

УДК 004.91+004.4

ПРЕПРИНТЫ ИПМ ИМ. М. В. КЕЛДЫША: КОНВЕРТАЦИЯ ИЗ MS WORD В HTML

А. А. Воробьев¹ [0000-0002-6849-8867], Р. Ю. Скорнякова² [0000-0001-7372-3574]

^{1, 2}Институт прикладной математики им. М. В. Келдыша РАН, г. Москва, Россия

¹voraa@yandex.ru, ²rirmaskorn@keldysh.ru

Аннотация

В последние годы широкое распространение получило представление полных текстов научных статей в формате HTML, обладающем для онлайн-публикаций рядом преимуществ по сравнению с традиционно используемым форматом PDF за счет имеющихся в HTML более развитых средств для структуризации материала, встраивания мультимедийного контента и реализации разного рода интерактивных и динамических возможностей. В связи с этим актуальной становится задача преобразования рукописей из традиционно используемых авторами форматов MS Word и LaTeX в полноценную HTML-версию, способную реализовать преимущества такого формата. В работе представлены результаты применения к препринтам ИПМ им. М. В. Келдыша подхода к конвертации научных статей из формата MS Word в HTML, предложенного в предыдущих работах. Описаны интерактивные возможности полученных HTML-версий.

Ключевые слова: HTML-версия научной статьи, преобразование научных статей из формата .docx в .html, препринты ИПМ им. М.В. Келдыша, JATS XML.

ВВЕДЕНИЕ

В последние годы онлайн-формат стал доминирующим для научных журналов, обеспечивая более эффективное распространение научной информации. Современные электронные журналы предоставляют полные тексты статей на условиях открытого доступа или на коммерческой основе. Традиционно полные тексты научных статей публикуются в формате PDF, однако значительная часть научных журналов, в особенности журналы с открытым доступом, предоставляют полные тексты также и в формате HTML, более подходящем для представления ин-

формации в веб-среде. В работах [1–3] подробно изложены преимущества и недостатки обоих форматов для научных публикаций. Основные преимущества формата HTML заключаются:

- в лучшей структуризации материала, что позволяет быстро ориентироваться в нем и находить нужный контент;
 - в возможности адаптации под различные размеры экрана;
 - в возможности автоматического перевода на другие языки, предоставляемой браузерами;
 - в наличии форматов масштабируемого представления формул, пригодных для машинной обработки и поиска;
- и, самое существенное,
- в лучших средствах для встраивания мультимедийного контента и расширения функционала разного рода интерактивными и динамическими возможностями.

Преимущество формата PDF состоит в неизменности макета и форматирования, он удобнее для чтения офлайн, обмена содержимым статей и печати. Ни один из форматов — ни PDF, ни HTML — на текущий момент не обладает абсолютным преимуществом перед другим, поэтому издательства стараются предоставлять контент научных статей в обоих форматах.

Если для получения PDF-версии статьи из рукописи, представленной в формате часто используемого авторами редактора MS Word, имеются стандартные средства, то для получения полноценной HTML-версии, раскрывающей преимущества этого формата, таких стандартных средств нет.

Для реализации преимуществ формата структура HTML-документа должна отражать структуру научной статьи. В нем должны быть выделены такие семантические единицы статьи, как заглавие, авторы, аннотация, разделы и подразделы с заголовками, библиографические ссылки, объединенные в список, формулировки и доказательства теорем, строчные и блочные формулы и т. п. Должны присутствовать также элементы-контейнеры для иллюстративного материала: рисунков с подписями, таблиц с заголовками и пр. Выделение структурных элементов позволяет организовать единый дизайн для всех публикаций журнала и служит основой для реализации динамических возможностей.

В формате HTML такая структура создается с помощью вложенных друг в друга элементов и назначенных им классов и атрибутов, тогда как в формате MS Word полноценные возможности для описания подобных структур отсутствуют. Частично семантику научного документа в редакторе MS Word можно выразить с помощью пользовательских стилей, таких как «Автор», «Аннотация», «Библиографическая ссылка» и т. п. Однако возможности для реализации многоуровневых иерархических структур в Word ограничены. Можно с помощью пользовательского стиля «Формула» пометить абзац как соответствующий блочной формуле, но указать, что этот абзац входит в формулировку теоремы, с помощью стиля уже нельзя, поскольку к каждому абзацу может быть применен только один стиль. Из-за этого ограничения алгоритм преобразования научной статьи из формата MS Word в документ HTML, структурированный должным образом, становится непростым.

Дополнительные усилия требуются, если формулы в документе Word представлены в формате MTEF редактора формул MathType [4], часто используемого авторами. Для обеспечения машинной обработки формул и возможности эффективного поиска научных статей по математическим выражениям формулы в HTML-документе должны быть представлены либо в формате MathML [5], либо в формате TeX [6] / LaTeX [7]. Поэтому необходим специальный конвертер, обеспечивающий перевод из двоичного формата MTEF в один из указанных форматов.

В работах [8, 9] описаны методы и программные инструменты, используемые издательствами для получения HTML-версии научной статьи из рукописи в формате MS Word. Доминирующим подходом является так называемый подход XML First, предполагающий предварительное создание XML-версии статьи, отражающей ее структуру (чаще всего в соответствии со стандартом NISO JATS [10]), с последующим преобразованием в форматы HTML и PDF. Основное преимущество представления научной статьи в XML-формате состоит в отделении контента от его визуального представления, что упрощает хранение статей, обмен ими и их преобразование в различные форматы. XML-версии полных текстов могут храниться в базе данных, а HTML-версии генерироваться динамически по запросу, что позволяет беспрепятственно изменять макет HTML-документа.

Однако получение XML-версии научной статьи из рукописи в формате MS Word является непростой задачей. Причины те же, что и при создании надлежащим образом структурированной HTML-версии: отсутствие в формате MS Word полноценных средств для выражения семантики научного документа и наличие формул в двоичном формате. Существующие конвертеры либо непригодны для обработки статей со сложным содержимым, либо требуют значительных финансовых затрат. К последней категории относятся широко используемые (в основном крупными издательствами) продукты eXtyles [11] компании Inera¹, основанные на применении пользовательских стилей MS Word, а также продукты [13] компании Ictect, использующие технологии искусственного интеллекта. Предложено также решение [14], в котором HTML служит промежуточным форматом при преобразовании текста научной статьи в формат JATS XML. Этот подход предполагает первоначальное преобразование документа Word в HTML с сохранением базового форматирования и последующее внесение семантики научного документа с помощью Word-подобного HTML-редактора. После этого производится преобразование HTML-документа в формат JATS XML. Однако конвертер [15], применяемый в этом решении, не поддерживает формулы в формате MTEF редактора MathType, что делает невозможным его использование для препринтов ИПМ, поскольку именно этот редактор формул преимущественно используется нашими авторами.

В итоге, после анализа рынка программных продуктов для создания HTML-версии научной статьи было принято решение разработать собственный конвертер, который был бы применим к препринтам ИПМ им. М. В. Келдыша и пригоден для использования в условиях небольшой редакции с ограниченными ресурсами. К настоящему моменту создана альфа-версия конвертера. В работе представлены результаты ее применения.

¹ Компания Wiley Partner Solutions, приобретая компанию Inera, объявила о прекращении разработки и поддержки eXtyles как лицензионного продукта в августе 2026 г. Вместо этого она будет предоставлять услуги по конвертации научных статей из формата MS Word в формат JATS XML. На замену продукту Inera eXtyles компания Typefi, долгое время тесно сотрудничавшая с компанией Inera, предлагает новый продукт Typefi Orion [12].

РЕАЛИЗАЦИЯ КОНВЕРТЕРА

В работах [16, 17] описан подход, выбранный нами для реализации конвертера. Преобразование в формат HTML производится после создания PDF-документа. Так же, как и в продуктах Inera eXtyles [11], мы используем пользовательские стили, но преобразуем рукопись в формате MS Word не в формат JATS XML, а в HTML-документ, структура научной статьи в котором выделяется с помощью классов и атрибутов, соответствующих элементам и атрибутам JATS. Для разметки семантическими стилями текст рукописи копируется в пустой документ MS Word, созданный на основе специального шаблона. Отдельный документ Word необходим, поскольку семантическая разметка изменяет исходное форматирование. На первом этапе реализации такого подхода HTML-документ, полученный из размеченного стилями документа Word, рассматривается как конечный. В дальнейшем, аналогично подходу [14], предполагаются создание конвертера из формата HTML в формат JATS XML и организация работы с базой данных для хранения полученных XML-версий.

