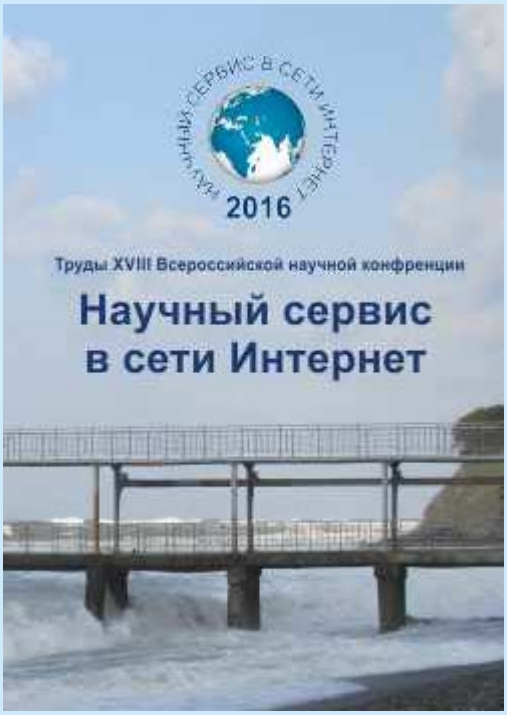
ИПМ им.М.В.Келдыша РАН

Абрау-2016 • Труды конференции



И.А. Мбого, Д.Е. Прокудин, А.В. Чугунов

Разработка инструментов интеграции научной информации в пространстве разнородных информационных систем

Рекомендуемая форма библиографической ссылки

Мбого И.А., Прокудин Д.Е., Чугунов А.В. Разработка инструментов интеграции научной информации в пространстве разнородных информационных систем // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2016. — С. 249-258. — doi:10.20948/abrau-2016-44

Размещена также презентация к докладу

Разработка инструментов интеграции научной информации в пространстве разнородных информационных систем

И.А. Мбого[1,2], Д.Е. Прокудин[1,2], А.В. Чугунов[2]

1 Санкт-Петербургский государственный университет
2 Университет ИТМО

Аннотация. Развитие научных исследований в условиях современного информационного общества характеризуется неуклонным ростом различных информационных систем, в которых отражаются результаты этих исследований. Для научных сообществ, как и для самих исследователей важнейшую роль играет оперативность в распространении результатов научных исследований через информационные системы. Некоторые из этих систем поддерживают автоматизированный обмен метаданными научных публикаций. Но существуют и такие, в которые метаданные попадают в результате ручного ввода, что сказывается как на оперативности, так и на качестве распространяемой информации. В данном исследовании предлагается один из возможных подходов к разработке инструментария, позволяющего максимально снизить дублирование информации и, как следствие, уменьшить возникновение ошибок ввода; а также автоматизировать процесс формирования метаданных публикаций для представления в разнородные информационные системы.

Ключевые слова: распространение научных знаний, автоматизированный обмен метаданными, разнородные информационные системы, шаблон, парсер

1. Введение

Развитие технологий информационного общества привело к тотальной информатизации, в том числе информатизации научной деятельности. Одним из важнейших аспектов научных исследований является оперативное распространение их результатов, которое в информационном обществе достигается в рамках реализации различных концепций и инициатив [1-7, 10, 11]. В основном для этого предназначены институциональные репозитории открытого доступа, которые используются, как правило, для самоархивирования результатов научных исследований сотрудников соответствующих научных организаций. Кроме этого, для распространения результатов научных исследований большую роль играют агрегаторы, которые аккумулируют не сами публикации, а метаданные о них, что позволяет в таких сетевых каталогах производить более эффективный поиск в отличии от поиска

по многим разрозненным репозиториям. При этом для сбора информации агрегаторы и репозитории используют средства автоматизации, например, протокол обмена метаданными OAI-PMH и формат представления метаданных Dublin Core. Однако, существуют информационные системы, не поддерживающие автоматизированный обмен метаданными, что создаёт проблемы для оперативного распространения результатов научных исследований. К тому же при ручном вводе метаданных в такие информационные системы возникают ошибки и опечатки, что ведёт к искажению информации, в том числе и при размещении её в наукометрических информационных системах.

