УДК 004.91, 004.4

ПОДХОД К СОЗДАНИЮ HTML-ВЕРСИИ НАУЧНОЙ СТАТЬИ ИЗ РУКОПИСИ В ФОРМАТЕ MS WORD ДЛЯ ИЗДАТЕЛЬСТВА С МАЛЫМ БЮДЖЕТОМ

Р. Ю. Скорнякова[0000-0001-7372-3574]

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

rimmaskorn@gmail.com

Аннотация

Наиболее распространенным подходом к созданию HTML-версии журнальной статьи среди научных издательств является предварительное создание XML-версии статьи в соответствии с NISO стандартом Journal Article Tag Suite (JATS) с дальнейшим автоматическим преобразованием в форматы HTML и PDF. Однако получение XML-версии статьи из рукописи в формате .docx текстового процессора MS Word, часто используемого авторами, при наличии в ней большого числа сложных формул и таблиц является непростой задачей. Имеющиеся программные средства либо не справляются с ней в полном объеме, либо обходятся дорого и не доступны для малых издательств с ограниченным бюджетом.

В настоящей работе предложен подход к созданию HTML-версии журнальной статьи из рукописи в формате .docx, содержащей формулы в формате MathType, который не требует от издательства значительных финансовых и временных затрат, и описан реализованный на данный момент прототип лежащего в основе этого подхода конвертера научных статей из формата .docx в форматы HTML и JATS XML, применимый для препринтов ИПМ им. М.В. Келдыша.

***Ключевые слова**: HTML-версия научной статьи, XML-версия научной статьи, JATS XML, преобразование научных статей из формата .docx в html.*

ВВЕДЕНИЕ

В настоящее время подавляющее большинство научных журналов имеет онлайн-версии и предоставляет полные тексты статей для открытого доступа или на коммерческой основе. Помимо традиционной формы представления полных

текстов – формата PDF – многие издательства публикуют полные версии научных статей в HTML-формате. Каждый из этих форматов имеет свои преимущества и недостатки, подробно изложенные в работах [1, 2]. Основные преимущества HTML-формата

– в лучшей структуризации материала, что позволяет быстрее ориентироваться в нем и находить нужный контент;

– в возможности адаптации под различные размеры экрана;

– в предоставляемой браузерами возможности автоматического перевода на другие языки;

– в наличии форматов масштабируемого представления формул, пригодного для машинной обработки и поиска;

– и, самое существенное, в возможности добавления мультимедийного контента и расширения функционала разного рода интерактивными и динамическими возможностями, такими как всплывающие подсказки с текстом библиографической ссылки, список ссылающихся публикаций, динамически обновляемая дата последней редакции в ссылке на живую публикацию и др.

PDF-формат более удобен для чтения офлайн и обмена содержимым статей. Ни один из форматов PDF или HTML на данном этапе не обладает абсолютным преимуществом перед другим, поэтому издательства стараются предоставлять контент в обоих форматах. В связи с этим весьма актуальной является задача организации процесса получения двух синхронизованных между собой версий научной статьи из материала, присланного автором. Несмотря на то, что история публикаций полных текстов научных статей в HTML-формате насчитывает уже порядка 30 лет, единого подхода к организации этого процесса и доступного для всех инструментария за это время не выработано.

В работе предложен один из возможных вариантов организации процесса получения PDF- и HTML-версий научной статьи из исходного материала в формате текстового процессора MS Word, ориентированный на малые издательства с ограниченным бюджетом, и описан прототип конвертера научных статей из формата .docx в форматы HTML и JATS XML, лежащего в основе такого подхода, применимый к препринтам Института прикладной математики имени М.В. Келдыша РАН (далее ИПМ).

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ СУЩЕСТВУЮЩИХ ПОДХОДОВ И ПРОГРАММНЫХ ИНСТРУМЕНТОВ

В работах [3, 4] описаны применяемые издательствами подходы к созданию синхронизированных между собой PDF- и HTML-версий журнальных статей и программные инструменты, используемые для рукописей в формате текстового процессора MS Word.

Чтобы можно было реализовать преимущества формата HTML для научной статьи, HTML-код должен удовлетворять определенным требованиям. В нем при помощи классов и атрибутов должны быть выделены основные структурные единицы статьи, такие как аннотация, библиографический список, разделы и подразделы и т. п. Внутри основного текста статьи должны быть выделены рисунки с относящимися к ним подписями и описаниями, таблицы с номерами и заголовками, формулы и группы формул с относящимися к ним номерами. Внутри библиографического списка должны быть выделены отдельные библиографические ссылки с метками. Желательно, чтобы были проставлены ссылки из основного текста на элементы библиографии, рисунки, формулы, таблицы, чтобы можно было по ним переходить и реализовывать всплывающие подсказки. Чтобы формулы были пригодны для машинной обработки и поиска, они должны быть представлены в формате MathML или TeX. Подобного рода структуру удобно представлять, используя формат XML. Поэтому вполне естественно, что наиболее распространенным подходом к созданию синхронизированных между собой PDF- и HTML-версий научной статьи стал так называемый подход XML-First, состоящий в предварительном создании XML-версии журнальной статьи в соответствии со стандартом, принятым в издательстве, с последующим автоматическим преобразованием ее в форматы PDF и HTML (Рис. 1).

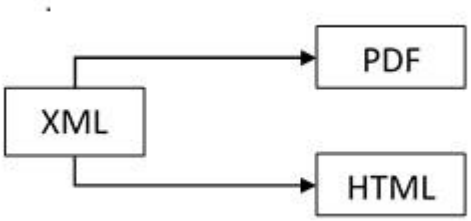

Рис. 1. Подход XML-First

XML-разметка, применяемая в издательствах, отражает структуру статьи и обычно включает деление на титульную часть, содержащую метаданные, тело статьи (ее содержание) и справочную часть, которая содержит библиографический список и может включать приложения, глоссарий, благодарности и т. п. В ней, как правило, предусматриваются отдельные элементы для названия, авторов, аннотации, разделов статьи и их заголовков, рисунков, формул, таблиц, библиографических ссылок и т. д.

