To XML: Convert DOCX to unstyled, valid, “semantic” XHTML *1.0, 1.1 or 5.0.* URL: https://xmlmind.com/w2x/docx_to_xhtml.html.

61. Doc Converter Pro. URL: https://docconverter.pro/.

62. XSweet. URL: https://xsweet.org/.

63. Open XML PowerTools.
URL: https://github.com/OpenXmlDev/Open-Xml-PowerTools/.

64. Opensagres XDocReport. URL: https://github.com/opensagres/xdocreport.

65. Mammoth. URL: https://mike.zwobble.org/projects/mammoth/.

66. *Siegman T., Young B.* HTML-First at Wiley // BookNet Canada blog. 14.02.2018. URL: https://www.booknetcanada.ca/blog/2018/2/14/html-first-at-wiley.

67. Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles / Peroni, Silvio [at al.] // PeerJ Computer Science. 2017. No. 3. Article no. e132. https://doi.org/10.7717/peerj-cs.132.

68. PubCSS. URL: https://github.com/thomaspark/pubcss.

69. dokieli. URL: https://dokie.li/.

70. *Capadisli S., Guy A., Verborgh R., Lange C., Auer S., Berners-Lee T.* Decentralised authoring, annotations and notifications for a read-write web with dokieli // Proceedings of the 17th international conference on web engineering. Cham. 2017. Springer. P. 469–481. https://doi.org/10.1007/978-3-319-60131-1_33.

71. RASH Framework. URL: https://rash-framework.github.io/

72. *Spinaci G., Peroni S., Di Iorio A., Poggi F., Vitali F.* The RASH JavaScript Editor (RAJE): A Wordprocessor for Writing Web-first Scholarly Articles // Proceedings of the 2017 ACM Symposium on Document Engineering. 2017 (DocEng 2017). P. 85–94. https://doi.org/10.1145/3103010.3103018

METHODS AND TOOLS USED FOR PREPARATION SCIENTIFIC ARTICLES PUBLICATIONS IN HTML FORMAT

R. Y. Skornyakova[0000-0001-7372-3574]

Keldysh Institute of Applied Mathematics (Russian Academy of Sciences)

rimmaskorn@gmail.com

Abstract

Along with the traditional form of electronic presentation of full texts scientific articles – the PDF format, the HTML format has become increasingly widespread in recent years. It has a number of advantages for online publications due to the available means for better content structuring, adding multimedia and implementing of various interactive and dynamic features. In this regard, the task of getting an HTML version of a scientific article from the original format sent by the author becomes highly topical.

The article discusses various approaches to preparing HTML versions of full texts scientific articles and describes the software used in this process. The main attention is paid to the tools used for source materials in the Word format.

The paper also outlines the basics of the JATS XML standard, which is widely used in the preparation of online publications of journal articles.

***Keywords**: HTML version of a scientific article, XML version of a scientific article, standard for the exchange of scientific articles, JATS, conversion of scientific article formats*

REFERENCES

1. *Chebukov D.E.* Ob HTML versii polnogo teksta nauchnoj stat'I // Trudy XX Vserossiiskoi nauchnoi konferentsii «Nauchnyi servis v seti Internet», g. Novorossiisk, 17–22 sentiabria 2018 g. M.: IPM im. M.V. Keldysha: 2018. S. 487–498.
URL: https://keldysh.ru/abrau/2018/theses/16.pdf, doi:10.20948/abrau-2018-16.

2. Animaciya i video v nauchnoj publikacii / *M.M. Gorbunov-Posadov [i dr.]* // Preprinty IPM im. M.V. Keldysha. 2014. № 104. 32 s.
URL: https://library.keldysh.ru/preprint.asp?id=2014-104.

3. *Kitaev E.L., Skornyakova R.Yu.* Skrejping «na letu» vneshnih veb-resursov, upravlyaemyj razmetkoj HTML-stranicy // Preprinty IPM im. M.V. Keldysha. 2019. № 20. 31 s. https://doi.org/10.20948/prepr-2019-20
URL: https://library.keldysh.ru/preprint.asp?id=2019-20.

4. *Gorbunov-Posadov M.M.* Zhivaia publikatsiia // Otkrytye sistemy. 2011. № 4. S. 48–49. URL: https://keldysh.ru/gorbunov/live.htm.

5. *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.
URL: https://library.keldysh.ru/preprint.asp?id=2017-82 doi:10.20948/prepr-2017-82.

6. *Aalbersberg I.J.* PDF versus HTML – which do researchers prefer? // Elsevier connect. 9 Jul 2013.
URL: https://www.elsevier.com/connect/pdf-versus-html-which-do-researchers-prefer.

7. *Kasdorf W.E.* The XML revolution // Learned Publishing. 2001. Vol. 14, No. 3. P. 223–231. https://doi.org/10.1087/095315101750240485.

8. *Young D., Madans P.* XML: Why Bother? // Publishing Research Quarterly. 2009. No. 25. P. 147–153. https://doi.org/10.1007/s12109-009-9120-4.

