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).

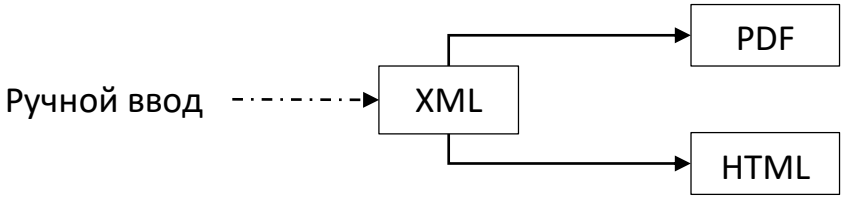

Рис. 1. Схема процесса XML-First

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

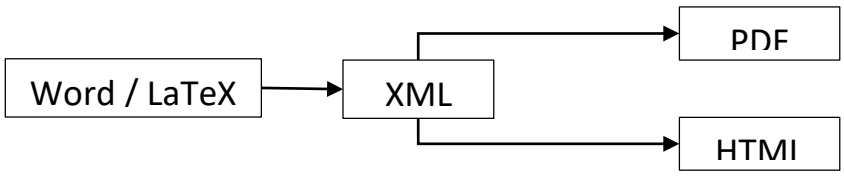

Рис. 2. Схема процесса XML-Middle (для PDF и HTML)

[16] https://www.sspnet.org/

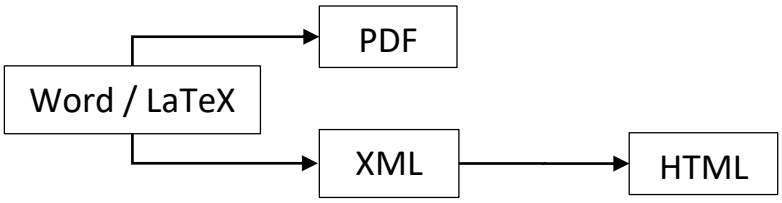

Рис. 3. Схема процесса XML-Middle (для HTML)

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.

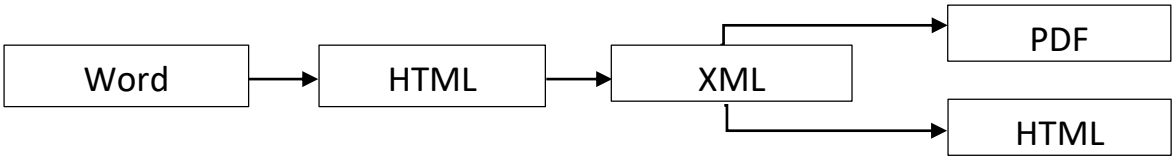

Рис. 4. Схема процесса XML-Middle с использованием промежуточного HTML

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.

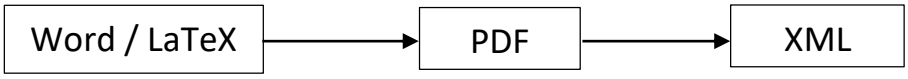

Рис. 5. Схема процесса XML-Last (без HTML)

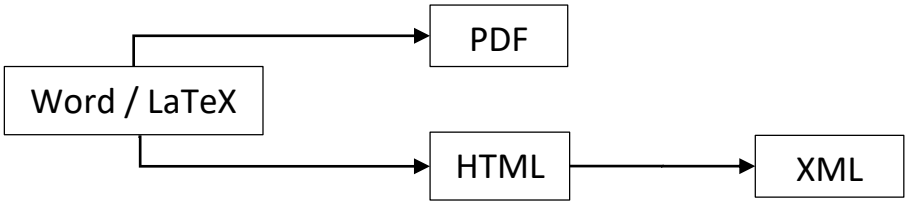

Рис. 6. Схема процесса XML-Last

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.

