Дискуссионная площадка: «Инфраструктура и сервисы цифровой экосистемы библиотек»

doi:10.18720/SPBPU/2/k19-143

РАБОТА С ЦИФРОВЫМИ ДОКУМЕНТАМИ В ОТКРЫТОМ АРХИВЕ НАУЧНОГО ЦЕНТРА

DIGITAL PUBLICATIONS IN THE OPEN ARCHIVE OF THE SCIENTIFIC CENTER

Ковязина Елена Васильевна, научный сотрудник, к.т.н., Институт вычислительного моделирования Сибирского отделения Российской академии наук, elena@icm.krasn.ru

Kovyazina Elena, researcher, PhD, Institute of Computational Modeling of Siberian Branch of Russia Academy of Science elena@icm.krasn.ru

Аннотация. Наличие полнотекстового архива публикаций сотрудников научного центра позволяет вести их учет и продвижение, обеспечить открытость результатов научных исследований для мирового сообщества. Однако цифровые документы имеют ряд качественных особенностей по сравнению с традиционными печатными документами. В докладе представлена попытка определить проблемы обеспечения ссылочной и содержательной целостности цифровых документов в открытом архиве и возможные пути их решения.

Abstract. The presence of a full-text archive of employee publications at a research center allows them to be kept in record and advance, and to ensure the openness of research results to the global community. However, digital documents have a number of qualitative features in comparison with traditional printed documents. An attempt to determine the problems of ensuring the linking and content integrity of digital documents and possible solutions is presented in this report.

Ключевые слова. Цифровые репозитории, цифровые документы, содержательная и ссылочная целостность.

Keywords. Digital repositories, digital documents, linked data, content integrity.

Введение. Современное состояние научного взаимодействия характеризуется, в частности, формированием инфраструктуры открытой науки в глобальных сетях. Разработка и предоставление в пользование сервисов, предназначенных для оценки публикационной активности научных

организаций и отдельных ученых, а также качества публикуемых научных статей стали существенной частью деятельности индексов научного цитирования (Web of Science, Scopus, РИНЦ) и разработчиков программных систем проверки на заимствования (например, Антиплагиат). Существенный вклад в эту работу вносят и корпоративные инициативные проекты, как отечественные, например, [1], так и зарубежные, например, [2-3]. Интенсивное развитие инфраструктуры открытой науки влечет за собой необходимость продвижения результатов научных исследований академических институтов и университетов в публичную среду, и, как следствие, определяет востребованность работы библиотек соответствующих организаций в этом направлении. Эволюционное развитие библиографических баз данных трудов сотрудников научно-исследовательских организаций через полнотекстовые базы данных к открытым архивам требуют квалифицированной работы сотрудников библиотек и информационных служб с цифровыми документами, которые все чаще не являются простыми электронными копиями бумажных публикаций, а представляют собой полноценные цифровые объекты, обладающие совершенно новыми качественными особенностями.

С точки зрения научного взаимодействия для обеспечения цифровой трансформации науки любой цифровой объект, будь то документ или набор данных, должен быть FAIR (Findable, Accessible, Interoperable and Re-usable) [4], т.е. должен быть легко найден, доступен любому, у кого есть Интернет-связь, обеспечивать интероперабельность при манипуляциях с ним и давать возможность многократного использования. Это наиболее общие требования, связанные с внешним воздействием на объект и взаимодействием с ним. Однако в указанном контексте практически не отражено состояние самого цифрового объекта. Как это выглядит на практике? Одним из наиболее очевидных свойств цифрового документа в Web-пространстве является наличие внутри него активных гипертекстовых ссылок. Цифровой документ практически всегда документ гипертекстовый. Как указывают авторы [5]: «Реализация взаимных ссылок в цифровых документах не представляет большой сложности, однако при этом электронный документ приобретает новое качество. Во-первых, электронный объект с реализованными связями уже не совсем соответствует своему печатному оригиналу. Это уже другой объект. Этот факт должен быть учтен всеми юридическими нормами. Во-вторых, внедренные в объект связи должны быть гарантировано актуальными. Как следствие, появляется отличное от традиционных библиотек требование обеспечения ссылочной целостности данных. Это очень жесткое требование, которое тяжело обеспечить даже в хорошо формализованных системах управления базами данных». И далее уточняют, что каждый цифровой документ есть «новый цифровой объект как самосогласованное хранилище цифрового контента, или база данных

цифровых объектов». Несмотря на то, что приведенное определение было сделано относительно давно - статья датируется 2010 годом, с течением времени оно не только не утратило злободневность, а, напротив, стало еще более актуальным из-за роста разнообразия ссылок и связей между цифровыми объектами в веб-пространстве. Задача построения теории и моделей документов как цифровых объектов, включая методы работы с ними, глубока и обширна, и все еще ожидает своих исследователей. Здесь ограничимся небольшим перечнем практических проблем, возникающих в библиотеках научно-образовательных учреждений при учете и обслуживании цифровых публикаций их сотрудников.