2. Пути решения проблемы распространения результатов научных исследований научными сообществами

Все означенные проблемы характерны и для научных сообществ, деятельность которых объединяется проводимыми научными конференциями. Основным формальным результатом работы научной конференции является издание сборника её трудов, отражающего результаты представленных научных исследований. Распространение результатов научных исследований, отражённых в сборниках трудов конференций, происходит через размещение электронных версий сборников или статей (тезисов докладов) с метаданными на сайтах научных конференций, а также в открытых репозиториях организаций.

При этом, как правило, сборники размещаются целиком (в виде одного файла) без приведения метаданных к каждой статье, что затрудняет как поиск по статьям, так и дальнейшее распространение информации во внешние информационные системы. Размещение отдельных статей в открытых репозиториях организаций ограничивается аффилиацией авторов с этой организацией, поэтому в данном случае будут представлены статьи отдельных авторов, что ведёт к фрагментации представления информации.

С подобными проблемами столкнулось сообщество исследователей технологий информационного общества, объединённых проводимой ежегодно научной конференцией «Интернет и современное общество». Для их решения с 2014 года стало формироваться комплексное информационное пространство [9]. Его развитие связано с решением следующих основных задач:

- информационная поддержка организации и проведения ежегодной конференции «Интернет и современное общество»;
- автоматизация процессов подачи заявок для участия в конференции;
- автоматизация процессов подготовки статей и метаданных для последующей интеграции в информационные системы комплексного информационного пространства;
- размещение материалов конференции в собственном тематическом репозитории;

- оперативное распространение материалов конференции, отражающих результаты актуальных исследований.

Для этого в качестве ядра системы был построен открытый репозиторий материалов конференции (http://ojs.ifmo.ru/index.php/IMS), для чего была выбрана программная платформа Open Journal Systems (OJS) по следующим основным её возможностям:

- свободно распространяемое программное обеспечение;
- возможность представления каждого отдельного сборника в виде контейнера, включающего набор отдельных статей с метаданными и полными текстами или ссылками на внешние файлы;
- реализации протокола обмена мета данными OAI-PMH на уровне провайдера для автоматического обмена с внешними информационными системами;
- реализация формата Dublin Core как стандарта представления метаданных;
- наличие модулей экспорта метаданных в форматах основных агрегаторов, не поддерживающих автоматизированный обмен [12].

Выбор этого решения позволил разместить в ручном режиме материалы сборников конференций с 2011 по 2014 годы и по протоколу OAI-PMH интегрировать метаданные статей и тезисов из них такие агрегаторы как OAIster (http://www.worldcat.org/search?q=on:DGCNT+http://ojs.ifmo.ru/index.php/index/oai+IMS+RUITM&qt=results_page) и Соционет (https://socionet.ru/collection.xml?h=repec:rus:ims000). Также материалы, размещённые на платформе OJS, стали индексироваться в поисковой системе Академия Google (https://scholar.google.ru).

Организаторы конференции «Интернет и современное общество» не рассматривали такие информационные системы как Web of Science, Scopus в силу того, что они являются коммерчески ангажированными и не поддерживают инициативы свободного доступа к научной информации. Другие информационные системы (например, ORCID, Academia.edu, ResearchGate) позволяют размещать информацию о публикациях самим авторам в своих аккаунтах, что тоже исключило их из круга рассмотрения организаторов конференции. Наиболее перспективным для организаторов является присваивание статьям, размещённым в репозитории Университета ИТМО, универсального идентификатора цифрового объекта (DOI). Однако, в этом вопросе требуется единое корпоративное решение для возможности присваивания DOI публикациям издателя материалов конференции, которым является Университет ИТМО.