При преобразовании рукописи из формата MS Word в формат HTML проще выявлять ошибки применения пользовательских стилей Word, чем при непосредственном преобразовании в JATS XML, поскольку семантика JATS, реализованная в HTML-документе посредством классов и атрибутов, легко визуализируется в браузере при помощи каскадных таблиц стилей (CSS). В этом, на наш взгляд, состоит преимущество данного подхода по сравнению с подходом [11]. Преимущество нашего подхода по сравнению с подходом [14], где семантика научного документа вносится при помощи специализированного HTML-редактора, состоит в том, что сотрудник редакции остается работать в привычной для него среде MS Word и не должен осваивать новые инструменты. К недостаткам можно отнести необходимость параллельно вносить изменения в два документа Word – исходный и размеченный стилями, если потребуются коррекция текста после публикации.

В основе конвертера препринтов ИПМ из формата MS Word в формат HTML лежит конвертер Mammoth [18], разработанный английским программистом Майклом Уильямсоном. Он позволяет настраивать преобразование с помощью таблицы соответствия стилей: абзацу, единым образом отформатированному

фрагменту текста (run), а также таблице с определенным стилем можно поставить в соответствие элемент или набор вложенных друг в друга элементов HTML с определенными классами и атрибутами. Тем самым создается основа для вне-сения в HTML-документ семантики, отражающей структуру научной статьи.

Конвертер Mammoth не поддерживает формулы – для преобразования формул, содержащихся в документе MS Word в формате MTEF, нами был создан отдельный конвертер, основанный на библиотеке MathType SDK, который каждую формулу формата MTEF, содержащуюся в документе Word, преобразует в формат MathML и записывает в текстовый файл с именем, соответствующим ее порядковому номеру. Полученные файлы затем используются конвертером из формата .docx в формат HTML: при преобразовании из формата .docx MathML-код формулы считывается из файла и вставляется в нужное место HTML-документа.

ФОРМАТИРОВАНИЕ СТРУКТУРНЫХ ЭЛЕМЕНТОВ ПРЕПРИНТОВ ИПМ ИМ. М. В. КЕЛДЫША И ИХ РАЗМЕТКА СТИЛЯМИ В MS WORD

Препринты ИПМ имеют довольно сложный и разнообразный контент. Характерным является наличие большого числа сложных формул, а также рисунков и графиков, объединенных в группы. В табл. 1 представлены структурные элементы препринтов, встретившиеся нам на данный момент, и соответствующие им элементы и атрибуты JATS.

Табл. 1. Структурные элементы препринтов ИПМ им. М. В. Келдыша.

Структурный элемент препринта	Элемент JATS	Атрибут JATS
Титульная часть	<front>	
Заглавие	<article-title>	
Автор	<contrib>	contrib-type="author"
Аннотация	<abstract>	
Ключевые слова	<kwd-group>	
Заглавие на английском	<trans-title>	xml:lang="en"
Автор на английском	<contrib>	contrib-type="author" xml:lang="en"
Аннотации на английском	<trans-abstract>	xml:lang="en"
Ключевые слова на английском	<kwd-group>	xml:lang="en"
Сведения о финансировании	<funding-statement>	

Основное содержание	<body>	
Раздел	<sec>	
Заголовок (раздела, таблицы, рисунка...)	<title>	
Абзац	<p>	
Аббревиатура	<abbr>	
Внутристроковая формула	<inline-formula>	
Отдельно стоящая (блочная) формула (может включать метку – номер формулы)	<disp-formula>	
Метка (номер) – может относиться к разделу, формуле, таблице, рисунку, библиографической ссылке и др.	<label>	
Группа формул (включает несколько формул, может иметь свою метку)	<disp-formula-group>	
Программный код	<code>	
Внутритекстовая ссылка (на формулу, таблицу, элемент библиографии...)	<xref>	ref-type – тип ссылки (на что ссылается)
Список	<list>	
Элемент списка	<list-item>	
Контейнер для таблицы (включает контейнер для текстового описания и саму таблицу)	<table-wrap>	
Контейнер для текстового описания объекта (включает метку и заголовок)	<caption>	
Таблица	<table>	
Шапка таблицы	<thead>	
Ячейка в шапке таблицы	<th>	
Строка таблицы	<tr>	
Ячейка таблицы	<td>	
Контейнер для изображения (включает изображение и текстовое описание – подрисуночную подпись)	<fig>	

Изображение	<graphic>	
Контейнер для содержимого вне основного потока текста (используется для небольших рисунков, которые обтекает текст)	<boxed-text>	
Контейнер для группы изображений (включает набор контейнеров для изображений и текстовое описание группы – подпись под группой)	<fig-group>	
Формулировка теоремы	<statement>	content-type = "Theorem"
Формулировка леммы	<statement>	content-type = "Lemma"
Доказательство	<statement>	content-type = "Proof"
Формулировка условия	<statement>	content-type = "Case"
Отрицательный результат	<named-content>	content-type = "Fail"
Справочная часть	<back>	
Группа приложений	<app-group>	
Приложение	<app>	
Группа сносок (включает отдельные сноски)	<fn-group>	
Сноска	<fn>	
Глоссарий (включает список сокращений)	<glossary>	
Список сокращений	<def-list>	list-content = "abbreviations"
Библиографический список (включает заголовок и отдельные библиографические ссылки)	<ref-list>	
Библиографическая ссылка (включает метки и текст ссылки)	<ref>	
Текст библиографической ссылки	<mixed-citation>	

Для распознавания этих структурных элементов в документе MS Word используются 42 стиля абзацев и символов и 3 стиля таблиц; 40 из используемых стилей абзацев и символов представлены в галерее стилей (рис. 1); не представлены только два: «Текст сноски», который применяется по умолчанию при

вставке сноски, и «MTDisplayEquation», используемый для блочных формул плагином MathType. Стили таблиц применяются в тех случаях, когда таблицы использовались авторами для форматирования формул, групп формул и групп рисунков, чтобы отличать форматизирующие таблицы от собственно таблиц. Число стилей увеличилось по сравнению с предложенным в работе [16], поскольку добавилась поддержка новых структурных элементов, таких как теорема, условие, программный код и др.



Рис. 1. Галерея стилей абзацев и символов.

Авторы препринтов ИПМ, как правило, не используют встроенные в MS Word средства для автоматической нумерации объектов и создания перекрестных ссылок, поскольку текст автоматически созданной ссылки начинается с прописной буквы, а простая замена прописной буквы на строчную не срабатывает – нужны специальные усилия, чтобы заменить прописную букву строчной. Кроме того, конвертер Mammoth, лежащий в основе нашего конвертера, хотя и декларирует поддержку перекрестных ссылок, на практике выдает некорректный результат. Поэтому разработанный нами конвертер не предполагает использование автоматической нумерации. Для оформления нумерации объектов и перекрестных ссылок в шаблоне MS Word предусмотрены специальные символные стили. Номера (метки) формул, рисунков, таблиц, теорем и других объектов оформляются символьным стилем «Label», а для перекрестных ссылок используются различные символные стили в зависимости от типа объекта, на который

направлена ссылка. Например, для оформления ссылки на рисунок применяется стиль «(рис. N)», а для оформления ссылки на теорему используется стиль «(теор N)». Чтобы ссылки были визуально различимы в документе Word, стили перекрестных ссылок разного типа имеют разные цвета. На внешнем виде PDF-версии препринта это не отражается, поскольку разметка препринта для преобразования в HTML происходит в копии исходного документа Word после получения PDF-документа.

Разноцветные стили используются и для оформления форматированных таблиц. В исходном документе ячейки форматированных таблиц не имеют границ. Мы применяем к форматированным таблицам стили с цветными границами, чтобы разметка была заметна. Пример такой разметки приведен на рис. 2: группы формул оформлены табличным стилем с зелеными границами, одиночные формулы – табличным стилем с синими границами.

Под группой изображений мы понимаем набор изображений, имеющих общий числовой номер. Как правило, такая группа оформляется авторами как таблица с чередующимися строками: в нечетных строках помещаются изображения, в четных – метки с подписями или только метки, или, наоборот, в нечетных строках – метки (возможно, с подписями), в четных – изображения. В последней строке иногда располагается подпись ко всей группе рисунков. Для таких таблиц мы применяем табличный стиль с оранжевыми границами (рис. 3). При конвертации в HTML таблица такого формата преобразуется в набор фигур, вложенных в обрамляющую фигуру-контейнер.