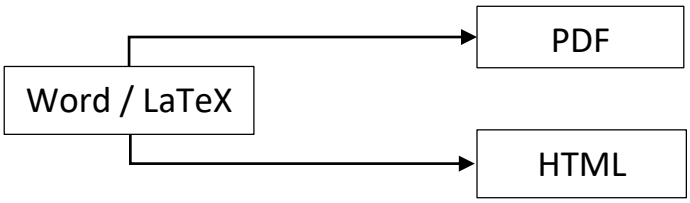

Рис. 7. Схема процесса без использования XML

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

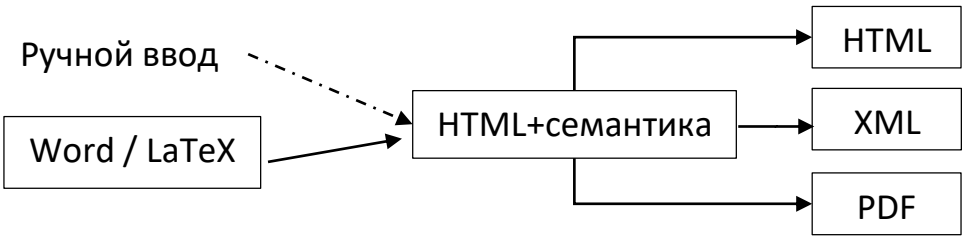

Рис. 8. Подход HTML-First

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

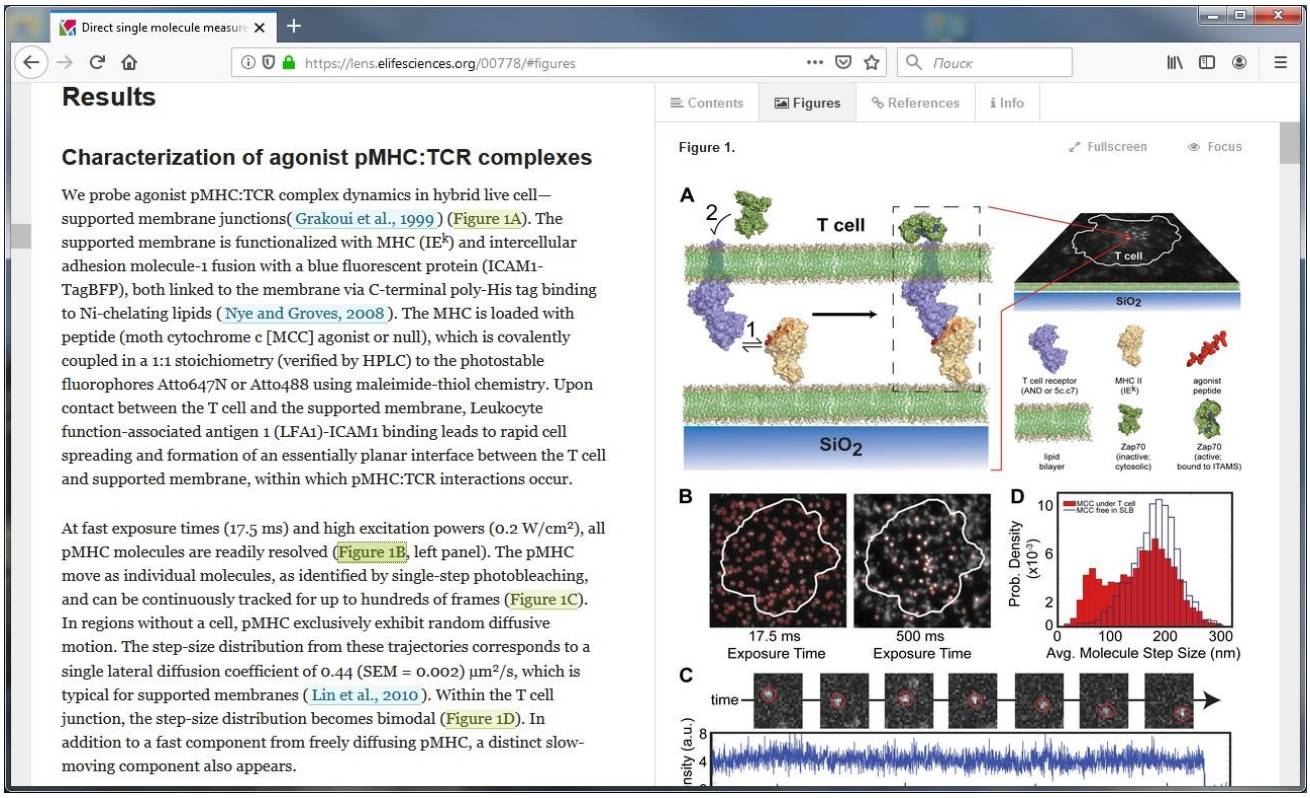