Для информационной поддержки организации и проведения конференции был создан информационный сайт научной конференции «Интернет и современное общество» (http://ims.ifmo.ru), программно-аппаратная платформа которого разработана в Университете ИТМО. В системе реализован механизм

подачи заявок на участие в конференции и загрузка рукописей статей и тезисов. Интеграция этой информационной системы с информационной системой управления (ИСУ) Университета ИТМО позволяет сотрудникам университета свои публикации отображать в корпоративных профилях, что учитывается в показателях их научной деятельности.

Изменившиеся с 2015 года требования Университета ИТМО привели к необходимости размещения публикаций сотрудников в открытый репозиторий ИМТО (http://www.openbooks.ifmo.ru). В связи с этим полные тексты статей сборников материалов конференции размещаются в этой информационной системе. Репозиторий Университета ИТМО не поддерживает автоматизированный обмен метаданными с другими информационными системами, поэтому в 2015 году пришлось вручную заполнить таблицу с метаданными (формат XLS), предложенную разработчиками программной платформы открытого репозитория Университета ИТМО.

Для размещения метаданных статей в открытом архиве материалов конференции «Интернет и современное общество» совместно с разработчиками открытого репозитория Университета ИТМО были сформированы XML-файлы соответствующего формата как результат экспорта из базы данных открытого репозитория Университета ИТМО. Из-за некорректного распознавания тэгов метаданных при импорте в OJS с использованием стандартного модуля импорта в формате XML пришлось проводить дополнительную ручную работу.

Размещение метаданных научных публикаций в единой наукометрической базе данных Российский индекс научного цитирования (РИНЦ), функционирующей на платформе Научной электронной библиотеки (НЭБ), также было проведено в 2015 году в ручном режиме с использованием онлайн инструмента XML-разметки «Артикулус». Возможность автоматизации этого процесса ограничивалась использованием механизмов «drag-and-drop» и «copy-past».

3. Проблема подготовки метаданных научных публикаций в материалах научных конференций

Ещё одной немаловажной проблемой, с которой столкнулись организаторы ежегодной научной конференции «Интернет и современное общество», явилась необходимость многочисленного ручного ввода метаданных подаваемых на конференцию материалов, при вводе метаданных в открытый репозиторий университета ИТМО и в Научную электронную библиотеку. При этом неизбежно возникали ошибки и опечатки. К тому же происходило дублирование одной и той же информации, а сам процесс удлинялся.

С подобными проблемами сталкиваются большинство организаторов научных конференций, когда сначала потенциальными участниками подаются тексты статей или тезисы докладов, которые сопровождаются минимально необходимым набором метаданных на русском и английском языках: имена

авторов; названия организаций, с которыми они аффилированы; адреса электронной почты; названия статьи или тезисов; аннотация; набор ключевых слов. Этих данных, как правило, достаточно для подготовки и публикации сборников материалов конференции. Но для размещения метаданных и текстов материалов сборников во внешних информационных системах необходим ввод дополнительной информации. Так, например, для размещения сборников в Научной электронной библиотеке (НЭБ) необходима дополнительные метаданные: название раздела сборника, коды классификации научной информации по принятым в России стандартам, SPIN-коды авторов в системе РИНЦ (для автоматической привязки статей к их профилям) и т.п. В качестве примера можно привести научную конференцию «Научный сервис в сети Интернет». По итогам работы конференции в 2015 году перед изданием сборника трудов авторам была разослана форма для заполнения метаданными. Некоторые из них дублировали информацию, полученную организаторами в шаблонах тезисов докладов и статей. Также были добавлены метаданные, необходимые для идентификации авторов в РИНЦ (SPIN-код) и классификации текстов (коды УДК). Это говорит о том, для требуется дополнительный сбор информации, которая, зачатую, дублируется. Организаторам конференции «Интернет и современное общество» также пришлось в 2015 году после публикации сборников и окончания конференции разослать авторам формы для внесения недостающей информации, что потребовало дополнительной ручной работы, удлинило процесс сбора метаданных (зависимость от скорости реакции авторов) и не позволило оперативно разместить статьи с метаданными в НЭБ и открытом репозитории Университета ИТМО.