9. *Rech D.A.* Instituting an XML-First Workflow // Publishing Research Quarterly. 2012. No. 28. P. 192–196. https://doi.org/10.1007/s12109-012-9278-z.

10. *Kasdorf W.E.* The Columbia Guide to Digital Publishing. NYC: Columbia University Press, 2003. 816 p.

11. *Murray-Rust P., Rzepa H.S.* Scientific publications in XML – towards a global knowledge base // Data Science. 2002. No. 1. P. 84–98. https://doi.org/10.2481/dsj.1.84.

12. Standardized Markup for Journal Articles: Journal Article Tag Suite (JATS) // NISO website, 07.07.2021. URL: https://www.niso.org/standards-committees/jats.

13. *Lapeyre D.A.* Introduction to JATS (Journal Article Tag Suite) // XML.com. 12.10.2018. URL: https://www.xml.com/articles/2018/10/12/introduction-jats/.

14. *Usdin B.T., Lapeyre D.A.* JATS/BITS/NISO STS // Proceedings of the Symposium on Markup Vocabulary Ecosystems. Balisage Series on Markup Technologies, vol. 22 (2018), Washington, DC, USA, 30.07.2018. https://doi.org/10.4242/BalisageVol22.Usdin01.

15. *Beck J.* NISO Z39.96 The Journal Article Tag Suite (JATS): What Happened to the NLM DTDs? // The Journal of Electronic Publishing. 2011. Vol. 14, issue 1. https://doi.org/10.3998/3336451.0014.106.

16. *Donohoe P., Sherman J., Mistry A.* The Long Road to JATS // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 21–22, 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK279831/.

17. *Gafurova P. O., Elizarov A. M., Lipachev E. K.* Bazovye servisy fabriki metadannyh cifrovoj matematicheskoj biblioteki Lobachevskii-DML // Elektronnye biblioteki. 2020. T. 23. № 3. S. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

18. *Lizzi V.* Improving JATS for multilingual articles // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, May 3–4, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK579699/.

19. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3 // National Center for Biotechnology Information.
URL: https://jats.nlm.nih.gov/archiving/tag-library/1.3/chapter/set-intro.html.

20. Journal Publishing Tag Library NISO JATS Version 1.3 // National Center for Biotechnology Information.
URL: https://jats.nlm.nih.gov/publishing/tag-library/1.3/chapter/journal-tag-set-intro.html.

21. Article Authoring Tag Library NISO JATS Version 1.3 // National Center for Biotechnology Information.
URL: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.3/chapter/set-intro.html.

22. Journal Article Tag Suite // National Center for Biotechnology Information URL: https://jats.nlm.nih.gov/.

23. JATS4R (JATS for Reuse). URL: https://jats4r.org/.

24. *Kasdorf W.E.* Getting from Word to JATS XML // The Association of Learned and Professional Society Publishers blog. 18.10.2018.
URL: https://blog.alpsp.org/2018/10/getting-from-word-to-jats-xml.html.

25. *Adam L.R., Perera C.* eXtyles, Typefi, and the NLM Journal Publishing DTD // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, November 1–2, 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK47080/.

26. *Piez W.* HTML First?: Testing an alternative approach to producing JATS from arbitrary (unconstrained or "wild") .docx (WordML) format // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 25–26, 2017.
URL: https://www.ncbi.nlm.nih.gov/books/NBK425546/.

27. Oxygen XML Editor. URL: https://www.oxygenxml.com/.

28. Fonto Editor. URL: https://www.fontoxml.com/.

29. LiXuid Manuscript.
URL: https://www.ariessys.com/blog/introduction-lixuid-manuscript-xml/.

30. XEditor. URL: https://www.xpublisher.com/products/xeditor.

31. Texture JATS XML editor. URL: https://github.com/substance/texture.

32. *Garnett A., Aufreiter M., Buchtala O., Alperin J. P.* Introducing Texture: An Open Source WYSIWYG Javascript Editor for JATS // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 25–26, 2017.

URL: https://www.ncbi.nlm.nih.gov/books/NBK425544/.

33. Libero Editor. URL: https://gitlab.coko.foundation/libero/editor.

34. PMC XML Validator // National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/pmc/tools/xmlchecker/.

35. JATS4R Validator // JATS4R, NISO Working Group. URL: https://validator.jats4r.org/.

36. PMC Style Checker // National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/pmc/tools/stylechecker/.

37. ScienceCentral Style Checker // The Korean Federation of Science and Technology Societies. URL: https://www.e-sciencecentral.org/tools/stylechecker/.

38. JATS Preview Stylesheets // GitHub.com. URL: https://github.com/ncbi/JATSPreviewStylesheets.

39. *Piez W.* Fitting the Journal Publishing 3.0 Preview Stylesheets to Your Needs: Capabilities and Customizations // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, November 1–2, 2010. URL: https://www.ncbi.nlm.nih.gov/books/NBK47104/.

40. *Graham T.* Formatting JATS: as easy as 1-2-3 // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 1–2, 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK189779/.