Основное преимущество XML-представления статьи состоит в отделении контента от его визуального представления. Это упрощает хранение статей, обмен ими и преобразование в различные форматы. Полные тексты статей в XML-формате можно хранить в базе данных и формировать HTML-представление динамически по запросу, например, при помощи XSL-преобразования. Хранение полных текстов статей в базе данных делает возможным поиск не только по метаданным, но и по содержанию. К преимуществам XML-формата можно отнести также возможность добавления семантики предметных областей за счет включения элементов и атрибутов, отражающих конкретные научные понятия с использованием определенных словарей и онтологий, что способствует продвижению в направлении формализации научного знания.

Изначально разные издательства использовали разные XML-схемы, но необходимость обмена журнальными статьями и хранения их в электронных библиотеках потребовала выработки для этой цели единого стандарта. Такой стандарт был разработан в 2003 году в Национальной медицинской библиотеке США. Первоначально он предназначался для хранения полных текстов статей онлайн-архива PubMed Central, поддерживаемого этой библиотекой, но затем стал использоваться и другими организациями. В процессе эксплуатации выяснилось, что формат удобен не только для хранения и обмена, но и для подготовки научных статей к публикации. После доработки совместно с другими организациями, этот формат, получивший название Journal Article Tag Suite, сокращенно JATS, в 2012 году стал стандартом NISO [5]. К настоящему времени JATS де факто стал международным стандартом. Он используется более чем в 25 странах, в том числе в России.

Стандарт включает в себя три модели, имеющие разные назначения:

– «архивно-обменную» – Journal Archiving and Interchange Tag Set – для обмена журнальными статьями и хранения их в репозиториях, объединяющих статьи из разных изданий;

– «издательскую» – Journal Publishing Tag Set – для разметки статьи, публикуемой в конкретном журнале;

– и «авторскую» – Article Authoring Tag Set – для первоначального ввода контента статьи без привязки к конкретному журналу.

Эти модели в значительной степени совпадают, но имеются и отличия, обусловленные их назначением. Например, в «авторской» модели нет таких элементов, как название журнала, ISSN и т. п.; в «архивно-обменной» модели тело статьи не является обязательным элементом – для хранения и обмена могут быть предназначены только метаданные.

На рис. 2 представлен пример журнальной статьи в формате JATS XML. Корневым элементом является элемент <article>, который включает в себя элементы-контейнеры:

– <front>, представляющий титульную часть статьи и включающий журнальные метаданные (<journal-meta>), такие как название, ISSN, и метаданные статьи (<article-meta>), такие как название, сведения об авторах, аннотация;

– <body>, представляющий основное содержание статьи;

```
<article xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mml="http://www.w3.org/1998/Math/Mat
  <front>
    <journal-meta>...</journal-meta>
    <article-meta>...</article-meta>
  </front>
  <body>
    <sec sec-type="intro">...</sec>
    <sec>
      <title>Related Works</title>
      <p>The growing interest in the publication of Web-first research papers has resulted in
      <p>...</p>
      <p>...</p>
      <table-wrap id="table-1" orientation="landscape">
        <object-id pub-id-type="doi">10.7717/peerjcs.132/table-1</object-id>
        <label>Table 1</label>
        <caption><title>A comparison among existing HTML-oriented formats for scholarly papers
        <alternatives>
          <graphic mimetype="image" mime-subtype="png" xlink:href="https://peerj.com/articles/
          <table>...</table>
        </alternatives>
        <table-wrap-foot>...</table-wrap-foot>
      </table-wrap>
      <sec>...</sec>
      <sec>
        <title>HTML-oriented WYSIWYG editors</title>
        <p>One of the most important and recent proposals, which is compliant with the princip
        <p>Fidus Writer (<ext-link ext-link-type="uri" xlink:href="https://www.fiduswriter.org
        <p>Authorea (<ext-link ext-link-type="uri" xlink:href="https://www.authorea.com">https
      </sec>
    </sec>
    <sec>...</sec>
    <sec>...</sec>
    <sec>...</sec>
    <sec>...</sec>
    <sec sec-type="conclusions">...</sec>
    <sec sec-type="supplementary-m" id="supplemental-in">...</sec>
  </body>
  <back>
    <ack>...</ack>
    <sec sec-type="additional-info">...</sec>
    <ref-list content-type="authoryear">
      <title>References</title>
      <ref id="ref-1">...</ref>
      <ref id="ref-2">...</ref>
      <ref id="ref-3">...</ref>
      <ref id="ref-4">...</ref>
      <ref id="ref-5">...</ref>
      <ref id="ref-6">...</ref>
      <ref id="ref-7">...</ref>
      <ref id="ref-8">...</ref>
```

Рис. 2. Пример представления статьи в формате JATS XML

– <back>, представляющий справочную часть, в которую входит раздел благодарностей (<ack>), раздел с дополнительной информацией о работе (sec-type = “additional-info”), библиографический список (<ref-list>), содержащий отдельные библиографические ссылки (<ref>), а также могут входить глоссарий, приложения и т. п.

Элемент <body> включает элементы <sec>, представляющие разделы статьи, которые в свою очередь включают заголовки (<title>), параграфы <p>, контейнеры таблиц <table-wrap>. Контейнер <table-wrap> включает собственно таблицу (<table>), ее номер (<label>), заголовок (<caption>), а также альтернативное представление таблицы в виде изображения (<graphic>).

JATS XML – гибкий стандарт, обязательных элементов в нем не очень много, можно использовать только необходимое подмножество. Разные издательства и порталы могут иметь свои спецификации JATS, отличающиеся требованиями к наличию тех или иных элементов и атрибутов, их порядку и т. п.

Подход XML-First безусловно имеет много преимуществ – при наличии XML-представления преобразование в другие форматы становится делом техники, однако получение самой XML-версии является непростой задачей, в особенности, если статья имеет сложный контент со множеством формул. Издательствам было бы удобно, если бы авторы присылали статьи, уже набранные в формате JATS XML, однако на практике, несмотря на наличие специализированных XML-редакторов, добиться этого не удается [6]: авторам проще набирать тексты в привычной среде Word или TeX, к тому же статья часто пишется до принятия решения, в каком журнале она будет опубликована, а в разных журналах могут быть свои особенности применения стандарта JATS. Поэтому издательства либо нанимают дополнительной персонал для ручного ввода статьи в формате XML, либо используют коммерческие или разработанные собственными силами конвертеры из исходных форматов в формат XML.