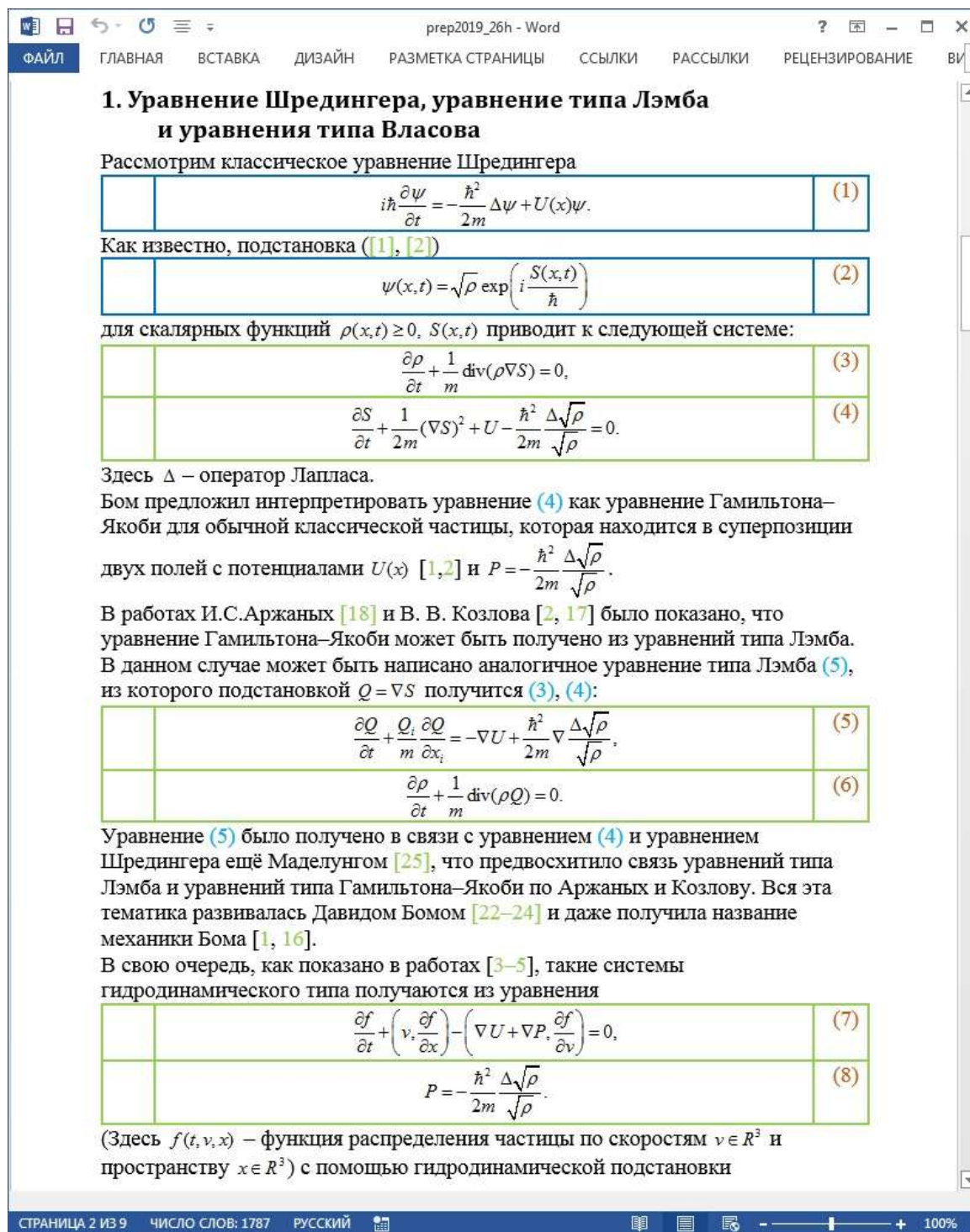


Рис. 2. Пример разметки страницы.

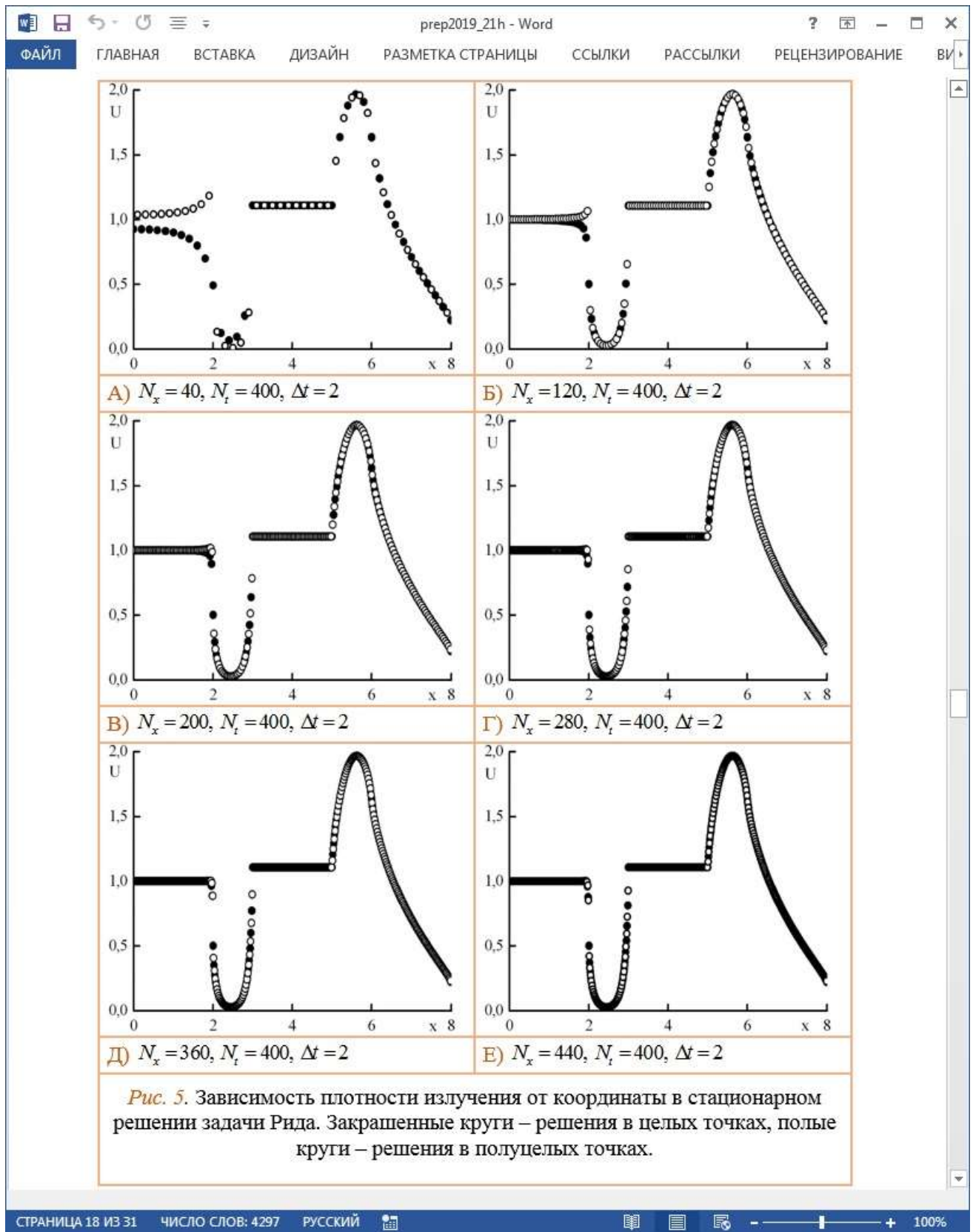


Рис. 3. Группа изображений, оформленная как таблица.

ПРЕОБРАЗОВАНИЕ МАТЕМАТИЧЕСКИХ ФОРМУЛ

Для представления математических формул мы используем формат MathML (Mathematics Markup Language), рекомендуемый консорциумом W3C (World Wide Web Consortium) в качестве стандарта разметки математических выражений в веб-документах. Для визуализации формул используется JavaScript-библиотека MathJax [19] версии 3.2.2.

Авторы препринтов ИПМ, присылающие рукописи в формате MS Word, как правило, вводят формулы с помощью WYSIWYG редактора MathType, внутренний формат для хранения формул которого – MTEF – является двоичным. Разработанный нами конвертер формул извлекает формулы в формате MTEF из документа MS Word и с помощью средств MathType SDK преобразует их в формат MathML, задействуя трансляторы, входящие в поставку редактора MathType. Используется версия 6.9 редактора MathType – последняя версия с бессрочной лицензией².

Так же, как и авторы работы [20], в которой подробно изложены вопросы конвертации математических формул из документа MS Word, мы столкнулись с ошибками в трансляторах MathType. Часть ошибок являются синтаксическими. Библиотека MathJax отображает места таких ошибок красным цветом (рис. 4), поэтому они легко обнаруживаются. Но бывают и семантические ошибки, которые выявляются только при сличении HTML-версии с исходным текстом (рис. 6). Встречаются также отклонения от исходного форматирования, на которые обращают внимание авторы препринтов (рис. 5). Часть ошибок и отклонений от форматирования мы правим вручную, часть – программным путем при конвертации исходного текста в формат HTML после вставки формул в HTML-документ.