Помимо чисто организационных проблем сбора и библиографического описания цифровых публикаций, юридических проблем, связанных с разделением имущественных и авторских прав, актуальной является также проблема сохранности цифрового архива, включающая выбор оптимального формата, в котором будет храниться файл, конверсию цифровых документов в выбранный формат и, в дальнейшем, отслеживание контрольных сумм цифровых объектов в хранилище. Несанкционированное изменение контрольной суммы служит индикатором порчи или разрушения файла, и влечет за собой необходимость его восстановления из резервной копии. Перечень проблем и затруднения в их решении существенно различаются в зависимости от того в какой программной и технологической среде формируется цифровой архив. Эффективнее всего вышеперечисленные проблемы решаются с помощью технологий открытых архивов (ОА), реализованных на специализированном программном обеспечении. Для определенности в дальнейшем в качестве такой системы будем иметь ввиду свободно распространяемую в сети Интернет программную платформу DSpace. Она обеспечивает поддержку хранилищ различного типа, включая трипл-хранилища, и включает периодическое отслеживание контрольных сумм цифровых объектов, на основе которого возможна организация служб оповещения пользователей и системы восстановления утраченной в результате технических сбоев и/или вирусных и хакерских атак информации.

Для понимания проблем ссылочной и содержательной целостности цифровых документов и решения возникающих при этом задач требуется знание особенностей структуры и жизненного цикла цифровых документов. Гипертекстовые ссылки формируют систему связей цифрового документа, как с внешними по отношению к нему цифровыми объектами, так и между фрагментами текста внутри документа. Соответственно, можно сказать, что существуют внешние ссылки, указывающие на другие цифровые объекты, связанные с исходным, и внутренние ссылки, указывающие на место или фрагмент внутри самого документа.

Ссылки на внешние объекты. Внешние ссылки предоставляют самые широкие возможности для научной деятельности. С их помощью реализуются перспективные технологии связывания цифровых объектов в семантическом веб-пространстве, работа с таксономиями, онтологиями и т.д. Связанные данные – это наиболее актуальный раздел работы с цифровыми документами, активно развивающийся и перспективный, что подтверждает большое количество публикаций на эту тему, например, [6], и выход методических указаний по работе с ними Американской библиотечной ассоциации (ALA) [7]. Внешние ссылки создают достаточно сложную систему связей между документами в сети, в то время как в традиционных библиотечных технологиях широко распространен только один тип связей - иерархический. Примерами внешних ссылок могут служить:

1. Пристатейные списки, создающие связи цитирования, широко используемые для наукометрических исследований;
2. Ссылки внутри статьи на внешние цифровые документы, например, поясняющие используемые в ней термины и определения, а также расширяющие область повествования;
3. Ссылки на внешние карты, схемы, таблицы и иные научные нетекстовые данные, не являющиеся частью документа, но используемые в нем.

Очевидной проблемой для внешних ссылок является утрата их актуальности вследствие удаления или перемещения из места хранения, на которое указывает ссылка. Эта проблема хорошо известна всем пользователям глобальных сетей. Проверка актуальности ссылок является трудоемким и весьма творческим процессом. Несмотря на то, что некоторые программные платформы, используемые для формирования цифровых документов, имеют встроенные или подключаемые (плагины) средства проверки актуальности ссылок, однако они, как правило, ограничиваются только проверкой на отсутствие веб-страницы (broken) по указанному в ссылке адресу. Отыскать, куда перемещен ресурс его владельцами, можно только вручную, даже при наличии по месту ссылки подсказки-указателя. Решением проблемы актуальности ссылок в научных публикациях является использование постоянных идентификаторов цифровых документов, таких как, например, DOI, или принятые в DSpace идентификаторы CRNI Handle System, используемые затем в URI цифрового документа и перемещаемые вместе с ним. Однако и это не избавляет полностью от проблемы, например, в случае полной утраты информационного ресурса или временной утраты подписки на него. Существует также и проблема «устаревания» цифрового документа. Например, контроль ссылочной целостности pdf-файла документации DSpace, «живого» и постоянно актуализируемого, за восемь последних лет показало утрату актуальности включенных в него внешних ссылок примерно на 1,5-2% в год, что в результате составило порядка 15-20%

утраченных ссылок. А научные публикации десятилетней давности в цифровом архиве Института вычислительного моделирования Сибирского отделения РАН утратили до 80% процентов внешних ссылок.