Ручной ввод требует значительных финансовых затрат на оплату дополнительного персонала и довольно много времени на сам ввод текста. Использование программ-конвертеров позволяет сократить временные затраты, однако имеющиеся в свободном доступе конвертеры из формата .docx не справляются с задачей в полном объеме (в них отсутствует возможность конвертации формул,

таблицы конвертируются только в изображения и т. п.), а коммерческие продукты довольно дороги, хотя и обходятся дешевле, чем ручной труд по вводу данных.

Несмотря на то, что структурно документ Word представляет собой zip-архив xml- и медиа-файлов, и на этом основании иногда считают, что его нетрудно преобразовать в другой XML-формат цепочкой XSL-преобразований, полностью автоматических конвертеров произвольных («сырых») документов Word в формат JATS XML не существует. Причина – в отсутствии в исходных xml-файлах семантики, отражающей структуру научной статьи. Для конвертации в формат JATS XML семантика вносится либо путем разметки документа Word специальными стилями, как в Inera eXtyles [7] – самом распространенном инструменте для преобразования документов Word в формат JATS XML, либо программным путем с использованием технологий искусственного интеллекта, как в разработке компании Ictect [8]. В первом случае нужна предварительная работа по разметке исходного документа Word стилями, отражающими семантику научной статьи, во втором – основное время уходит на доработку выходного документа XML, поскольку с помощью технологий искусственного интеллекта точного соответствия структуре JATS получить не удается. Стоит отметить также, что основанные на технологиях искусственного интеллекта конвертеры обучались на англоязычном материале и к русскоязычному материалу не применимы.

Число стилей для семантической разметки и, соответственно, время на подготовку документа к конвертации зависят от требуемой степени детализации JATS. На рис. 3, 4 представлены стили абзацев, используемые в инструменте Inera eXtyles для получения детального JATS XML, удовлетворяющего требованиям архива PubMed Central. Работа по разметке этими стилями требует значительного времени даже при использовании специального диалога, имеющегося в плагине к редактору Word, который предоставляет Inera eXtyles.

Рис. 3. Стили абзацев в Inera eXtyles JATS (Front, Trans, Back)

Рис. 4. Стили абзацев в Inera eXtyles JATS (Body, List, Object)

Несмотря на наличие различных программ-конвертеров и накопленный опыт работы с ними, преобразование текстов научных статей из формата MS Word в формат JATS XML остается довольно трудоемким и/или финансово затратным. Поэтому, помимо подхода XML-First, рассматриваются и другие подходы. Например, авторы работы [9] предлагают использовать HTML как промежуточный формат при преобразовании в JATS XML, мотивируя это, в частности, тем, что семантику в HTML проще визуализировать. Разработанный ими конвертер с открытым исходным кодом xSweet [10], основанный на XSL-преобразованиях, преобразует

«сырые» документы Word в формат HTML с сохранением подробной информации о форматировании, включая тип, размер и цвет шрифта, отступ в абзаце и т. п., так, чтобы внешний вид документа HTML совпадал с внешним видом документа Word. Семантику научной статьи предлагается вносить в документ HTML с помощью Word-подобного HTML-редактора Wax-JATS, встроенного в издательскую платформу Kotahi. Схема предлагаемого рабочего процесса представлена на рис. 5.

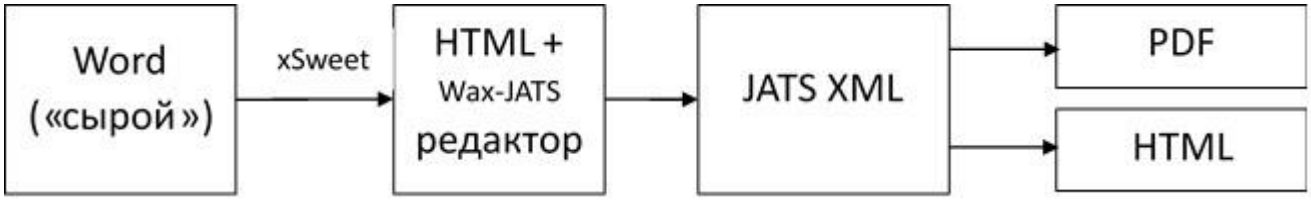

Рис. 5. Схема рабочего процесса с промежуточным HTML

К существенным недостаткам конвертера xSweet следует отнести отсутствие поддержки формул в формате широко используемого редактора формул MathType. Поддерживаются только формулы в формате OMML встроенного в MS Word редактора формул, возможностей у которого существенно меньше, чем у редактора MathType. Недостающие формулы можно только повторно ввести в редакторе Wax-JATS в формате TeX. Поэтому для рукописей, содержащих большое число формул в формате MathType, этот инструмент не подходит.

СХЕМА РАБОЧЕГО ПРОЦЕССА ДЛЯ ИЗДАТЕЛЬСТВА С МАЛЫМ БЮДЖЕТОМ

После анализа рынка программного обеспечения для автоматизации процесса получения HTML-версий научных статей нами было принято решение заняться разработкой собственного конвертера научных статей из формата .docx в HTML и JATS XML, ориентированного на маленькие издательства с ограниченным бюджетом, отказавшись от подхода XML-first из-за трудоёмкости процесса, его реализующего. Основные требования к конвертору:

– конвертор должен преобразовывать формулы в формате MathType в машиночитаемые форматы;

– конвертация не должна требовать длительной предварительной подготовки исходного текста.