Примером ошибочного преобразования, приводящего к синтаксической ошибке, является преобразование математического выражения с нижним (верхним) индексом, относящимся к выражению в скобках (рис. 4).

² Более поздние версии нам недоступны из-за действующих санкций

$$\left\{ \begin{array}{l} \frac{\partial \bar{u}_i}{\partial t} + \frac{1}{h}(W_{i+1} - W_i) + \overline{\kappa_{a,i} u_i} = \overline{Q_{0i}}, \\ \frac{\partial \bar{w}_i}{\partial t} + \frac{1}{h}(\text{msub} - \text{msub}) + \overline{\kappa_{t,i} w_i} = \overline{Q_{1i}}. \end{array} \right. \quad \left\{ \begin{array}{l} \frac{\partial \bar{u}_i}{\partial t} + \frac{1}{h}(W_{i+1} - W_i) + \overline{\kappa_{a,i} u_i} = \overline{Q_{0i}}, \\ \frac{\partial \bar{w}_i}{\partial t} + \frac{1}{h}((DU)_{i+1} - (DU)_i) + \overline{\kappa_{t,i} w_i} = \overline{Q_{1i}}. \end{array} \right.$$

Рис. 4. Формула с синтаксической ошибкой до и после правки MathML-кода программным путем.

Такое выражение транслятор MathType преобразует в элемент <msub> (<msup>), внутри которого содержится больше двух элементов, как в коде ниже, что является синтаксической ошибкой.

```
<msub>
  <mo stretchy="false">(</mo>
  <mi>D</mi>
  <mi>U</mi>
  <mo stretchy="false">)</mo>
  <mrow>
    <mi>i</mi>
    <mo>+</mo>
    <mn>1</mn>
  </mrow>
</msub>
```

Мы исправляем эту ошибку программным путем: включаем выражение со скобками в элемент <mrow>, как в коде ниже.

```
<msub>
  <mrow>
    <mo stretchy="false">(</mo>
    <mi>D</mi>
    <mi>U</mi>
    <mo stretchy="false">)</mo>
  </mrow>
  <mrow>
    <mi>i</mi>
    <mo>+</mo>
    <mn>1</mn>
  </mrow>
</msub>
```

В качестве примера отклонения от форматирования можно привести преобразование оператора косой дробной черты. Когда такая черта располагается между выражениями, содержащими интегралы или суммы, она становится плохо различимой из-за того, что ее размер не масштабируется под высоту окружающих выражений (рис. 5). Мы включили в конвертер программный код, добавляющий

ко всем операторам дробной черты атрибут `stretchy="true"`, после чего дробная черта во всех выражениях стала вытягиваться в соответствии с размерами окружающих символов.

$$\bar{D} = \sum_k c_k \mu_k^2 \bar{I}_k / \sum_k c_k \bar{I}_k. \quad \bar{D} = \sum_k c_k \mu_k^2 \bar{I}_k / \sum_k c_k \bar{I}_k.$$

Рис. 5. Формула с косой дробной чертой до и после правки MathML-кода программным путем.

Примером ошибки, потребовавшей ручной правки, является преобразование равенства двух матриц разного размера, соответствие частей которых определялось при помощи горизонтальных и вертикальных линий. Результирующее MathML-выражение не содержало кода для линий, его пришлось вставлять вручную (рис. 6).

$$\begin{array}{c} \mathbf{c} \quad \mathbf{A} \\ \mathbf{b}^T \end{array} = \begin{array}{ccc} 1 & 1 & \\ 1/3 & 0 & 1/3 \\ 1 & -1/12 & 3/4 & 1/3 \\ & -1/12 & 3/4 & 1/3 \end{array} \quad \begin{array}{c} \mathbf{c} \quad \mathbf{A} \\ \mathbf{b}^T \end{array} = \begin{array}{ccc|ccc} 1 & 1 & & & & \\ 1/3 & 0 & 1/3 & & & \\ 1 & -1/12 & 3/4 & 1/3 & & \\ & -1/12 & 3/4 & 1/3 & & \end{array}$$

Рис. 6. Равенство матриц до и после ручной правки MathML-кода.

Возможно, в более поздних версиях MathType подобные ошибки трансляторов устранены, и, когда последние версии MathType будут нам доступны, необходимость в корректирующем коде и ручной правке формул отпадет, но на данном этапе выявление ошибок и ручная правка занимают значительную часть времени подготовки HTML-версии препринта.

ИНТЕРАКТИВНЫЕ ВОЗМОЖНОСТИ HTML-ВЕРСИЙ ПРЕПРИНТОВ

Выделение структурных элементов препринта в документе HTML позволило с помощью языка JavaScript реализовать интерактивные возможности, создающие удобства для читателя, которых нет в PDF-версии:

- доступное из разных мест оглавление (рис. 7), дающее читателю возможность быстро переключаться между разделами;

- внутритекстовые гиперссылки на формулы, таблицы, рисунки, сноски, теоремы, библиографию, разделы;
- увеличение рисунков и формул по щелчку мыши;
- разного рода всплывающие подсказки, возникающие при наведении мыши на внутритекстовую гиперссылку.

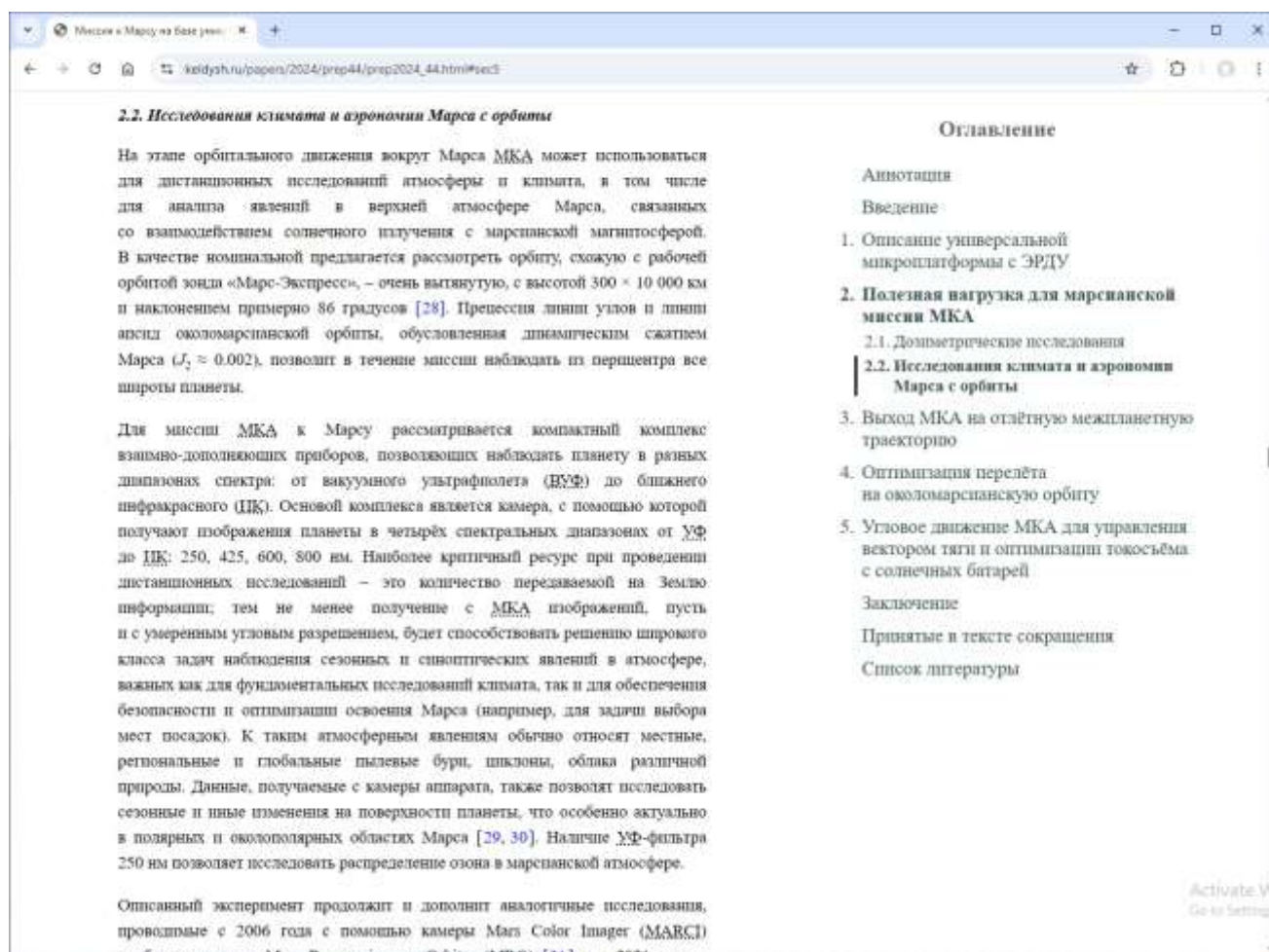


Рис. 7. Пример оглавления.