Рис. 9. Вид статьи во вьюере Lens

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.

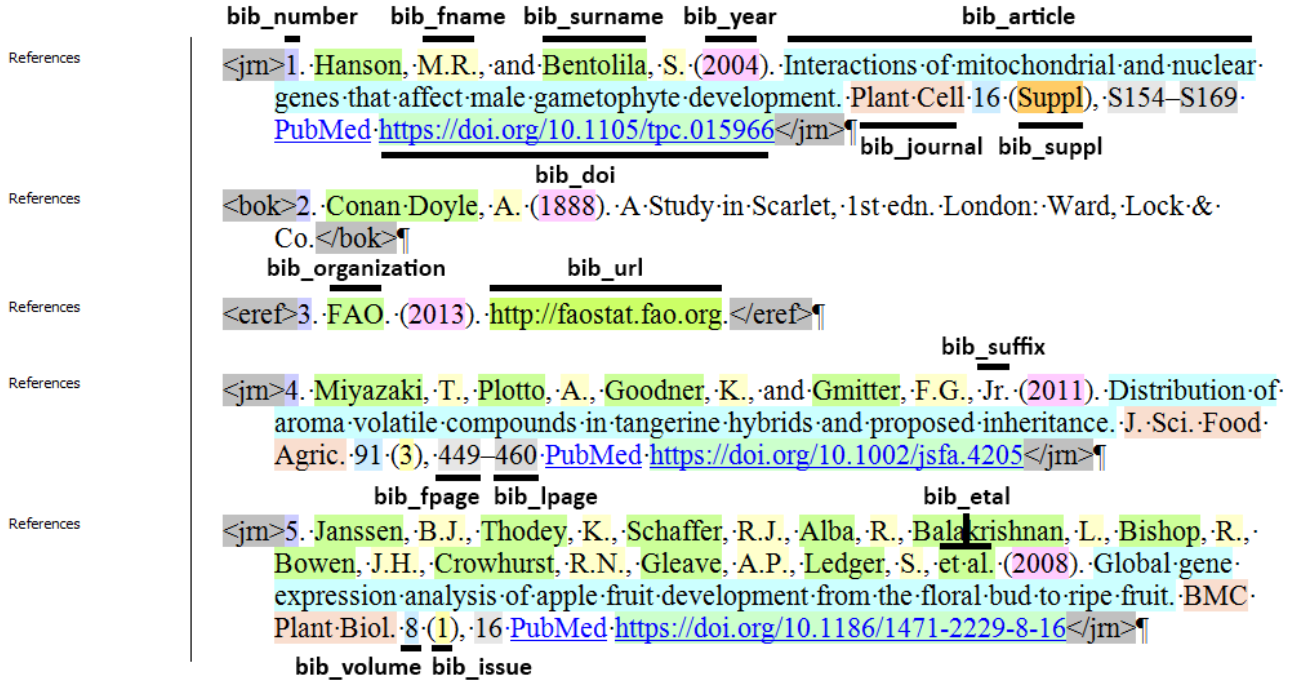

Рис. 12. Использование символьных стилей eXtyles JATS для элементов библиографической ссылки

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).