Так же, как и разработчики платформы Kotahi, мы считаем, что удобнее сначала конвертировать документ Word в документ формата HTML, а затем уже документ HTML конвертировать в формат JATS XML. Семантику научной статьи можно реализовать в HTML при помощи классов и атрибутов, соответствующих элементам и атрибутам JATS. В формате HTML, в отличие от формата XML, эту семантику можно легко визуализировать, создав, например, отладочный файл CSS, в котором элементы с нужными классами и атрибутами выделяются цветом. Однако, в отличие от подхода, реализуемого в платформе Kotahi, мы предлагаем вносить семантику не непосредственно в документ HTML, а изначально в документ Word, используя специальные стили, аналогично подходу, реализованному в инструменте Inera eXtyles. Эти стили используются при конвертации в формат HTML. При таком подходе работник редакции остается в привычной среде, ему нет необходимости осваивать новые инструменты.

Рабочий процесс предлагается выстроить следующим образом (Рис. 6).

Шаг 1. Присланный автором документ Word редактируется, преобразуется в формат PDF и публикуется так же, как это делалось и ранее, до внедрения HTML-версий.

Шаг 2. Работник редакции копирует содержимое документа Word в пустой документ, созданный на основе специально созданного шаблона, содержащего семантические стили, и производит разметку этими стилями.

Шаг 3. Размеченный семантическими стилями документ Word автоматически преобразуется в формат HTML; с помощью отладочного файла CSS выявляются ошибки семантической разметки документа Word и исправляются. После этого документ Word снова конвертируется в формат HTML.

Шаг 4. В отлаженный документ HTML вручную добавляются метаданные, отсутствующие в исходном документе Word, документ формата HTML конвертируется в формат JATS XML, и оба документа публикуются.

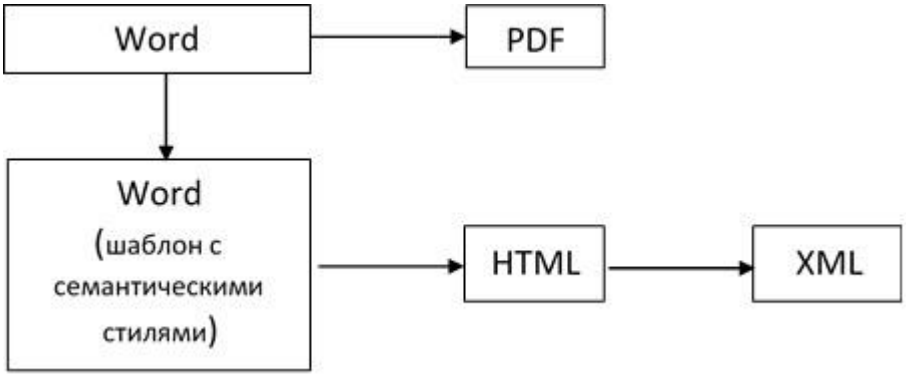

Рис. 6. Схема рабочего процесса для малобюджетного издательства

В этой схеме рабочего процесса полученный документ HTML является окончательным публикуемым документом, а документ формата JATS XML создается лишь с целью предоставления желающим возможности иметь копию статьи, пригодную для машинной обработки. В дальнейшем, когда будет накоплено достаточное число XML-документов, может быть создана база данных. При этом полученный конвертацией из документа Word HTML-документ будет нужен только как промежуточный, а окончательный документ HTML будет создаваться динамически преобразованием из JATS XML.

ПАЛИТРА СТИЛЕЙ И ЭЛЕМЕНТЫ JATS XML

Мы старались минимизировать набор стилей для семантической разметки, чтобы сократить время работы редактора. За основу были взяты стили из шаблона, рекомендованного авторам для оформления препринтов ИПМ, большинство из которых являются семантическими. В этом шаблоне стилей немного, порядка десятка, к ним было добавлено еще два десятка. На данный момент используются 24 стиля абзаца, 5 знаковых стилей (Рис. 7), стиль для сносок и 3 табличных стиля. Табличными стилями помечаются таблицы, используемые для форматирования формул, групп формул и набора рисунков, чтобы отличать их от собственно таблиц.

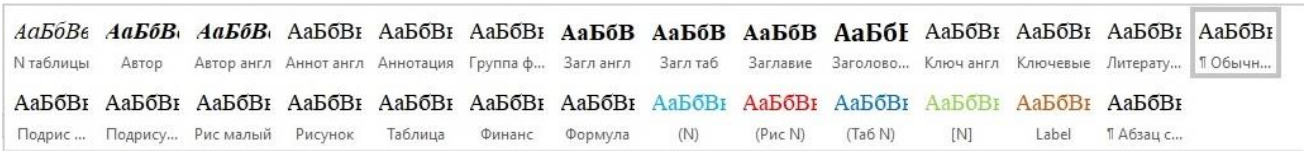

Рис. 7. Стили семантической разметки

Степень детализации JATS XML при таком числе стилей получается не очень высокой. Данный набор стилей позволяет конвертеру выделить следующие элементы:

среди метаданных

- заглавие и его английский вариант;
- авторов и перевод их фамилий и имен (отчеств) на английский;
- аннотацию и ее английский вариант;
- ключевые слова и их английский вариант;
- сведения о финансировании;

в теле статьи

- разделы с заголовками;
- абзацы;
- рисунки с подрисуночной подписью;
- контейнеры таблиц, включающие номер таблицы, заглавие и собственно таблицу;
- формулы с метками;
- группы формул с метками;
- сноски;
- ссылки на литературу, рисунки, таблицы, формулы и группы формул;

в справочной части

- библиографический список с заголовком;
- отдельные библиографические ссылки с метками.

Шаблон JATS XML-файла с соответствующими элементами и атрибутами представлен на рис. 8.