В всплывающих подсказках мы показываем полностью не только библиографические ссылки и сноски (как это принято во многих онлайн-журналах), но и содержимое формул (групп формул) (рис. 8), формулировок теорем (рис. 9) и условий (рис. 10), рисунков. Таблицы, как правило, занимают много места, поэтому для них, а также для разделов в подсказках отображаются только заголовки.

Численное моделирование

keldysh.ru/papers/2019/prep22/prep2019_22.html

Препринты ИПМ им.М.В.Келдыша

$$\frac{\partial}{\partial t} \{m (S_v S_w \rho_w + (1 - S_v) \rho_v \beta_w)\} + \text{div} [\rho_w \vec{V}_w] + q_w = 0, \quad (1)$$

$$\frac{\partial}{\partial t} \{m (S_v (1 - S_w) \rho_g + (1 - S_v) \rho_v (1 - \beta_w))\} + \text{div} [\rho_g \vec{V}_g] + q_g = 0. \quad (2)$$

$$\vec{V}_w = - \frac{k k_{rw}}{\mu_w} (\nabla P - g \rho_w \vec{k}), \quad (3)$$

$$\vec{V}_g = - \frac{k k_{rg}}{\mu_g} (\nabla P - g \rho_g \vec{k}), \quad (4)$$

$$\frac{\partial}{\partial t} \{m (S_v (S_w \rho_w c_w + (1 - S_w) \rho_g c_g) + (1 - S_v) \rho_v c_v) + (1 - m) \rho_s c_s\} + \text{div} \{ \rho_w \varepsilon_w \vec{V} + \rho_g \varepsilon_g \vec{V}_g + P (\vec{V}_w + \vec{V}_g) \} + \text{div} \vec{W} + q_c = 0. \quad (5)$$

$$\vec{W} = - \{m (S_v (S_w \lambda_w + (1 - S_w) \lambda_g) + (1 - S_v) \lambda_v) + (1 - m) \lambda_s\} \nabla T. \quad (6)$$

$$T = A \ln P + B. \quad (7)$$

того режима
стоянному, н.
торсионные
мятся к нулю.
всей области
енно газового
температуры

с графиками
от трехфазной
н с переходом

от системы уравнений (1)–(7) с температурой, являющейся функцией давления, к системе (9)–(12), в которой температура является независимой переменной. Более детально это показано на рис. 6, где представлена зависимость температуры от давления в разные моменты времени. На этих графиках выделяются два участка, один из которых соответствует зависимости (7), а другой – случаю отсутствия этой функциональной зависимости.

Рис. 4. Распределение температуры для моментов времени 500, 1000, 2000, 6000, 9000, 12000 с.

https://keldysh.ru/papers/2019/prep22/prep2019_22.html#fm12

Рис. 8. Всплывающая подсказка для системы уравнений.

Миссия к Марсу на базе уни... x Математическая модель расч... x +

← → ↻ 🏠 Not secure 195.209.147.157:81/papers.new/2024/prep45/prep2024_45.html ☆ 🗑️ 🔄 ⋮

Препринты ИПМ им.М.В.Келдыша ☰

...к решению краевой
...охладе для уравнения

Теорема 2. 1) $G_\varphi(r, s)$ непрерывна в квадрате $[r_0, r_1]^2$.

2) При фиксированном $s \in [r_0, r_1]$ функция Грина $G_\varphi(r, s)$ как функция от r удовлетворяет однородным граничным условиям $G_\varphi(r_0, s) = G_\varphi(r_1, s) = 0$ и её столбцы являются на $[r_0, s]$ и $(s, r_1]$ решениями однородного уравнения (а) из (36).

3) Всюду в $[r_0, r_1]^2$ существуют непрерывные частные производные $\frac{\partial G_\varphi}{\partial r}$, $\frac{\partial^2 G_\varphi}{\partial r^2}$, которые с треугольников Δ_\pm непрерывно продолжаются на главную диагональ, причём для продолжений имеет место формула

$$\frac{\partial G_\varphi}{\partial r}(r+0, r) - \frac{\partial G_\varphi}{\partial r}(r-0, r) = \frac{1}{r} A^{-1}, \quad r_0 < r < r_1.$$

4) Для функции (40) имеют место формулы

$$u'_\varphi(r) = \int_{r_0}^{r_1} \frac{\partial G_\varphi}{\partial r} g(s) ds, \quad u''_\varphi(r) = \frac{1}{r} A^{-1} g(r) + \int_{r_0}^{r_1} \frac{\partial^2 G_\varphi}{\partial r^2} g(s) ds, \quad (41)$$

причём подынтегральные функции суммируемы.

5) Отображение $C_2[r_0, r_1] \rightarrow C_2^2[r_0, r_1]$, $g(s) \rightarrow u_\varphi(r)$, задаваемое формулой, является биекцией, обратная к которой определяется формулой $u_\varphi \rightarrow A_1(D)u_\varphi = g$.

...для первого метода
... $n = 0, 1, 2, \dots$, (48)
...что из теорем 1, 2
...е, $u_z^{(n+1)}$, $u_\varphi^{(n+1)}$,
...обозначим норму в
...дно, $\|\cdot\|_\infty \leq \|\cdot\|_\omega$.
... $n \geq 1$,
... $n \geq 1$.
... $C_2^2[r_0, r_1]$, то

из [теоремы 2](#) следуют неравенства

$$\|u_z^{(n+1)} - u_z^{(n)}\|_\omega \leq |\Phi_*| \cdot \|A_2(D)^{-1}B\| \cdot \|u_\varphi^{(n)} - u_\varphi^{(n-1)}\|_\infty \leq q_2 \|u_\varphi^{(n)} - u_\varphi^{(n-1)}\|_\omega, \quad n \geq 1,$$

Аналогично выводится неравенство

$$\|u_\varphi^{(n+1)} - u_\varphi^{(n)}\|_\omega \leq q_1 \|u_z^{(n)} - u_z^{(n-1)}\|_\omega, \quad n \geq 1,$$

где $q_i = |\Phi_*| \cdot \|A_i(D)^{-1}B\|$, $i = 1, 2$.

Обозначим $x_n = \|u_z^{(n)} - u_z^{(n-1)}\|_\omega$, $y_n = \|u_\varphi^{(n)} - u_\varphi^{(n-1)}\|_\omega$, $n \geq 1$.

195.209.147.157:81/papers.new/2024/prep45/prep2024_45.html#theor2

Рис. 9. Всплывающая подсказка для формулировки теоремы.