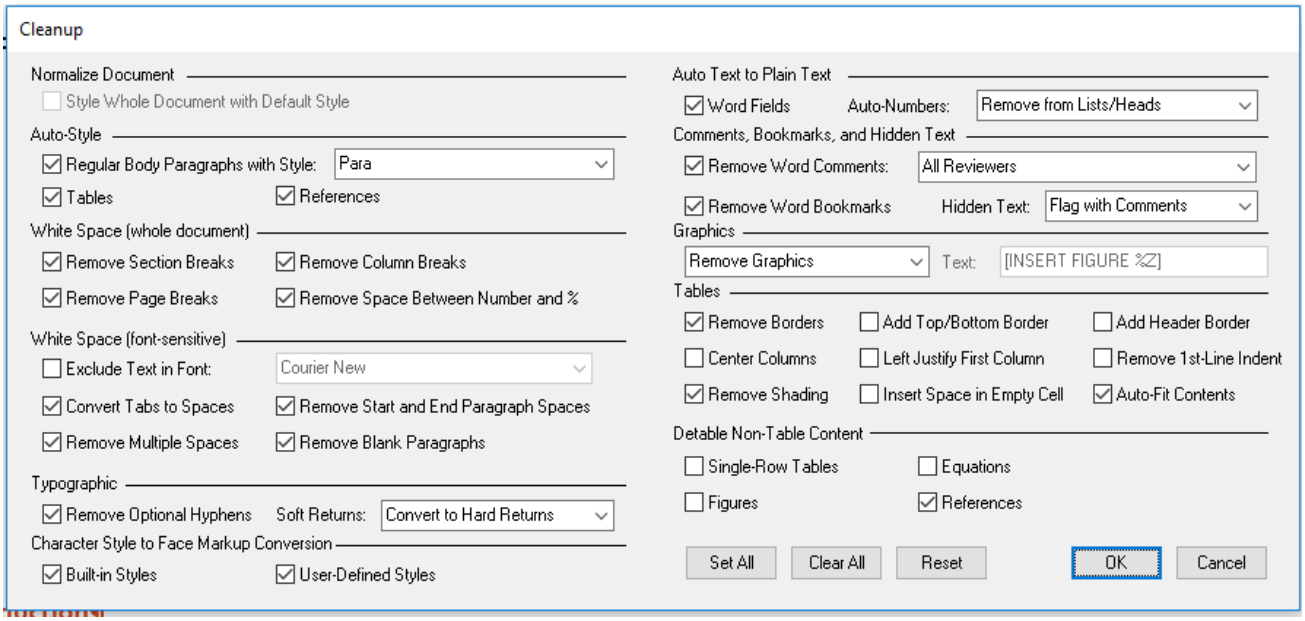

Рис. 13. Опции предварительного форматирования документа

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

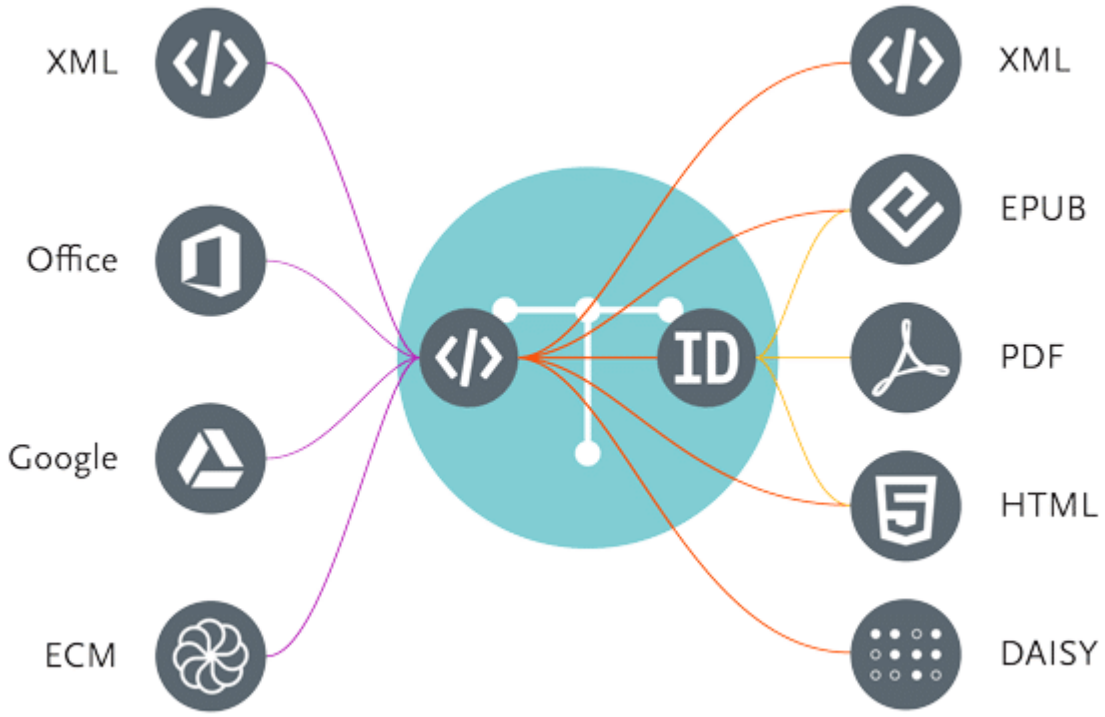

Рис. 14. Преобразования форматов, осуществляемые платформой Typefi