41. PeerJ/jats-conversion: *Conversion and validation for JATS XML* // GitHub.com URL: https://github.com/PeerJ/jats-conversion.

42. Seeing through the eLife Lens: A new way to view research // Inside eLife, Jun 6, 2013. URL: https://elifesciences.org/inside-elife/0414db99/seeing-through-the-elife-lens-a-new-way-to-view-research.

43. Lens // GitHub.com URL: https://github.com/elifesciences/lens.

44. Inera JATS Solutions. URL: https://www.inera.com/jats-solutions/.

45. Inera eXtyles Arc. URL: https://www.inera.com/extyles-arc/.

46. Typefi: Aut*omated publishing for print, online, and mobile.* URL: https://www.typefi.com/products-services/.

47. Q&A: End-to-end automation with eXtyles Arc and Typefi. URL: https://www.typefi.com/qa-end-to-end-automation-with-extyles-arc-and-typefi/.

48. *Eve M.P.* The Means of (Re-)Production: Expertise, Open Tools, Standards and Communication // Publications. 2014. No. 2. P. 38–43. https://doi.org/10.3390/publications2010038.

49. Ictect Intelligent Content for Journals. URL: https://www.ictect.com/JATS-XML.

50. SciSpace JATS XML Converter. URL: https://typeset.io/for-publishers/jats-xml/.

51. Pandoc. URL: https://pandoc.org/.

52. Transpect. An Open Source framework for converting and checking data. URL: https://transpect.github.io/.

53. *Ahmetov D.Yu., Elizarov A.M., Lipachev E.K.* Servis-orientirovannaya informacionnaya sistema nauchnogo zhurnala «Elektronnye biblioteki» // Elektronnye biblioteki. 2016. T. 19, № 1. S. 2-39. URL: https://rdl-journal.ru/article/view/377/468.

54. *Galyavieva M.S., Elizarov A.M., Lipachev E.K.* Cifrovaya infrastruktura elektronnogo nauchnogo zhurnala: avtomatizaciya redakcionno-izdatel'skih processov i sistema servisov // Elektronnye biblioteki. 2016. T. 19, № 5. S. 408–465. URL: https://rdl-journal.ru/article/view/404/489.

55. meTypeset. URL: https://github.com/withanage/meTypeset.

56. *Garnett A., Alperin J.P., Willinsky J.* The Public Knowledge Project XML Publishing Service and meTypeset: Don't call it "Yet Another Word-to-JATS Conversion Kit" // Journal Article Tag Suite Conference (JATS-Con) Proceedings 2015 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK279666/.

57. DocxToJats. URL: https://github.com/Vitaliy-1/docxToJats.

58. *Ekanger A., Odu O.* How we tried to JATS XML // Ravnetrykk. 2020. No. 39. P. 156–162. https://doi.org/10.7557/15.5517.

59. 13 Best Free Word to HTML Converter Software for Windows. URL: https://listoffreeware.com/free-word-to-html-converter-software-windows/.

60. XMLmind Word To XML: Convert DOCX to unstyled, valid, “semantic” XHTML *1.0, 1.1 or 5.0.* URL: https://xmlmind.com/w2x/docx_to_xhtml.html.

61. Doc Converter Pro. URL: https://docconverter.pro/.

62. XSweet. URL: https://xsweet.org/.

63. Open XML PowerTools.
URL: https://github.com/OpenXmlDev/Open-Xml-PowerTools/.

64. Opensagres XDocReport. URL: https://github.com/opensagres/xdocreport.

65. Mammoth. URL: https://mike.zwobble.org/projects/mammoth/.

66. *Siegman T., Young B.* HTML-First at Wiley // BookNet Canada blog. 14.02.2018.
URL: https://www.booknetcanada.ca/blog/2018/2/14/html-first-at-wiley.

67. Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles / Peroni, Silvio [at al.] // PeerJ Computer Science. 2017. No. 3. Article no. e132. https://doi.org/10.7717/peerj-cs.132.

68. PubCSS. URL: https://github.com/thomaspark/pubcss.

69. dokieli. URL: https://dokie.li/.

70. *Capadisli S., Guy A., Verborgh R., Lange C., Auer S., Berners-Lee T.* Decentralised authoring, annotations and notifications for a read-write web with dokieli // Proceedings of the 17th international conference on web engineering. Cham. 2017. Springer. P. 469–481. https://doi.org/10.1007/978-3-319-60131-1_33.

71. RASH Framework. URL: https://rash-framework.github.io/

72. *Spinaci G., Peroni S., Di Iorio A., Poggi F., Vitali F.* The RASH JavaScript Editor (RAJE): A Wordprocessor for Writing Web-first Scholarly Articles // Proceedings of the 2017 ACM Symposium on Document Engineering. 2017 (DocEng 2017). P. 85–94. https://doi.org/10.1145/3103010.3103018

СВЕДЕНИЯ ОБ АВТОРЕ

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

ORCID: 0000-0001-7372-3574

Материал поступил в редакцию 3 апреля 2023 года