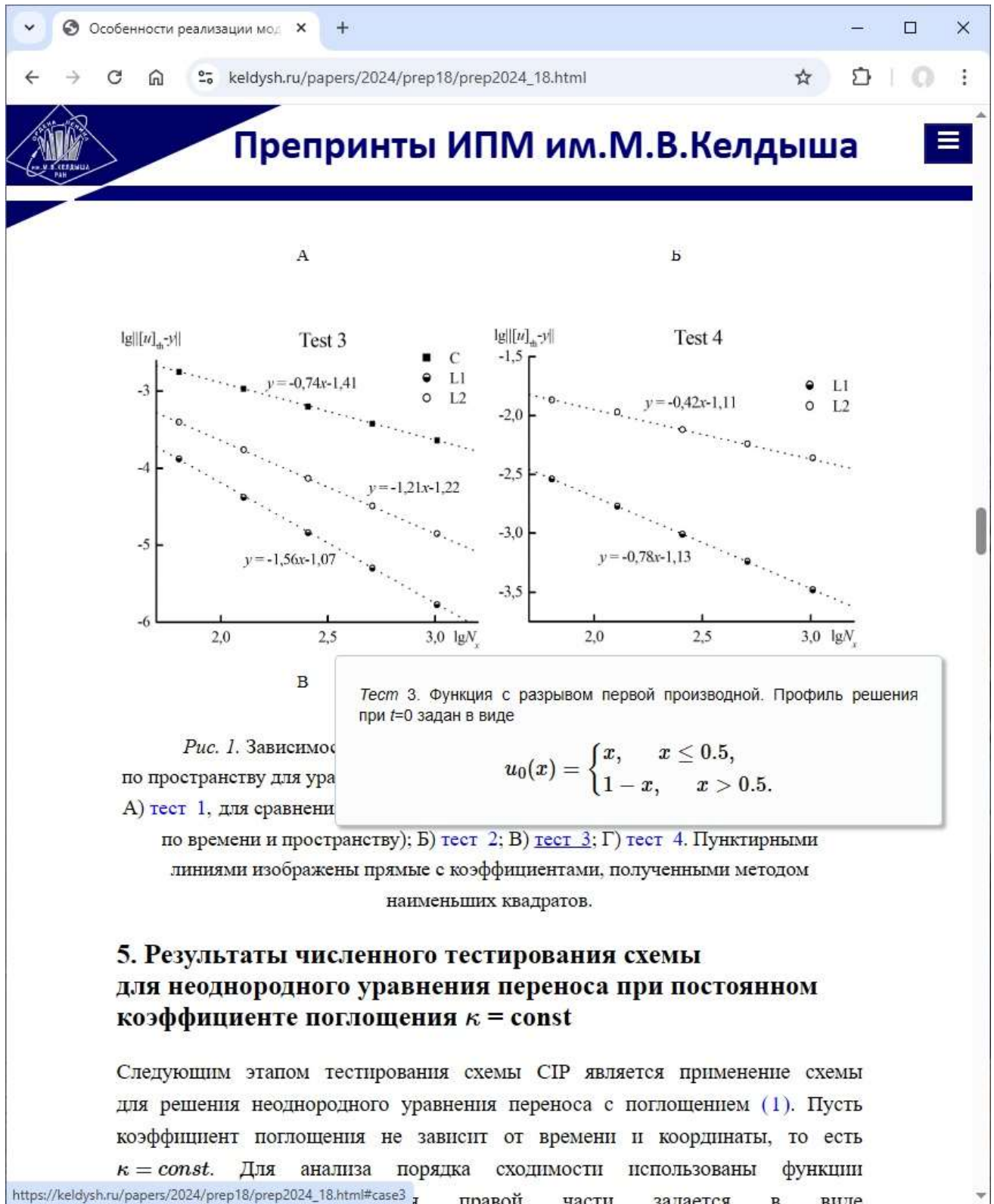


Рис. 10. Всплывающая подсказка для условия теста.

На наш взгляд, использование всплывающих подсказок для нумерованных математических формул и разного рода формальных утверждений позволяет существенно сократить время чтения, поскольку отпадает необходимость перемещаться по содержимому препринта для уточнения, о чем идет речь. Всплывающие подсказки для этих объектов, на наш взгляд, удобнее, чем демонстрация содержимого объекта в соседней колонке. Подсказка легко скрывается небольшим движением мыши, и при необходимости обращения к оглавлению не требуется дополнительных действий.

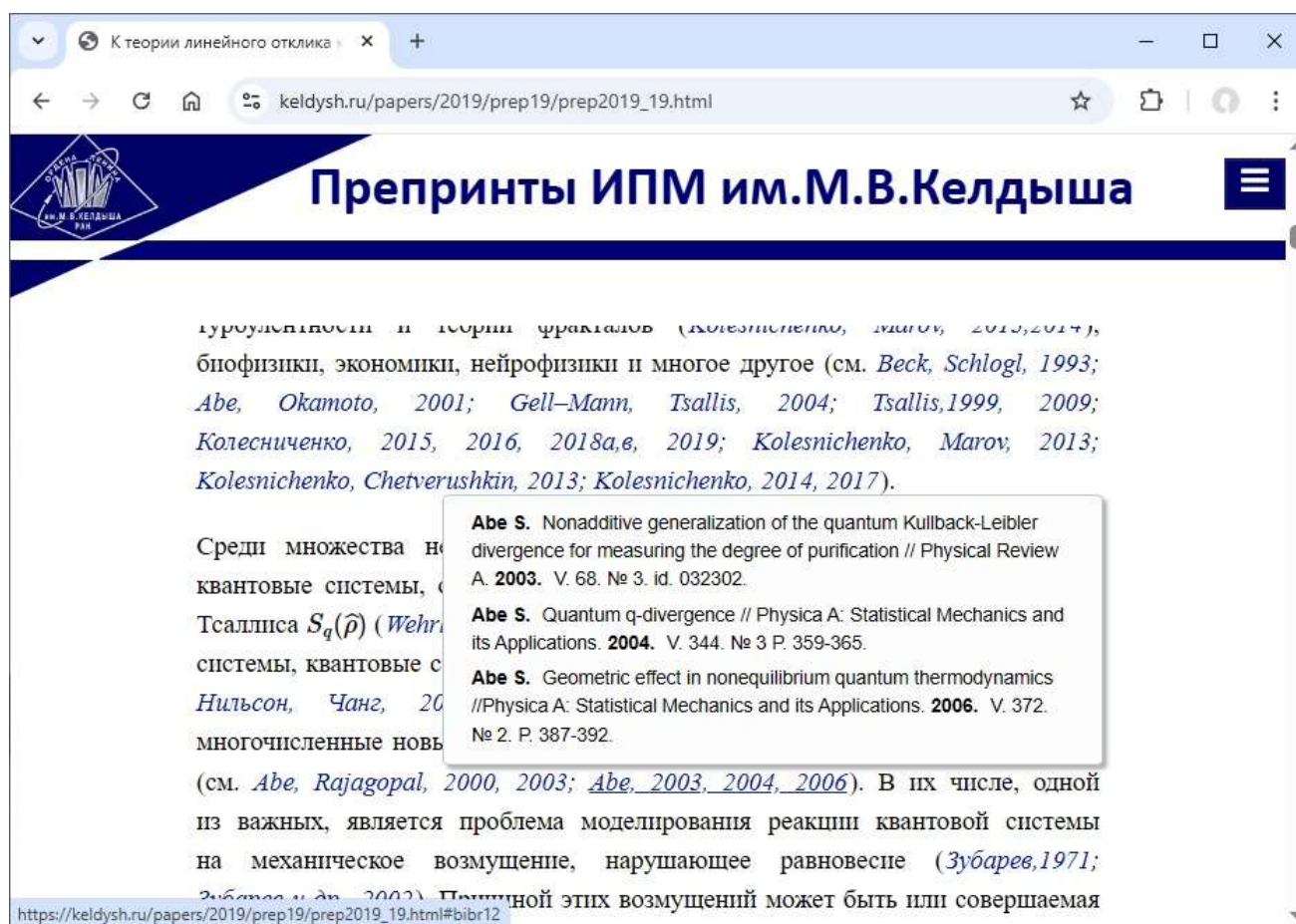


Рис. 11. Всплывающая подсказка для библиографических ссылок в гарвардском стиле.

Хотя для авторов существуют шаблон и инструкция [21] по оформлению препринтов ИПМ в редакторе MS Word, эти материалы носят рекомендательный характер, и им, к сожалению, следуют не все авторы. В частности, не всегда используется рекомендуемый ванкуверский стиль оформления библиографических

ссылок: некоторые авторы используют гарвардский стиль или его подобие. Для таких стилей мы реализовали гиперссылки на библиографию и всплывающие подсказки (рис. 11) так же, как и для ванкуверского стиля.

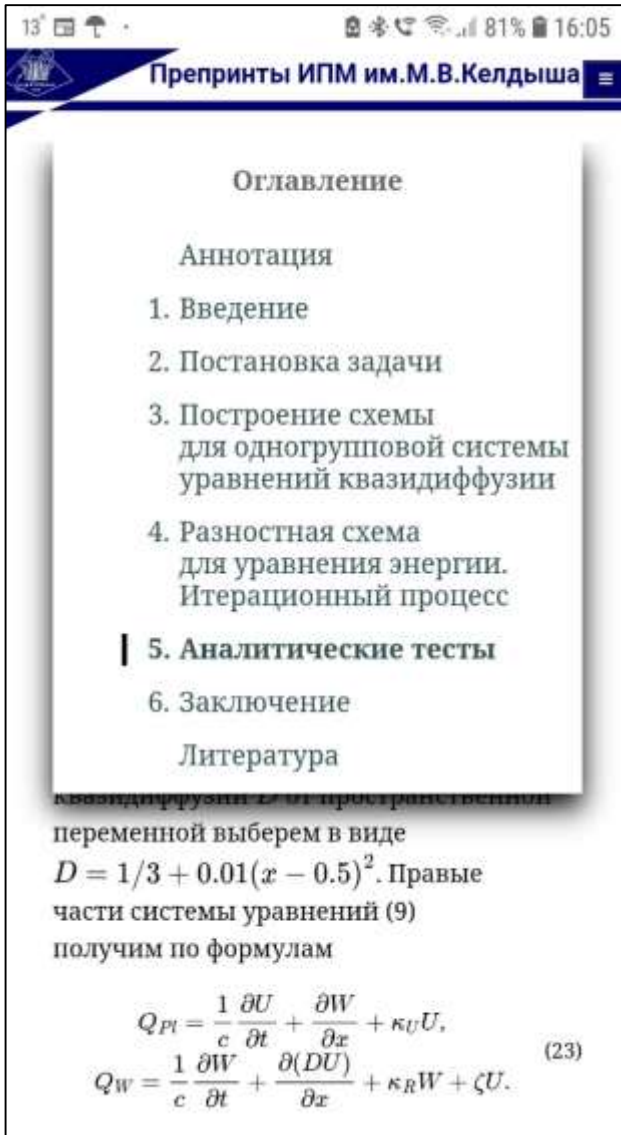


Рис. 12. Оглавление в мобильной версии.