Источник рисунка https://www.typefi.com/wp-content/uploads/lotus-diagram.png

[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

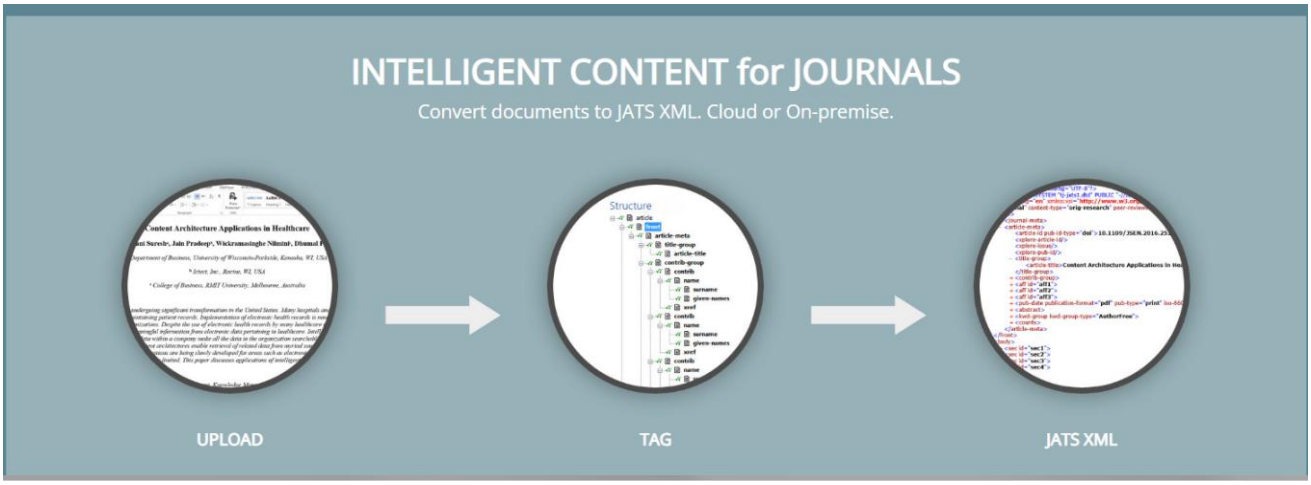

Рис. 15. Преобразование документа Word при помощи Ictect

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.