4. Разработка инструментов подготовки метаданных для размещения в разнородных информационных системах

Для решения рассмотренных проблем был изменён алгоритм всего процесса от подачи рукописей статей на конференцию «Интернет и современное общество» до публикации сборников материалов и размещении полных текстов и метаданных в различных информационных системах [8]. Основной задачей была поставлена разработка инструментария, который позволил бы:

- получить единовременно все метаданные от участников конференции;
- использовать первоначально введённые метаданные как для публикации сборников материалов конференции, так и для их размещения в комплексном информационном пространстве и во внешних информационных системах, не поддерживающих автоматизированный обмен метаданными.

Для решения первой задачи был разработан шаблон рукописи статьи в формате текстового файла с поддержкой макросов для текстового редактора MS Word (расширение .docm). Он доступен для загрузки на официальном сайте конференции «Интернет и современное общество» (http://ims.ifmo.ru/file/pages/10/ims-2016-template.docm). Там же размещена и

подробная инструкция по работе с этим шаблоном. В шаблон после основного текста статьи была размещена сводная таблица с полным набором метаданных, рассчитанная на внесение информации как о самой рукописи, так и её авторов (максимально – до трёх). Шаблон содержит набор стилей, необходимых для единообразного оформления статей и последующей предпечатной подготовки. Шаблон статьи включает в себя базовый набор метаданных, представленных на русском и английском языках: название статьи; фамилии авторов; инициалы имён и отчеств авторов; названия организаций, с которыми они аффилированы; адреса электронной почты; аннотации; ключевые слова и список использованной литературы.

Метаданные в таблице вносятся тремя основными способами (рис. 1):

1) с использованием специального макроса копируются автоматически из шаблона статьи;

2) вносятся автором вручную;

3) вручную заполняются организаторами конференции после издания сборника материалов конференции и размещении полных текстов в открытом репозитории Университета ИТМО.

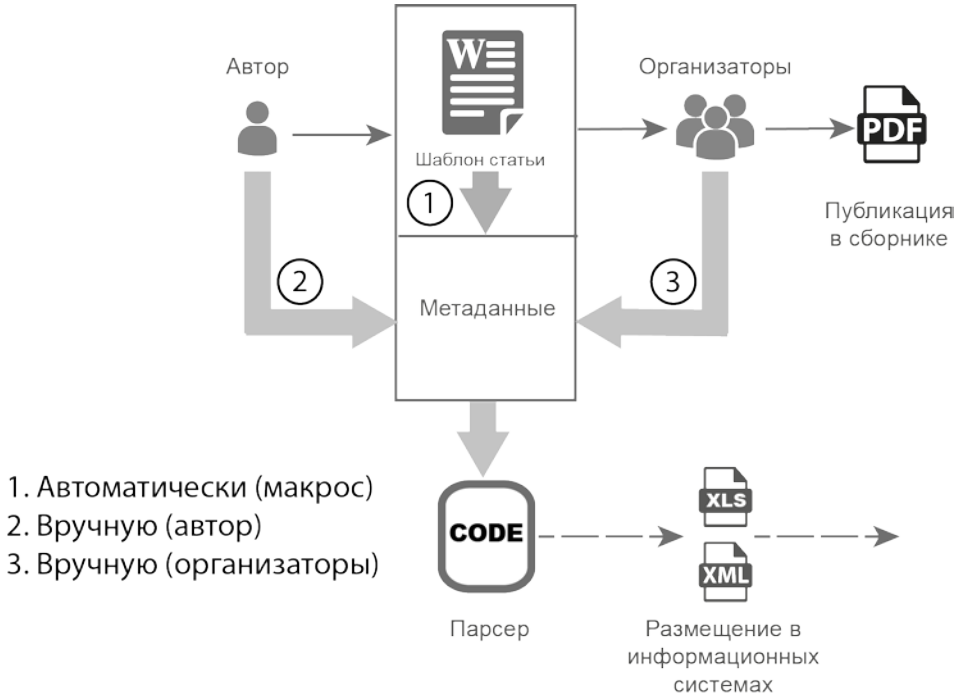


Рис. 1. Процессы подготовки метаданных для размещения в разнородных информационных системах