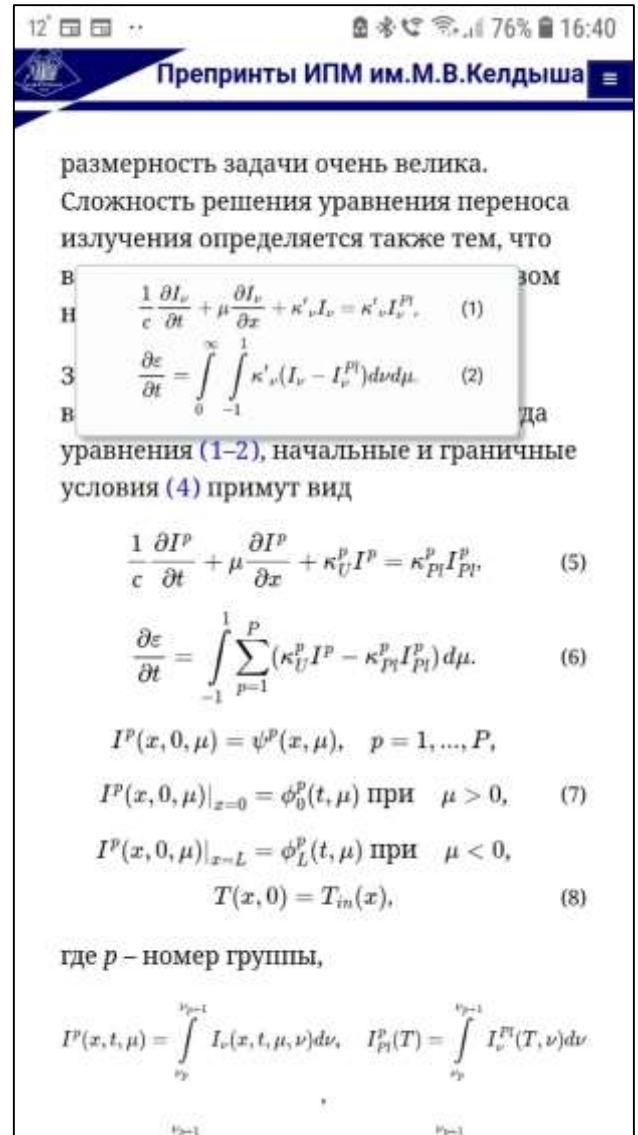


Рис. 13. Всплывающая подсказка в мобильной версии.

Описанные интерактивные возможности реализованы и в мобильной версии: оглавление появляется при нажатии на кнопку «гамбургер» (рис. 12), а всплывающая подсказка – при длительном нажатии на внутритекстовую гиперссылку (рис. 13).

Примеры HTML-версий препринтов можно посмотреть на сайте Института. Веб-адреса препринтов с указанием особенностей их визуального представления и реализованных интерактивных возможностей представлены в табл. 2.

Табл. 2. HTML-версии препринтов ИПМ на сайте ИПМ им. М. В. Келдыша

URL	Особенности визуализации, удобства для читателя
https://keldysh.ru/papers/2018/rep7/rep2018_7.html	Всплывающие подсказки для сносок и формул, в т. ч. по интервалу номеров. Двухуровневое оглавление
https://keldysh.ru/papers/2019/rep19/rep2019_19.html	Всплывающие подсказки для сносок, формул и библиографических ссылок в нестандартном стиле (подобии гарвардского)
https://keldysh.ru/papers/2019/rep21/rep2019_21.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров. Масштабирование рисунков
https://keldysh.ru/papers/2019/rep22/rep2019_22.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров. Масштабирование рисунков
https://keldysh.ru/papers/2019/rep26/rep2019_26.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров
https://keldysh.ru/papers/2020/rep121/rep2020_121.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров, и вариантов условий. Масштабирование рисунков
https://keldysh.ru/papers/2023/rep25/rep2023_25.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров. Масштабирование рисунков и формул
https://keldysh.ru/papers/2024/rep18/rep2024_18.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров, для отдельных рисунков внутри группы и для условий тестов. Масштабирование рисунков и формул

https://keldysh.ru/papers/2024/rep37/rep2024_37.html	Всплывающие подсказки для формул. Масштабирование рисунков
https://keldysh.ru/papers/2024/rep44/rep2024_44.html	Всплывающие подсказки для сокращений и сносок. Выделение неудачных результатов красным цветом. Рисунок, обтекаемый текстом. Двухуровневое оглавление. Масштабирование рисунков
https://keldysh.ru/papers/2024/rep64/rep2024_64.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров. Гиперссылки как на группу рисунков, так и на отдельные рисунки в группе, с соответствующими всплывающими подсказками. Двухуровневое оглавление. Масштабирование рисунков и формул
https://keldysh.ru/papers/2024/rep65/rep2024_65.html	Всплывающие подсказки для формул, в т. ч. по интервалу номеров, и условий тестов. Гиперссылки как на группу рисунков, так и на отдельные рисунки в группе, с соответствующими всплывающими подсказками. Масштабирование рисунков и формул
https://keldysh.ru/papers/2025/rep71/rep2025_71.html	Наличие приложений. Масштабирование рисунков
https://keldysh.ru/papers/2025/rep75/rep2025_75.html	Подсветка синтаксиса программного кода. Двухуровневое оглавление. Масштабирование рисунков

ЗАКЛЮЧЕНИЕ

В ходе настоящей работы опробован подход к получению HTML-версии научной статьи из рукописи в формате MS Word, предложенный ранее. Разработан конвертер, реализующий этот подход для препринтов ИПМ им. М. В. Келдыша. Апробация показала, что подход может быть использован для конвертации рукописей научных статей со сложным содержимым, включающим математические формулы в формате MathType, группы взаимосвязанных рисунков, формулировки условий и утверждений различного типа.

HTML-версии, создаваемые конвертером, отражают структуру научной статьи, что позволяет с помощью JavaScript-кода реализовать интерактивные элементы, расширяющие функционал электронной публикации по сравнению с традиционным PDF. К числу таких элементов относятся постоянно доступное оглавление, ускоряющее навигацию, и всплывающие подсказки для формул и утверждений, облегчающие восприятие материала.

Основные проблемы возникли при конвертации математических формул. Выявленные ошибки в работе трансляторов редактора формул MathType (версия 6.9) потребовали написания дополнительного корректирующего программного кода и, в ряде случаев, ручной правки сгенерированного кода MathML.

В настоящее время проект перешел в стадию практической реализации. С августа 2025 г. введена в опытную эксплуатацию альфа-версия конвертера. На сегодняшний день конвертировано 17 препринтов, из которых 14 размещены на сайте Института, еще 3 препринта требуют согласования с авторами. Планируем в течение года поставить конвертацию на поток, дорабатывая конвертер и включая по мере необходимости новые структурные элементы. В планах также включение обработки мультимедийного контента, доработка модуля конвертации из формата HTML в формат JATS XML, а также исправление выявленных ошибок непосредственно в трансляторах MathType.

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. *Горбунов-Посадов М.М.* Что дает формат HTML научной публикации // Труды 5-й Международной конференции «Проектирование будущего. Проблемы цифровой реальности», г. Москва, 3–4 февраля 2022 г. М.: ИПМ им. М. В. Келдыша, 2022. С. 216–222. URL: <https://keldysh.ru/future/2022/19.pdf>, <https://doi.org/10.20948/future-2022-19>

3. *Изаак А.Д., Икономов Н.Р., Мисюрин О.Г., Чебуков Д.Е.* Создание интерактивных HTML-версий полных текстов научных статей в российских математических журналах // 13-ая Международная научно-практическая конференция «Научное издание международного уровня — 2025: тенденции развития», г. Москва, 20–23 мая 2025 г.

URL: <https://www.mathnet.ru/php/seminars.phtml?presentid=46360>.

4. *Mathtype Equation Editor.* URL: <https://www.wiris.com/en/mathtype/>.

5. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Веб-технологии для математика: основы MathML: практическое руководство. М.: Физматлит, 2010. 192 с.

6. *Кнут Д.Э.* Все про TEX = The TeXbook. М.: Вильямс, 2003. 560 с.

7. *Lamport L.* LaTeX: a document preparation system. New York: Addison-Wesley Publishing Company, Inc., 1994. 273 с.

8. *Скорнякова Р.Ю.* Методы и инструменты, используемые при подготовке публикаций научных статей в формате HTML // Электронные библиотеки. 2023. Т. 26, № 2. С. 252–302. URL: <https://rdl-journal.ru/article/view/774/850>.

9. *Скорнякова Р.Ю.* Обзор программных средств для создания HTML-версии журнальной статьи из исходного материала в формате Word // Научный сервис в сети Интернет: труды XXV Всероссийской научной конференции (18–21 сентября 2023 г., онлайн). М.: ИПМ им. М. В. Келдыша, 2023. С. 332–344.

URL: <https://doi.org/10.20948/abrau-2023-38>.

10. Standardized Markup for Journal Articles: Journal Article Tag Suite (JATS) // NISO website, 31.10.2024. URL: <https://www.niso.org/standards-committees/jats>.

11. Inera JATS Solutions. URL: <https://www.inera.com/jats-solutions/>.

12. Typefi Orion. URL: <https://www.typefi.com/orion/>.

13. Ictect Intelligent Content for Journals.

URL: <https://www.ictect.com/JATS-XML>.

14. *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/>.

15. XSweet. XSweet docx to html extraction and more.

URL: <https://github.com/Coko-Foundation/XSweet>.

16. Скорнякова Р.Ю. Подход к созданию HTML-версии научной статьи из рукописи в формате MS Word для издательства с малым бюджетом // Электронные библиотеки. 2024. Т. 27, № 6. С. 1064–1089.

URL: <https://rdl-journal.ru/article/view/880/929>.

17. Скорнякова Р.Ю. Разработка конвертера препринтов ИПМ из формата .docx в форматы HTML и JATS XML // Научный сервис в сети Интернет: труды XXVI Всероссийской научной конференции (23–25 сентября 2024 г., онлайн). М.: ИПМ им. М. В. Келдыша, 2024. С. 250–263.

URL: <https://doi.org/10.20948/abrau-2024-20>.

18. Mammoth. .docx to HTML converter.

URL: <https://mike.zwobble.org/projects/mammoth/>.

19. MathJax. URL: <https://www.mathjax.org/>.

20. Gebhard C., Rosenblum B. Wrangling Math from Microsoft Word into JATS XML Workflows // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 12–13, 2016.

URL: <https://www.ncbi.nlm.nih.gov/books/NBK350572/>.

21. Шаблон-инструкция для оформления препринта ИПМ в редакторе MS Word. URL: <https://keldysh.ru/preprints/sample.docx>.

KELDYSH INSTITUTE OF APPLIED MATHEMATICS' PREPRINTS: CONVERSION FROM MS WORD TO HTML

A. A. Vorobjov¹[0000-0002-6849-8867], R. Y. Skornyakova²[0000-0001-7372-3574]

^{1,2}*Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow, Russia*

¹voraa@yandex.ru, ²rirmaskorn@keldysh.ru

Abstract

In recent years, the presentation of full-text scientific articles in HTML has become widespread. This format offers several advantages for online publication compared to the traditional PDF format, owing to its more advanced tools for structuring material, embedding multimedia content, and implementing various interactive and dynamic features. Therefore, the task of converting manuscripts from the traditionally used MS Word and LaTeX formats into a high-quality HTML version capable of realizing the advantages of this format has become relevant. This paper presents the results of applying the approach for converting scientific articles from MS Word to HTML, proposed in previous studies, to Keldysh Institute of Applied Mathematics' preprints. The interactive capabilities of the resulting HTML versions are described.

Keywords: *HTML version of a scientific article, conversion of scientific articles from .docx format to html, KIAM preprints, JATS XML.*

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. *Gorbunov-Posadov M.M.* Chto daet format HTML nauchnoi publikatsii // Trudy 5-i Mezhdunarodnoi konferentsii «Proektirovanie budushchego. Problemy tsifrovoi realnosti», g. Moskva, 3–4 fevralia 2022 g. M.: IPM im. M.V. Keldysha, 2022. S. 216–222. URL: <https://keldysh.ru/future/2022/19.pdf>, <https://doi.org/10.20948/future-2022-19>.

3. *Izaak A.D., Ikonomov N.R., Misiurina O.G., Chebukov D.E.* Sozdanie interaktivnykh HTML-versii polnykh tekstov nauchnykh statei v rossiiskikh matematicheskikh zhurnalakh // 13-aia Mezhdunarodnaia nauchno-prakticheskaja konferentsiia «Nauchnoe izdanie mezhdunarodnogo urovnia — 2025: tendentsii razvitiia», g. Moskva, 20–23 maia 2025 g.

URL: <https://www.mathnet.ru/php/seminars.phtml?presentid=46360>.

4. *MathType Equation Editor*. URL: <https://www.wiris.com/en/mathtype/>.

5. *Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Veb-tehnologii dlia matematika: osnovy MathML: prakticheskoe rukovodstvo. M.: Fizmatlit, 2010. 192 s.

6. *Knut D.E.* Vse pro TEX = The TeXbook. M.: Viliams, 2003. 560 s.

7. *Lamport L.* LaTeX: a document preparation system. New York: Addison-Wesley Publishing Company, Inc., 1994. 273 s.

8. *Skorniakova R.Iu.* Metody i instrumenty, ispolzuemye pri podgotovke publikatsii nauchnykh statei v formate HTML // Russian Digital Libraries Journal. 2023. T. 26, № 2. S. 252–302. URL: <https://rdl-journal.ru/article/view/774>.

9. *Skorniakova R.Iu.* Obzor programmnykh sredstv dlia sozdaniia HTML-versii zhurnalnoi stati iz iskhodnogo materiala v formate Word // Nauchnyi servis v seti Internet: trudy XXV Vserossiiskoi nauchnoi konferentsii (18–21 sentiabria 2023 g., onlain): IPM im. M.V. Keldysha: 2023. S. 332–344.

URL: <https://doi.org/10.20948/abrau-2023-38>.

10. Standardized Markup for Journal Articles: Journal Article Tag Suite (JATS) // NISO website, 31.10.2024. URL: <https://www.niso.org/standards-committees/jats>.

11. Inera JATS Solutions. URL: <https://www.inera.com/jats-solutions/>.

12. Typefi Orion. URL: <https://www.typefi.com/orion/>.

13. Ictect Intelligent Content for Journals.

URL: <https://www.ictect.com/JATS-XML>.

14. *Piez W.* 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/>.

15. XSweet. XSweet docx to html extraction and more.

URL: <https://github.com/Coko-Foundation/XSweet>.

16. *Skorniakova R.Iu.* Podkhod k sozdaniuu HTML-versii nauchnoi stati iz rukopisi v formate MS Word dlia izdatelstva s malym biudzhetom // Russian Digital Libraries Journal. 2024. T. 27, № 6. S. 1064–1089. URL: <https://rdl-journal.ru/article/view/880/929>.

17. *Skorniakova R.Iu.* Razrabotka konvertera preprintov IPM iz formata .docx v formaty HTML i JATS XML // Nauchnyi servis v seti Internet: trudy XXVI Vserossiiskoi nauchnoi konferentsii (23–25 sentiabria 2024 g., onlain). M.: IPM im. M.V. Keldysha, 2024. S. 250–263. URL: <https://doi.org/10.20948/abrau-2024-20>.

18. Mammoth. .docx to HTML converter.
URL: <https://mike.zwobble.org/projects/mammoth/>.

19. MathJax. URL: <https://www.mathjax.org/>.

20. *Gebhard C., Rosenblum B.* Wrangling Math from Microsoft Word into JATS XML Workflows // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, April 12–13, 2016.
URL: <https://www.ncbi.nlm.nih.gov/books/NBK350572/>.

21. Shablon-instruktsiia dlia oformleniia preprinta IPM v redaktore MS Word.
URL: <https://keldysh.ru/preprints/sample.docx>.

СВЕДЕНИЯ ОБ АВТОРАХ



ВОРОБЬЕВ Андрей Артурович – старший программист Института прикладной математики им. М. В. Келдыша РАН, специалист в области веб-технологий.

Andrey Arturovich VOROBYOV – Senior Programmer at the Keldysh Institute of Applied Mathematics RAS, specialist in web technologies.

email: voraa@yandex.ru

ORCID: 0000-0002-6849-8867



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@keldysh.ru

ORCID: 0000-0001-7372-3574

Материал поступил в редакцию 23 марта 2026 года