Внутренние ссылки цифрового документа. В эпоху цифровых копий печатных публикаций проблемы актуальности внутренних ссылок документа не существовало. Проблема внутренней ссылочной целостности документа, или его содержательной целостности, появилась с развитием технологий самодепонирования авторами своих публикаций в институциональных репозиториях. Техническая возможность постоянной корректировки документа автором открыла новые возможности, одновременно породив и новые проблемы. Поясним эти возможности на примере перспективных технологий работы с «живыми» документами, представленных в [1, 8]. Как пишут авторы статьи [8], «практика размещения самими учеными результатов своих исследований в форме научных статей и материалов в открытом доступе в сети Интернет постепенно получает организационную поддержку. … Научные статьи и материалы, депонируемые их авторами в электронном репозитории своей организации, становятся частью профессиональной информационной среды. Они цитируются наряду с "полноценными" публикациями в рецензируемых журналах. При этом онлайновые средства для электронного депонирования являются общедоступными и достаточно просты в использовании. Как следствие, авторы научных статей и материалов могут вносить в них изменения в общем случае в течение всей своей профессиональной жизни. При массовом использовании подобной практики электронные научные статьи и материалы получают статус "живого" документа (в зарубежной литературе “liquid publication”- «текучая (или неустойчивая) публикация»)». В этой же работе приведен целый ряд типов научных документов, которые в перспективе должны были бы стать «живыми», например:

- обзорная статья в некоторой области науки, которую автор может пожелать актуализировать с появлением новых значимых результатов в охватываемой данным обзором области;
- пополняемая научная библиография по какой-либо тематике исследований;
- описание пополняемой музейной коллекции в систематизирующих областях науки, например, ботанического гербария, коллекции насекомых, минералов и т.д.
- отчет, представляющий результаты многоэтапного научного проекта;
- опись архивных документов и т.п.

Работа с «живыми» цифровыми документами библиотечными технологиями практически не регламентирована. Требования ГОСТ 7.0.95-

2015 «Электронные документы» в отношении «живых» документов неоднозначны, нет четко формализованной границы между «редакцией» и «версией» цифрового документа, между «версией» и новым документом и т.п. Возникает много трудностей и неоднозначных решений при формировании описательных (библиографических) метаданных в системах автоматизации библиотек, построенных преимущественно на форматах семейства MARC. Анализ причин и деталей плохой применимости MARC-форматов для описания цифровых объектов приведен в [6, 9]. Технологии открытых архивов, использующие более приспособленные для цифровых документов форматы описания, например, QDC или MODS, позволяют работать с версиями и редакциями цифровых документов, используя встроенную систему контроля версий. Поиск в этом случае идет по последней версии документа, но пользователям доступна вся его история. Система самодепонирования построена на сетевых протоколах SWORD и позволяет авторам работать с документом в течение всего его жизненного цикла. Основные проблемы пользователей при работе с такими документами возникают в случае «ссылок цитирования», когда, например, фрагмент документа, на который ссылается цитирующая работа перестает существовать. Авторами статей [1, 8] предлагается целый ряд сервисных возможностей для корректного использования неустойчивых «живых» публикаций в сфере научной коммуникации. Например, возможно автоматическое оповещение авторов, цитировавших данную работу, об изменениях, внесенных в первоисточник, и приведенную цитату, в частности. При необходимости можно также автоматизировать внесение изменений одновременно во все публикации данного автора, если это касается, например, изменившихся фактографических данных. При более подробном размышлении нетрудно развить область применения указанных качеств, хотя очевидно, что сам факт существования «живых» документов создает массу проблем контроля на плагиат и научной этики.

Ссылки, формирующие цифровой документ. Цифровой документ может иметь сложную структуру, состоять из нескольких цифровых объектов (файлов), в том числе разнородных, и имеющих различные форматы хранения, которые требуют для работы с ними использования различных программных средств. Простой пример такого документа – упорядоченный набор tiff-страниц. Включение в цифровые документы аудио - и видеофрагментов, формирование гипермедийной среды Интернет, открывает новые возможности и качества документов, связанные с ней. Любопытное решение по поиску в аудиозаписях по ключевым звучащим словам было представлено, например, для Президентской библиотеки [10].

Существуют также дополнения к цифровому документу, прикрепленные к нему в качестве неотъемлемой части, например, существовавший до недавних пор в DSpace текст лицензии Creative Common.

Связи между элементами, составляющими цифровой документ в открытом архиве, регламентируются его структурными метаданными.