Шаблон реализует следующие инструменты тестового редактора MS Word:

- метаданные вводятся в текстовые поля, что позволило создать макрос для автоматического заполнения этих полей из текста через обновление при изменении соответствующей информации в полях ввода;

- поля в тексте размечены именованными закладками. Из таблицы с метаданными происходит обращение к полю ввода в виде ссылки на закладку {REF имя_закладки} и происходит копирование информации из поля на соответствующей закладке;

- макрос автоматически обновляет данные полей из текста в таблицу.

Остальные поля заполняются вручную, что указано в названии каждого поля. В зоне ответственности авторов находятся поля, к которым, например, относятся коды различных классификаторов информации (принятые в России и используемые в Научной электронной библиотеке); тематический раздел конференции (выбирается из выпадающего списка); дополнительная информация об авторах. Остальные поля заполняются организаторами конференции уже после формирования всех выходных данных сборников материалов конференции (например, номера страниц). Окончательное формирование полного набора метаданных происходит после размещения полных текстов статей в открытом репозитории Университета ИТМО и получения интернет-ссылки на каждую статью.

Для возможности дальнейшей пакетной обработки метаданных таблица вынесена в новый раздел. При этом каждое поле имеет свой стиль, а сами стили имеют вложенную структуру, например, author1-surname, name, initials ... author2 - surname, name, initials ... и т.д. В этом примере вложенность достигается применением стиля author1 к ячейкам таблицы, содержащим информацию об авторе. Сам же файл будет сохраняться в качестве фильтрованного HTML-формата. Ранее предполагалось использовать формат XML, но в процессе разработки шаблона при экспорте в этот формат в файле записывалось много лишних тэгов со сложной структурой, разбирать которую значительно сложнее. Поэтому был выбран формат HTML, структура которого значительно проще.

После процесса рецензирования и принятия положительного решения о допуске к участию в конференции из шаблона выделяется текст рукописи статьи, проходит его редактура и корректура, а затем он передаётся на макетирование в сборник материалов конференции и окончательно оформляется в файле формата PDF.

Для дальнейшей обработки метаданных, размещённых в шаблоне статьи, в настоящее время разрабатывается специальное программное обеспечение (парсер). Разрабатываемый парсер предполагает пакетную обработку метаданных статей и записи их в электронных форматах для дальнейшего импорта во внешних разнородных информационных системах, не поддерживающих автоматизированные процессы обмена метаданными:

- формирование таблицы в формате XLS для размещения метаданных вместе с полными текстами статей в формате PDF в открытом репозитории Университета ИТМО (http://openbooks.ifmo.ru). Структура таблицы предоставлена администраторами репозитория;

- формирование файла в формате XML для импорта метаданных сборников и статей в электронный архив материалов конференции «Интернет и современное общество» ((http://ojs.ifmo.ru/index.php/IMS/issue/archive). Структура файла и назначение тэгов взяты из результатов экспорта одного из сборников прошлых лет, размещённых в архиве материалов конференции. Для этого использовался стандартный модуль экспорта в формате XML;

- формирование файла в формате XML для импорта в Научную электронную библиотеку (http://elibrary.ru). Структура файла и назначение тэгов взяты из результатов выгрузки проекта одного из сборников прошлых лет, созданного в инструменте XML-разметки «Артикулус».

Так как в сборниках материалов конференций есть информация, относящаяся ко всему сборнику (например, выходные данные – ISBN, название издательства, название самого сборника и т.д.), то нет необходимости вносить её в метаданные каждой статьи. Поэтому в приложении будет реализована возможность вручную внести эту информацию уже после публикации самих сборников.