В дальнейшем набор стилей может быть расширен, однако существенное увеличение числа стилей приведет к тому, что в них сложно будет ориентироваться, и понадобится специальный плагин для ускорения процесса разметки, как это сделано в Inera eXtyles.

```
<article article-type="research-article" dtd-version="1.4" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mml="
  <front>
    <article-meta>
      <title-group xml:lang="ru" lang-variant="original" lang-source="author"> <article-title /> </title-group>
      <contrib-group content-type="authors" xml:lang="ru" lang-variant="original" lang-source="author">
        <contrib contrib-type="author" />
      </contrib-group>
      <abstract xml:lang="ru" lang-variant="original" lang-source="author" />
      <kwd-group xml:lang="ru" lang-variant="original" lang-source="author" />
      <title-group xml:lang="en" lang-variant="translation" lang-source="author"> <article-title /> </title-group>
      <contrib-group content-type="authors" xml:lang="en" lang-variant="translation" lang-source="author">
        <contrib contrib-type="author" />
      </contrib-group>
      <abstract xml:lang="en" lang-variant="translation" lang-source="author" />
      <kwd-group xml:lang="en" lang-variant="translation" lang-source="author" />
      <funding-group> <funding-statement /> </funding-group>
    </article-meta>
  </front>
  <body>
    <sec>
      <title />
      <p> <inline-formula /> <xref ref-type="" />  </p>
      <list> <title /> <list-item /> </list>
      <disp-formula>
        <alternatives>
          <graphic />
          <mml:math />
        </alternatives>
        <label />
      </disp-formula>
      <disp-formula-group>
        <disp-formula>...</disp-formula>
        <disp-formula>...</disp-formula>
        <label />
      </disp-formula-group>
      <fig>
        <graphic />
        <caption> <title /></caption>
      </fig>
      <table-wrap>
        <caption> <label /> <title /> </caption>
        <table />
      </table-wrap>
    </sec>
  </body>
  <back>
    <ref-list> <title />
      <ref> <label /> <mixed-citation /> </ref>
    </ref-list>
  </back>
```

Рис. 8. Элементы JATS XML, получаемые с помощью палитры стилей

РЕАЛИЗАЦИЯ КОНВЕРТЕРА

За основу для разработки конвертера научных статей из формата .docx в HTML был взят инструмент с открытым исходным кодом Mammoth [11], написанный английским программистом Майклом Уильямсоном. При преобразовании из

формата .docx в формат HTML конвертер Mammoth использует только информацию о стилях, игнорируя такие детали, как шрифт, размер текста, цвет и т. п., например, встроенный стиль «Заголовок 1» по умолчанию преобразуется в элемент h1. Имеется возможность настройки преобразования при помощи таблицы соответствия стилей: стилю можно поставить в соответствие элемент или набор вложенных друг в друга элементов с классами и атрибутами. Например, абзацу со стилем «Рисунок» можно поставить в соответствие элемент div с классом jats-graphic, последовательности символов со стилем «[N]» – элемент a с классом jats-xref и атрибутом data-jats-ref-type='bibr', который будет работать как ссылка на элемент из списка литературы, и т. д. Тем самым создается возможность внесения в HTML-документ семантики, отражающей структуру научной статьи. Наличие такой возможности явилось определяющим фактором при выборе этого инструмента среди бесплатных конвертеров .docx в HTML. Отметим также, что Mammoth поддерживает основные структурные единицы документа – списки, изображения, таблицы (правда, без форматирования), сноски – и позволяет произвести настройку преобразования изображений путем переопределения метода convertImage.

Конвертер Mammoth имеет реализации на нескольких языках программирования: Python, JavaScript, Java и C#, однако последняя не написана вручную, а получена автоматически из реализации на Java. Код C#, полученный автоматическим преобразованием, не читабелен, его трудно модифицировать вручную, из-за чего теряются преимущества открытого исходного кода. К тому же, обновления C#-реализации производятся с большим опозданием по сравнению с другими реализациями. Поэтому мы от нее отказались. Из реализаций, написанных вручную, для нас более удобной оказалась реализация на Java.

Разрабатываемый нами конвертер на первом шаге преобразует документ Word в формат HTML с помощью конвертера Mammoth, используя таблицу соответствия стилей, представленную на рис. 9, а затем осуществляет автоматическую постобработку для достижения полного соответствия документа HTML нужной структуре. Постобработкой осуществляются разбиение на разделы, удаление таблиц, которые применялись для форматирования формул, объединение рисунков

с их подписями в одном контейнере, расстановка ссылок на литературу, рисунки, таблицы, формулы и т. д.

```
r[style-name='Label']      => span.jats-label
r[style-name='(Рис N)']    => a.jats-xref[data-jats-ref-type='fig'][href='#fig']
r[style-name='(Таб N)']    => a.jats-xref[data-jats-ref-type='table'][href='#tab']
r[style-name='(N)']        => a.jats-xref[data-jats-ref-type='disp-formula'][href='#fml']
r[style-name='[N]']        => a.jats-xref[data-jats-ref-type='bibr'][href='#bibr']

p[style-name='Заглавие']   => section.jats-article-meta > h1.jats-article-title
p[style-name='Автор']      => section.jats-article-meta > ul.jats-contrib-group[data-jats-content-type='authors']
p[style-name='Аннотация']  => section.jats-article-meta > section.jats-abstract > p:fresh
p[style-name='Ключевые']   => section.jats-article-meta > p.jats-kwd-group
p[style-name='Загл англ']  => section.jats-article-meta > h2.jats-trans-title[lang='en']
p[style-name='Автор англ'] => section.jats-article-meta > ul.jats-contrib-group[data-jats-content-type='authors'][lang='en']
p[style-name='Аннот англ'] => section.jats-article-meta > section.jats-trans-abstract[lang='en'] > p:fresh
p[style-name='Ключ англ']  => section.jats-article-meta > p.jats-kwd-group[lang='en']
p[style-name='Финанс']     => section.jats-article-meta > section.jats-funding-group > p.jats-funding-statement:fresh

p[style-name='Heading 1']       => h2.jats-title
p[style-name='Нумерованный 1']  => h2.jats-title

p[style-name='Рисунок']         => div.jats-graphic
p[style-name='Подрисуночный']   => figcaption.jats-caption > p.jats-title
p[style-name='Рис малый']       => p > aside:fresh         > div.jats-graphic
p[style-name='Подрис малый']    => p > aside               > div.jats-title

p[style-name='Формула']         => figure.jats-disp-formula[id='fml']       > div:fresh
p[style-name='Группа формул']   => figure.jats-disp-formula-group[id='fml'] > div:fresh

p[style-name='List Paragraph']  => ul.jats-list > li.jats-list-item:fresh

table[style-name='ТабФорм']     => figure.jats-disp-formula[id='fml'] > table
table[style-name='Таб гр форм'] => figure.jats-disp-formula-group[id='fml'] > table
table[style-name='Таб рис']     => figure.jats-fig[id='fig']:fresh > table.c-graphic

p[style-name='N таблицы']       => figure.jats-table-wrap[id='tab']:fresh > figcaption.jats-caption > p.jats-label
p[style-name='Загл таб']        => figure.jats-table-wrap[id='tab']       > figcaption.jats-caption > p.jats-title
table                           => figure.jats-table-wrap[id='tab']       > table.jats-table
p[style-name='Таблица']         => span.c-table-data

p[style-name='Литература']      => footer.jats-back  > section.jats-ref-list > ol > li.jats-ref[id='bibr']:fresh
p[style-name='Footnote Text']   => p.jats-fn
```