Ссылочная целостность метаданных. Определенная часть гипертекстовых ссылок цифрового документа может быть представлена как в самом документе, так и в его описательных метаданных. Это могут быть ссылки на полный текст документа на сайте правообладателей или ссылки цитирования из списка библиографии. Текст лицензии, условия эмбарго, авторитетные данные и классификационные схемы – все это может существовать в виде ссылок в метаданных цифрового документа. Технологии открытых архивов, как правило, содержат средства обеспечения ссылочной целостности метаданных цифровых объектов.

Заключение. Развитие технологий, программных средств и сервисов инфраструктуры научных коммуникаций в перспективе предоставляет широкие возможности для интеллектуальной и исследовательской деятельности, интегрируя в единое целое документы и данные, включая их корректировку, пополнение и обработку в течение всего времени их существования. Традиционные технологии работы с документами в библиотеках не обеспечивают полноценную интеграцию имеющихся информационных ресурсов в развивающуюся цифровую среду, а значит требуют изменений и трансформации в соответствии с качественными особенностями цифровых документов. Для таких изменений требуются методические разработки, рекомендации и инструкции для персонала библиотек научно-образовательных организаций, и, безусловно, новые компетенции сотрудников, работающих с цифровыми объектами. Освоение новых принципов и подходов, технологических решений, систем и сервисов, используемых мировым библиотечным сообществом, повлечет за собой новое осмысление работы с информацией, ее будущих перспектив, и роли библиотекарей в этом процессе.

СПИСОК ЛИТЕРАТУРЫ

1. Kogalovskii, M.R. and Parinov, S.I., A Virtual Scientific Communication Environment Based on a Semantic Scientific Information System // Automatic Documentation and Mathematical Linguistics – 2016. - vol.50. - no.5. - pp.189–194. – Режим доступа: https://link.springer.com/content/pdf/10.3103%2FS0005105516050046.pdf (дата обращения 3.06.2019).
2. Brinckman Adam, Chard Kyle, Gaffney Niall et. Computing environments for reproducibility: Capturing the “Whole Tale”// Future Generation Computer System. – 2019. - №94. - pp.854-867. Режим доступа: https://www.sciencedirect.com/science/article/pii/S0167739X17310695 (дата обращения 3.06.2019).
3. Castelli Donatella, Manghi Paolo, Thanos Constantino. A vision towards Scientific Communication Infrastructures // International Journal of Digital Libraries. – 2013. - №13. - pp.155-169. – Режим доступа:

https://link.springer.com/content/pdf/10.1007%2Fs00799-013-0106-7.pdf (дата обращения 3.06.2019).

4. Ayris P., Ignat T. Defining the role of libraries in the Open Science landscape: a reflection on current European practice // Open Information Science. 2018. № 2. pp. 1-22. URL: https://doi.org/10.1515/opis-2018-0001 (дата обращения 3.06.2019).
5. Жижимов О. Л., Мазов Н. А., Федотов А. М. Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей. - 2010. - Т.3. - № 7. - С. 55-63. - ISSN 2073-3984.
6. Gonzales Brighid M. Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record // Information Technology and Libraries. – 2014. – декабрь. – Режим доступа: https://ejournals.bc.edu/ojs/index.php/ital/article/view/5631 (дата обращения 3.06.2019).
7. Mitchell, Eric T. Library Linked Data: Early Activity & Development / Eric T. Mitchell // Library Technology Reports. – 2016. - Vol.52. - № 1. – pp. 18-23. – Режим доступа: https://journals.ala.org/ltr/issue/download/534/290 (дата обращения 3.06.2019).
8. Паринов С. И. «Живые» документы в электронных библиотеках [Электронный ресурс] / С. И. Паринов, М. Р. Когаловский // Прикладная информатика. – 2009. - № 6. – URL: http://www.cemi.rssi.ru/mei/articles/koga-pari09-2.pdf (дата обращения 3.06.2019).
9. Clobridge Abby. Libraries in Transition: From book collections & union catalogues to open access & digital repositories // ProInflow : Časopis pro informační vědy. – 2011. - № 2. – pp. 121-132.
10. Поляков А. Ю. Комплексное решение на основе речевых технологий для президентской библиотеки имени Б.Н.Ельцина [Электронный ресурс] / А. Ю. Поляков, Д. В. Дырмовский, А. В. Рыбаков // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: материалы конф. – Электрон. дан. – М.: ГПНТБ России, 2009. – 1 электрон. опт. диск (CD-ROM). – Систем. требования: IBM PC, Windows 2000 или выше. – Загл. с этикетки диска. – ISBN 978-5-85638-132-9. – № гос. регистрации 0320900806.