5. Заключение

Предложенный подход к автоматизации процессов подготовки метаданных публикаций результатов научных исследований может быть использован как один из возможных, который определяется логикой построения конкретного комплексного информационного пространства поддержки междисциплинарного научного направления и элементами, его составляющими. Специфика этого подхода позволила оптимизировать процессы сбора, обработки и размещения метаданных в информационные системы, которые не поддерживают их автоматизированный обмен. Кроме этого, одним из принципиальных моментов является убеждённость авторов предложенного подхода в том, что в электронном виде текст научной статьи должен быть размещён только в одной информационной системе, а в других системах размещаются только метаданные этой публикации и ссылка на полный текст. Такой принцип должен способствовать уменьшению информационного шума и, как следствие, вести к снижению общей информационной энтропии. Это особенно важно в информационную эпоху, когда нарастают темпы увеличения информации. В научном информационном пространстве это ведёт к тому, что, с одной стороны, происходит дублирование научной информации, с другой – результаты актуальных научных исследований тонут в море маловажной (с научной точки зрения) информации, генерируемой для достижения наукометрических показателей, спускаемых «сверху» чиновниками от науки.

Литература

1. Burgelman, J.C., Luber, S., Von Schomberg, R., Lusoli, W. Open Science: Public consultation on "Science 2.0: Science in transition". Key results, insights and possible follow up. 2015. URL: http://www.science20-conference.eu/wp-content/uploads/2015/04/01_Jean-Claude_Burgelman_-_Open_Science__outcome_of_the_public_consultation_on__Science-20_science_in_transition.pdf (дата обращения: 25.04.2016).
2. Mukherjee, A., Stern, S. Disclosure or secrecy? The dynamics of Open Science. International Journal of Industrial Organization. 2009. Volume 27, Issue 3, May 2009. 449-462, DOI=http://dx.doi.org/10.1016/j.ijindorg.2008.11.005.
3. Nielsen, M. Reinventing Discovery: The New Era of Networked Science. Princeton. 2011. N.J.: Princeton University Press.
4. Open Access 2020. http://oa2020.org (дата обращения: 25.04.2016).
5. Open Science [электронный ресурс]. http://openscience.com (дата обращения: 25.04.2016).
6. Parsons, J. Welcome to Science 2.0 | Open Access in Action. LIBRARY JOURNAL. March 15, 2016. URL: http://lj.libraryjournal.com/2016/03/oa/welcome-to-science-2-0-open-access-in-action/ (дата обращения: 25.04.2016).
7. Ten years on from the Budapest Open Access Initiative: setting the default to open. URL: http://www.budapestopenaccessinitiative.org/boai-10-recommendations (дата обращения: 25.04.2016).
8. Кудрявцева М.В., Мбого И.А., Прокудин Д.Е. Реализация подхода к автоматизации информационных процессов поддержки междисциплинарного научного направления в пространстве разнородных информационных систем // Информационное общество: образование, наука, культура и технологии будущего: сборник научных статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург, 22 – 24 июня 2016 г. — СПб: Университет ИТМО, 2016. — 200 с. С. 87-99.
9. Мбого И.А., Прокудин Д.Е., Чугунов В.А. Формирование информационного пространства междисциплинарного научного направления: подходы и решения // Межотраслевая информационная служба. 2015. №1. С. 36-44.
10. Паринов С.И. Открытая наука // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (г. Новороссийск, 21-26 сентября 2015 г.). М.: ИПМ им. М.В. Келдыша, 2015. С. 267-278. URL: http://keldysh.ru/abrau/2015/proc.pdf#page=267 (дата обращения: 26.04.2016).
11. Паринов С.И. Развитие электронных библиотек – путь к Открытой Науке [Электронный текст] // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции RCDL'2009. Петрозаводск: КарНЦ РАН, 2009. С. 225-234.

URL: http://rcdl.ru/doc/2009/225_234_Invited-2.pdf (дата обращения: 24.04.2016).
12. Прокудин Д.Е. Через открытую программную издательскую платформу к интеграции в мировое научное сообщество: решение проблемы оперативной публикации результатов научных исследований // Научная периодика: проблемы и решения. 2013. № 6 (18). С. 13-18. DOI= http://dx.doi.org/10.18334/np36109.