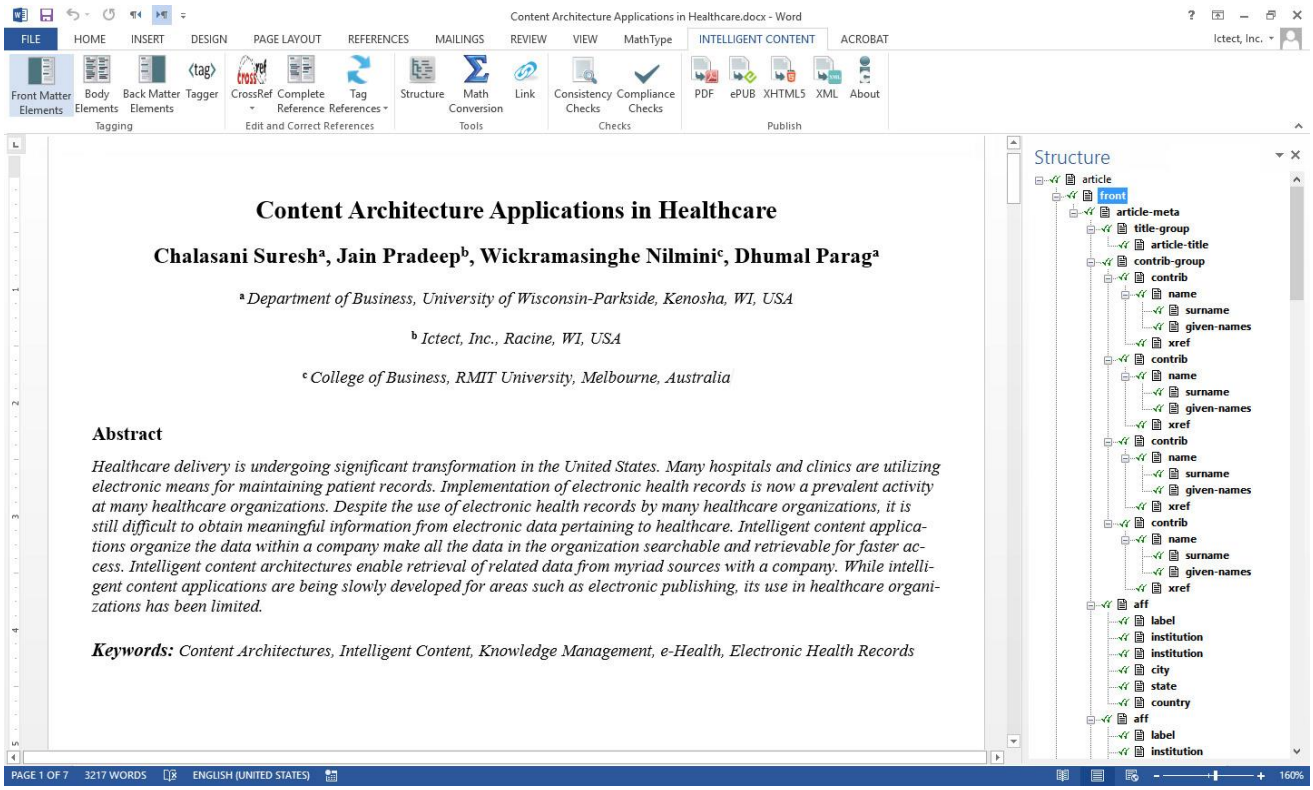

Рис. 16. JATS структура документа Word в плагине icTools

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.

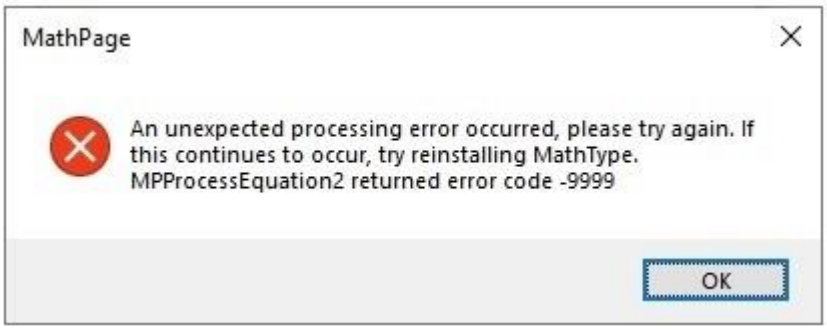

Рис. 18. Ошибка при попытке создания MathPage

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