Рис. 9. Таблица соответствия стилей

Конвертер Mammoth не поддерживает формулы, поэтому конвертацию формул из формата MathType пришлось реализовывать самим. Был написан отдельный конвертер формул, основанный на C#-библиотеке MathType SDK, который каждую формулу формата MathType, содержащуюся в документе Word, преобразует в формат MathML и записывает в текстовый файл с именем, соответствующим ее порядковому номеру. Эти файлы используются затем конвертером из формата .docx в формат в HTML: формулы считываются из файлов и записываются в нужные места.

На сегодняшний день реализован прототип конвертера .docx в HTML, который работает с документами Word, при условии соблюдения ряда ограничений:

– подписи к рисункам должны быть расположены под рисунками;

– номера формул должны быть справа от формул;

– номера и заглавия таблиц должны быть сверху от таблиц;

– библиографические ссылки должны либо все иметь метки, либо каждая библиографическая ссылка должна содержаться в одном абзаце;

– не допускаются многоуровневые разделы;

– не допускаются рисунки в формате wmf.

Последние два ограничения в дальнейшем могут быть сняты. Конвертер не поддерживает OLE-объекты, фигуры SmartArt и т. п. Подобные объекты должны быть преобразованы в изображения или таблицы.

ПРИМЕР РАЗМЕТКИ И РЕЗУЛЬТАТ ПРЕОБРАЗОВАНИЯ

На рис. 10 изображен фрагмент препринта ИПМ в редакторе MS Word, а на рисунках 11 и 12 – результат его автоматического преобразования в формат HTML: код и его представление в браузере.

Форматирование группы формул автор препринта выполнил при помощи таблицы. Перед преобразованием в HTML редактор применил к этой таблице специальный табличный стиль, соответствующий группе формул, так что форматирующая таблица оказалась выделенной цветом. Специальные семантические стили были применены к метке с номером формулы, ссылкам на формулы и на литературу. Они также выделены цветом.

В результате автоматического преобразования таблица с группой формул перешла в элемент figure с классом jats-disp-formula-group и атрибутом id="fml5", формулы были преобразованы в MathML, а номер формулы – в элемент span с классом jats-label. Ссылки на формулы и литературу перешли в элементы с классом jats-xref и атрибутами data-jats-ref-type="disp-formula" и data-jats-ref-type="bibr" соответственно, а также атрибутом href, указывающим на соответствующую формулу или элемент библиографического списка. Тем самым автоматически реализована возможность переходить по ссылкам на формулы и литературу. В дальнейшем наличие подобных ссылок на литературу позволит написать JavaScript-код, реализующий всплывающие подсказки с текстом библиографической ссылки при наведении «мыши» на ее номер внутри основного текста препринта.

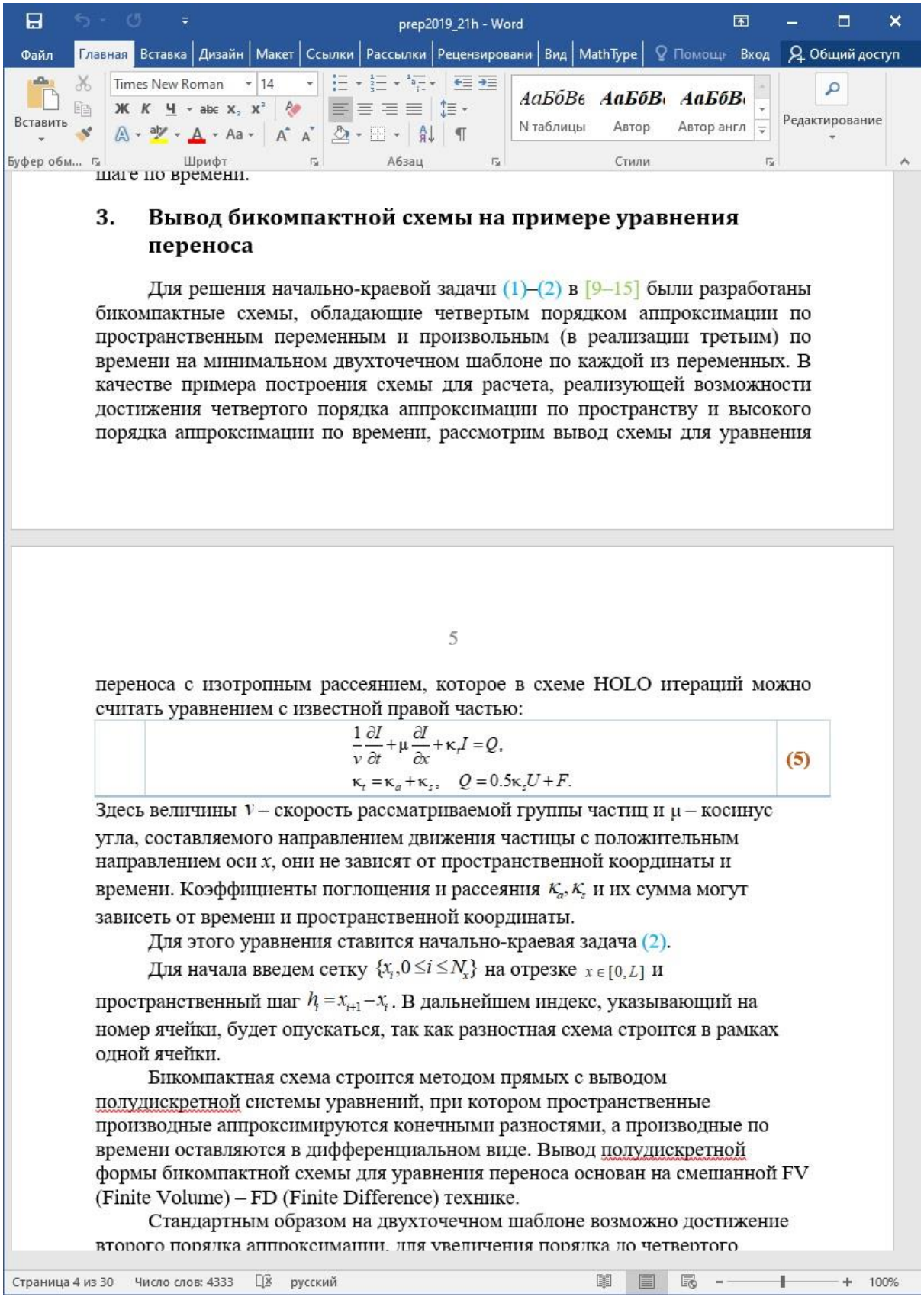

Рис. 10. Фрагмент препринта в редакторе MS Word

```
<section class="jats-sec" id="sec3">
    <h2 class="jats-title">3. Вывод бикомпактной схемы на примере уравнения переноса</h2>
    <p>
        Для решения начально-краевой задачи
        <a class="jats-xref" data-jats-ref-type="disp-formula" href="#fml1">(1)</a>
        <strong>–</strong>
        <a class="jats-xref" data-jats-ref-type="disp-formula" href="#fml2">(2)</a> в
        <a class="jats-xref" data-jats-ref-type="bibr" href="#bibr9">[9–15]</a>
        были разработаны бикомпактные схемы, обладающие четвертым порядком аппроксимации по
    </p>
    <figure class="jats-disp-formula-group" id="fml5">
        <div>
            <div>
                <p><span><math>...</math></span></p>
            </div>
            <div>
                <p><span class="jats-label">(5)</span></p>
            </div>
        </div>
    </figure>
    <p>
        Здесь величины <span><math>...</math></span> – скорость рассматриваемой группы части
        <span><math>...</math></span> – косинус угла, составляемого направлением движения ча
        <span><math>...</math></span> и их сумма могут зависеть от времени и пространственно
    </p>
    <p>Для этого уравнения ставится начально-краевая задача
        <a class="jats-xref" data-jats-ref-type="disp-formula" href="#fml2">(2)</a>.
    </p>
    <p>
        Для начала введем сетку <span><math>...</math></span> на отрезке
        <span><math>...</math></span> и пространственный шаг <span><math>...</math></span>.
        В дальнейшем индекс, указывающий на номер ячейки, будет опускаться, так как разностн
    </p>
</section>
```

Рис. 11. Фрагмент результата автоматической конвертации в HTML. Код

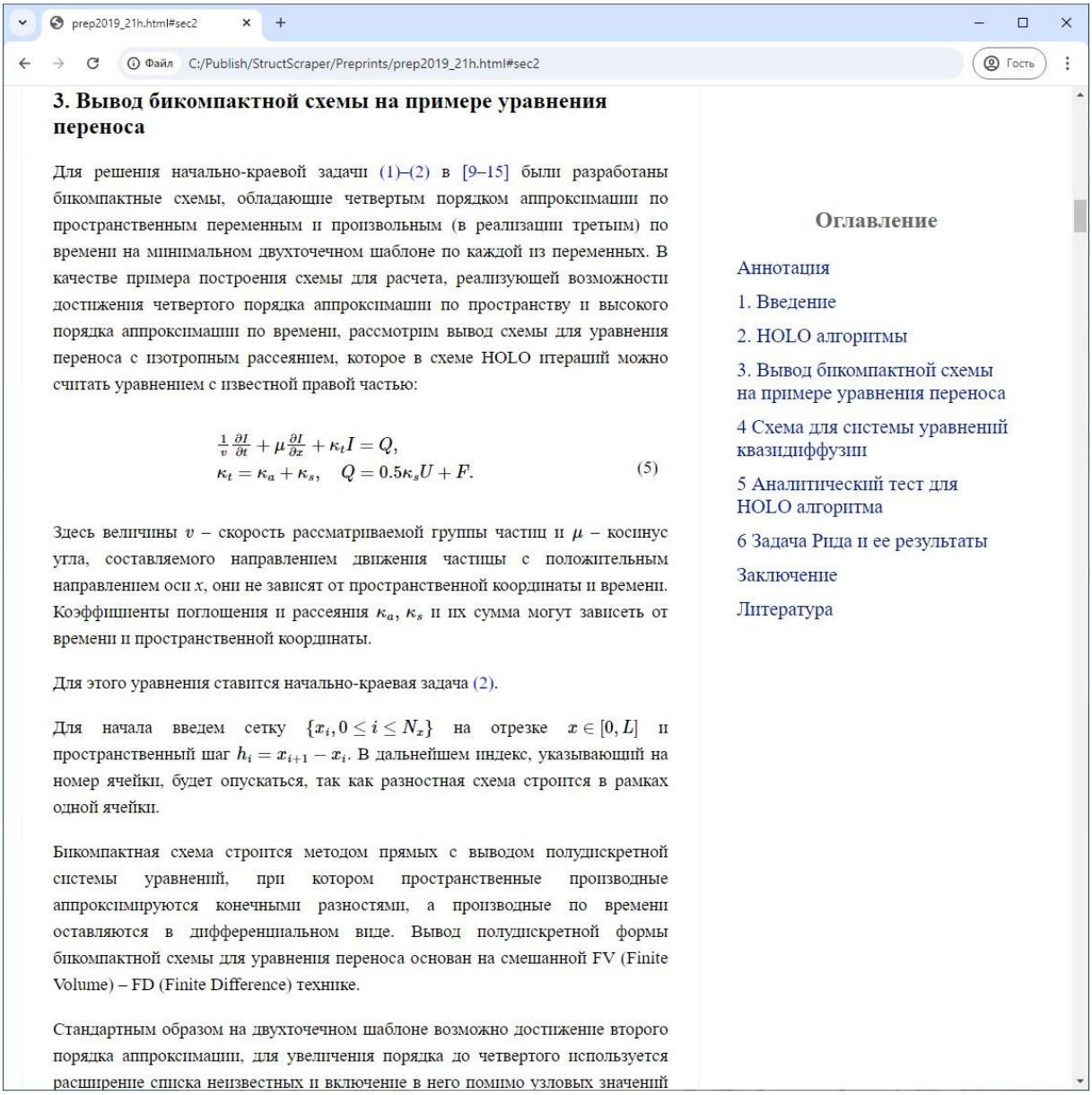

Рис. 12. Фрагмент результата автоматической конвертации в браузере

ЗАКЛЮЧЕНИЕ

Формат PDF, традиционно используемый для представления полных текстов научных статей, изначально предназначался для печатных изданий. Для онлайн-публикаций, доля которых существенно выросла по сравнению с печатными, более подходящим является формат HTML, обладающий рядом преимуществ за счет имеющихся в нем средств для лучшей структуризации материала, вставки мультимедийного контента и реализации разного рода интерактивных и

динамических возможностей. Для реализации преимуществ формата HTML контент в нем должен быть структурирован в соответствии со структурой научной статьи, а формулы представлены в машиночитаемых форматах MathML или TeX.

Наиболее распространенным подходом к созданию HTML-версии научной статьи является предварительное создание ее XML-версии в соответствии со стандартом JATS XML, в котором предусмотрены элементы и атрибуты для отражения структуры научной статьи, а затем ее автоматическое преобразование в формат HTML. Формат .docx текстового процессора MS Word, часто используемый авторами, не является специализированным форматом для научных документов и, хотя и представляет собой zip-архив xml- и медиа-файлов, в его XML-модели нет элементов, отражающих структуру научной статьи. Чтобы можно было автоматически преобразовать статью в формате .docx в формат JATS XML, ее структурные элементы в документе Word обычно выделяют с помощью пользовательских стилей. Ошибки стилевой разметки приводят к ошибкам в выходном XML-файле.

Мы предлагаем подход, при котором документ Word, размеченный специальными стилями, сначала преобразуется в формат HTML, где структурные элементы научной статьи выделяются при помощи классов и атрибутов, а затем уже документ формата HTML преобразуется в формат JATS XML. При таком подходе проще выявлять ошибки стилевой разметки, т. к. соответствующие ошибки в структуре выходного документа проще выявить в формате HTML, чем в формате XML, за счет визуализации структурных элементов при помощи каскадной таблицы стилей CSS.

Разрабатываемый нами программный инструмент, реализующий этот подход, включает шаблон MS Word для разметки стилями, конвертер математических формул из формата MathType в формат MathML, конвертер из формата .docx в формат HTML и конвертер из формата HTML в формат JATS XML. Инструмент ориентирован на малые издательства с небольшим числом сотрудников. Для сокращения времени подготовки исходного документа к конвертации в шаблон включено небольшое число стилей, достаточное, чтобы охватить структурные элементы, встречающиеся в большинстве исходных документов.

На данный момент реализованы конвертер математических формул и прототип конвертера из формата .docx в формат HTML. Конвертер из формата HTML

в формат JATS XML реализован частично. Планируем в течение года создать рабочие версии конвертеров и начать их внедрение в ИПМ им. М.В. Келдыша.

В дальнейшем планируется расширить возможности инструмента, добавив автоматическое выделение структурных элементов библиографических ссылок, имеющих DOI, задействовав для этого программный интерфейс Crossref REST API.

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. *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/404/489.

4. *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.

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

6. *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.

7. Inera JATS Solutions. URL: https://www.inera.com/jats-solutions/.

8. Ictect Intelligent Content for Journals. URL: https://www.ictect.com/JATS-XML.

9. *Visel D., Hyde A., Whitmore B. Kotahi: a new JATS production system* // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, May 3–4, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK579686/.

10. XSweet. The open .docx to HTML conversion tool. URL: https://xsweet.org/.

11. Mammoth. .docx to HTML converter.
URL: https://mike.zwobble.org/projects/mammoth/.

AN APPROACH TO CREATING AN HTML VERSION OF A SCIENTIFIC ARTICLE FROM A MANUSCRIPT IN MS WORD FORMAT FOR A LOW-BUDGET PUBLISHER

R. Y. Skornyakova[0000-0001-7372-3574]

Keldysh Institute of Applied Mathematics (Russian Academy of Sciences)

rimmaskorn@gmail.com

Abstract

The most common approach to creating an HTML version of a journal article among scientific publishers is to first create an XML version of the article in accordance with the NISO Journal Article Tag Suite (JATS) standard, followed by automatic conversion to HTML and PDF formats. However, obtaining an XML version from a manuscript in the .docx format of the MS Word word processor, often used by authors, when it contains a large number of complex formulas and tables is a difficult task. The existing software either does not cope with it in full or is expensive and inaccessible to small publishers with a limited budget. This paper proposes an approach to creating an HTML version of a journal article from a manuscript in .docx format containing formulas in MathType format, which does not require significant financial and time costs from the publisher. It also describes a currently implemented prototype of an underlied this approach converter of scientific articles from .docx format to HTML and JATS XML formats, which is applicable for KIAM preprints.

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

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.
URL: https://library.keldysh.ru/preprint.asp?id=2019-20.

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

4. *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.

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

6. *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.

7. Inera JATS Solutions. URL: https://www.inera.com/jats-solutions/.

8. Ictect Intelligent Content for Journals.
URL: https://www.ictect.com/JATS-XML.

9. *Visel D., Hyde A., Whitmore B.* Kotahi: a new JATS production system // Proceedings of Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA, May 3–4, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK579686/.

10. XSweet. The open .docx to HTML conversion tool. URL: https://xsweet.org/.

11. Mammoth. .docx to HTML converter. URL: https://mike.zwobble.org/projects/mammoth/.

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 6 ноября 2024 года